A Method for Analysis of Node Position in the Network of Internet Users

Katarzyna Musiał

Institute of Computer Science
Wroclaw University of Technology

Thesis submitted for the Degree of Doctor of Philosophy in the
Wroclaw University of Technology

· 2009 ·

Contents

Acknowledgments **xiii**

Important Symbols **xiv**

Abbreviations **xv**

Abstract **xvi**

1 Introduction **1**

2 Social Networks **7**

2.1 Regular Social Network 7

2.2 Network of Internet Users 14

2.2.1 Concept of the Network of Internet Users 15

2.2.2 Taxonomy of the Networks of Internet Users 19

2.3 Internet Identity — Node of the Network of Internet Users . . 21

2.3.1 Concept of the Internet Identities 21

2.3.2 Individual and Group Internet Identity 26

2.3.3 Internet Identities Integration 28

2.4 Internet Relationships — Edges of the Network of Internet Users 30

2.4.1 Concept of the Internet Relationship 30

2.4.2 Types of the Internet Relationships 31

2.4.3 Directness of the Internet Relationship 34

2.4.4 Ties 36

2.5 Examples of the Networks of Internet Users 38

2.5.1 Electronic Mail Services 38

2.5.2 Instant Messengers 38

2.5.3 Blogs Services 39

2.5.4 Social Networking Sites 39

2.5.5 Multimedia Sharing Systems 40

2.5.6 Auction Systems 41

2.5.7 Social Search Engines 42

2.5.8 Social Bookmarking and Cataloging 42

2.5.9 Homepages 43

2.5.10 Knowledge Sharing Systems 43

2.5.11 Virtual Worlds and Multiplayer Online Games 44

2.5.12 Collaborative Authoring Systems — Wikis 44

2.5.13 Friend Of A Friend Project 45

2.5.14 Complex Communications Systems 45

2.5.15 Comparison 45

3 User Position in Social Network 51

3.1 Social Network Analysis 51

3.2 User Position 55

3.2.1 Prestige 55

3.2.2 Centrality 57

3.3 Comparison of the User Position Indices 59

4 Node Position in Network of Internet Users 64

4.1 General Concept 64

4.2 Commitment Evaluation 67

4.3 Node Position Calculation 69

4.3.1 PIN Nodes 70

4.3.2 PIN Edges 71

4.3.3 PIN Hybrid 72

4.4 Example of Node Position Calculation 74

4.4.1 The Influence of ε 74

4.4.2 The Influence of Initial Node Positions 79

4.4.3 The Convergence of Node Position Method 80

4.4.4 Example — Conclusions 83

5 Formal Analysis of the Node Position Method 84

5.1 Total and Average Node Position 84

5.2 Convergence 86

5.3 Interval of Limit Values 89

6 Research 92

6.1 Thurman Network 93

6.1.1 Data Description and Preparation 93

6.1.2 Comparison of Node Position with Other Centrality Methods 95

6.2 Enron Network 99

6.2.1 Data Description and Preparation 99

6.2.2 Characteristics of Node Position 101

6.3 Wroclaw University of Technology Network 106

6.3.1 Data Description and Preparation 106

6.3.2 Influence of stopping condition τ and ε coefficient . . . 106

6.3.3 Comparison with Other User Position Indices 114

6.4 Efficiency Tests 117

6.4.1 Influence of ε on Processing Time 117

6.4.2 Influence of Network Size on Processing Time 118

6.4.3 Efficiency of Node Position versus Other User Position Indices 126

6.5 Possible Application of Node Position Method 127

6.5.1 Method for Bridges' Properties Analysis 128

6.5.2 Experiments 129

7 Conclusions and Future Work 134

References 138

Appendix A 149

Appendix B 151

Appendix C 161

List of Figures

2.1 A regular social network 9

2.2 The division of social networks based on the type of the relationship 12

2.3 An example of relationships between people working in a company 13

2.4 The division of social networks based on the type of the communication channel 13

2.5 Homogeneous *HSN*, system–based (*SSN*), and internet multimodal social network (*ISN*) 15

2.6 Examples of homogeneous *HSN*, system–based (*SSN*), and internet multimodal social network (*ISN*) 16

2.7 Real time vs. non real time networks 20

2.8 Mapping of social entities into internet identities 22

2.9 The concept of internet identities merging 26

2.10 Integration of two system–based social networks by means of internet identity merging 29

2.11 Taxonomy of internet relationships 32

2.12 The direction of relationships [33] 33

2.13 The direct internet relationship in the social network on the Internet 35

2.14 The quasi–direct internet realtionship in the social network on the Internet 35

2.15 The object–based internet relationship with equal roles: commentator (a), and different roles: commentator and author (b) 36

2.16 The indirect internet relationship in the social network on the Internet 36

2.17 The tie concept in the social network 37

2.18 Main functions of the social networking site related to relationship maintenance 40

3.1 Methods of social network analysis 54

4.1 Example of the network of Internet users *NIU* with the assigned commitment values 65

4.2 Distribution of the commitment for an inactive member y equally among all y's acquaintances 66

4.3 The network of Internet users that can be extracted from the email communication 74

4.4 The values of members' node positions in relation to ε value . 77

4.5 The value of social position in relation to ε 78

4.6 The minimum, maximum, average, and standard deviation of node position calculated for the same community but for different values of ε 78

4.7 The linear regression for average node position and standard deviation 79

4.8 The values of node position after every iteration for $\varepsilon = 0.95$. 81

4.9 The values of the member x's node position after every iteration for various ε 81

4.10 The convergence of the node position sum for various ε and various initial sums 82

5.1 The community where individual x has the greatest node position 89

5.2 The chart of the function $f(\varepsilon)$ for the network that consists of m Internet users 91

5.3 The range of the node position values 91

6.1 Graph representation of the classic Thurman network 93

6.2 The comparison of centrality measures to node position in the Thurman network 96

6.3 The comparison of prestige measures to node position in the Thurman network 96

6.4 The comparison of eigenvector measures with node position in the Thurman network 97

6.5 The values of Kendall's coefficient for the pairs of rankings from Table 6.4 100

6.6 Average NP and $NPwTF$ as well as their standard deviations in the Enron dataset, calculated for different values of ε 102

6.7 The percentage of users with $NP < 1$ and $NP \geq 1$ within the Enron network in relation to ε 102

6.8 The distribution of NP depending on ε 103

6.9 Percentage of duplicates within the set of node measures, separately for node position with different values of ε, degree prestige (IDC), and degree centrality (ODC) 103

6.10 The percentage contribution of members with $NP \geq NPwTF$ and $NP < NPwTF$ within the Enron social network in relation to ε 105

6.11 The mean squared error between NP and $NPwTF$ for the Enron dataset in relation to ε 106

6.12 Social network discovered from the email communication between employees of WUT 107

6.13 Processing time [min] of the PIN^{edges} depending on different values of ε coefficient and stop condition τ 109

6.14 Processing time [min] of the PIN^{edges} depending on different values of coefficient ε 109

6.15 Processing time [min] of the PIN^{edges} depending on different value stop condition τ 110

6.16 Number of required iterations depending on different values of ε coefficient and stopping condition τ 111

6.17 Processing time [min] of the PIN^{edges} depending on different values of ε coefficient and stop condition τ 112

6.18 Exponential fitting function for the relation between ε and number of iterations 113

6.19 Exponential fitting function for the relation between ε and number of duplicates 115

6.20 Processing time of the PIN^{nodes} depending on different size of network 120

6.21 Processing time of the PIN^{edges} depending on different size of network 121

6.22 Processing time of the PIN^{hybrid} depending on different size of network 122

6.23 Processing time depending on the number of nodes in the network for fixed number of edges (50,000) 123

6.24 Processing time depending on the number of edges in the network for fixed number of nodes (50,000) 125

6.25 Crucial steps of the proposed method of bridges properties analysis 128

6.26 Structure of peripheral nodes in Thurman network and bridges that connect PN with the regular cliques 131

6.27 Structure of peripheral clique in Thurman network and bridges that connect PC with the regular cliques 131

7.1 The relation layers in Flickr 152

7.2 Relation layers extracted from contact lists: R^{c} (a), R^{rc} (b), and R^{coc} (c) 153

7.3 Extraction of tag–based relations 155

7.4 Extraction of relations based on favorites 156

List of Tables

2.1 Existing definitions of an actor, a relation, and a regular social network 10

2.2 Layers in system-based social networks on the Internet 46

2.3 Features of the system–based social networks 48

3.1 Example of simple social data 52

3.2 Example of social network data. 0 means that person A does not know person B, 1 means that person A knows person B . 52

3.3 Groups of the user position indices 59

3.4 Comparison of User Position Indices 60

3.5 Advantages and Disadvantages of User Position Indices 62

4.1 Node position calculation for the network from Figure 4.3; $\varepsilon = 0.1$ 75

4.2 Node position calculation for the network from Figure 4.3; $\varepsilon = 0.5$ 75

4.3 Node position calculation for the network from Figure 4.3; $\varepsilon = 0.9$ 75

4.4 The number of iterations in relation to the value of ε 76

4.5 Node position calculation for different sets of their initial values ($\varepsilon = 0.5$) 80

4.6 Node position calculation for different sets of their initial values ($\varepsilon = 0.9$) 80

4.7 The initial node positions for set *NP_01* and *NP_02* 82

6.1 The values of commitment function within the Thurman network 94

6.2 The values of the analyzed measures in the Thurman office social network for each member of the network 95

6.3 The positions in rankings for the analyzed measures in the Thurman office social network, for $\varepsilon = 0.9$ 98

6.4 Kendall's coefficient for each pair of rankings from Table 6.3 . 99

6.5 Average NP and $NPwTF$ in the Enron dataset, calculated for different values of ε 101

6.6 Standard deviations of NP and $NPwTF$ in the Enron dataset, calculated for different values of ε 101

6.7 Kendall's coefficient for Enron 105

6.8 The statistical information for the WUT datasets 107

6.9 Processing time [min] of the PIN^{edges} depending on different values of ε coefficient and stopping condition τ 108

6.10 Number of iterations depending on different values of ε coefficient and stop condition τ 111

6.11 Number of duplicates depending on different values of ε coefficient and stopping condition τ 112

6.12 Parameters values for correlation function $n(\varepsilon)$ 114

6.13 Parameters values for correlation function $d(\varepsilon)$ 115

6.14 Number of duplicates and processing time for different centrality indices 116

6.15 Processing time of the PIN algorithm depending on different values of ε coefficient 117

6.16 Average processing time of one iteration and its standard deviation for different PIN algorithms 118

6.17 The random networks generated for the needs of the efficiency tests 119

6.18 Processing time of the PIN^{nodes} depending on different size of network 119

6.19 Processing time of the PIN^{edges} depending on different size of network 120

6.20 Processing time of the PIN^{hybrid} depending on different size of network 121

6.21 Processing time in relation to the number of nodes in the network for fixed number of edges (50,000) 122

6.22 The relation of processing times of PIN^{edges} to other PIN algorithms for fixed number of edges (50,000) 123

6.23 The ratio of processing time and number of nodes for different PIN algorithms for constant number of edges (50,000) 124

6.24 Processing time depending on the number of nodes in the network for fixed number of nodes (50,000) 124

6.25 The relation of processing times of PIN^{edges} to other PIN algorithms for fixed number of nodes (50,000) 125

6.26 The ratio of processing time and number of edges for different PIN algorithms for constant number of nodes (50,000) 125

6.27 Processing time for different centrality indices 127

6.28 The values of commitment function within the Thurman network130

6.29 The node position values and number of edges in Thurman network 132

Acknowledgments

I would like to start with the most interesting part of this dissertation — the acknowledgments. There are many people who contributed to my thesis and many events that influenced my work during the last few years. I would like at least to mention them here.

First of all, I want to thank Przemysław Kazienko. I am very grateful that he appeared on my scientific way and encouraged me to conduct the research in the field of complex network systems. I would also like to thank my advisor Aleksander Zgrzywa for guiding my work during the last few years.

I am grateful to my friends whom I worked with on the problems presented in this dissertation: Krzysztof Juszczyszyn for discussing issues related to social networks and for always being ready for debates on my research and not only; Piotr Bródka who helped me a lot during the experiments on the node position measure.

Last, but not least, I am grateful to my Family as well as Friends, especially Ania Musiał and Ryszard Bojarski for their continuous and unconditional support. This dissertation could not have been brought to completion without their patience, understanding and love.

Important Symbols

ε – the constant coefficient from the range $[0, 1]$;

$C(x, y)$ — the commitment function from user x to y;

m_x — the number of x's acquaintances;

m — the number of the nodes within the network of Internet users;

$NP(x)$ — the node position of user x;

τ — the stopping condition for the *PIN* algorithms

Abbreviations

BC — Betweenness Centrality

C — Commitment

CC — Closeness Centrality

CMC — Computer—Mediated Communication

CSSN — Computer—Supported Social Network

DC — Degree Centrality

DSSN — Device Supported Social Network

FOAF — Friend Of A Friend

HSN — Homogeneous Social Network

IDC — Indegree Centrality

IID — Internet Identity

IM — Instant Messenger

ISN — Internet Multimodal Social Network

MO — Multimedia Object

NIU — Network of Internet Users

NP — Node Position

NPwTF — Node Position with Time Factor

ODC — Outdegree Centrality

ODCE — Outdegree Eigenvector Centrality

PIN — Position In Network (name of the algorithm)

PP — Proximity Prestige

PPE — Proximity Prestige Eigenvector

RP — Rank Prestige

RSN — Regular Social Network

R — Internet Relationship

SN — Social Network

SNA — Social Network Analysis

SNS — Social Networking Sites

SSN — System–based Social Network

T — Tie

VIID — Virtual Internet Identity

Abstract

This dissertation deals with Internet–based social networks, where both nodes and relations have clear technical interpretation. However, well defined in technical terms, networks of Internet users are not well analyzed due to dynamics and complexity. Multidimensionality, hard–to–define before Internet, now may be investigated, but requires new algorithms and techniques. One of the algorithms proposed in this thesis that can be used in such a complex environment is the node position method that is used to discover the nodes that are important for a given Internet community. Important means that a node is perceived as the prominent by others and it is expressed by the fact that they communicate or share common activities with this node. Furthermore, the node is important if the nodes with high node position communicate with it because its position depends on the position of its neighbors. Moreover, the new node position method takes into account also the fact that it changes over time. The additional criterion that must be met by the developed method is its computational efficiency while applying it to large multirelational networks. Thus, it is necessary to provide the mechanism that enables to make a trade–off between the accuracy of the calculations and the time needed to perform them.

Chapter 1

Introduction

Overall characteristic of the research domain

With the development of the Internet and such concepts as Web 2.0 [100] or collective intelligence [83] as well as increasing popularity of social computing [109], [118], the complex social networks in the Internet have emerged as an important and promising field of research within computer science. All approaches to social networks have their origin in the concept of society by emphasizing the role of the connections between people not the individuals themselves [55], [119]. The social networks existing in the real world will be called in this thesis regular social networks (*RSN*). The general concept of the social network is quite simple and intuitive. It is a set of actors, i.e. a group of people or organizations or other social entities, which are the nodes of the network, and ties that link the nodes. The social network describes the ways in which actors are related to each other and defines the relationships between friends, co–workers, members of the particular society, relatives in the family, etc.

Although the general definitions of both regular social networks that can be extracted from the data about people and their interactions in the real world and social networks existing in the Internet are similar, their characteristics differ a lot. In result the knowledge derived from the studies on the regular social networks cannot be directly mapped onto the social networks existing in the virtual world. The Internet as a relatively new medium has created a new class of social networks that need to be analyzed and classified. Note that the Internet provides a vast amount of diverse data useful for social network analysis (*SNA*). Internet–based social networks can be either directly maintained by web systems like Friendster [16], MySpace [3], or LinkedIn [34] or extracted from data about user activities in the communication networks like e–mails, chats, blogs, homepages connected by hyperlinks, etc. [1]. The digital representation of the user referred to from now on **I**nternet **Id**entity (*IID*) and the connections between them called **I**nternet **R**elationships (*R*) can be characterized and described in many different ways, e.g. can be represented as a matrix or graph where Internet identity can be seen as nodes and relations as edges of the graph.

In the last few decades various methods of analysis have been developed in order to investigate the features of social networks [18], [19], [26], [110], [119]. Vast majority of these methods can be applied at such levels of analysis as single nodes, groups of vertices or a network as a whole. The characteristics used in the process of network analysis are e.g. centrality, density, cliques detection, etc. The decision regarding which method to use and at which of the enumerated levels depends on the knowledge that is needed from the researcher point of view.

One of the measures that is the object of the continuous interest of many researchers is the centrality index [18]. It serves to estimate the position of an individual in the network as a whole or in the group of people. Different ways of evaluating the value of this measure are utilized depending on the users needs. The most popular and well–known are as follows: degree centrality [119], closseness centrality [7], betweeness centrality [26], rank prestige [119], etc. It should be emphasized that all of the developed methods in the area of *SNA* are quite useful and effective in small and medium sized networks. However, most of them fail while applying them to the complex networks such as these existing in the Internet where we face a problem of vast amount of data.

Thesis Objectives and Contribution

In the Internet users can 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 **N**etwork of **I**nternet **U**sers (*NIU*) creation. One of the most meaningful and useful measures in the 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 the node position assessment becomes very complex because humans with their spontaneous and social behavior are hardly 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 position and probably the highest level of trust. These users can be representatives of the entire community. A small group of key persons can initiate new kinds of actions, disseminate new services or activate other network members. On the other hand, users with the lowest position should be stimulated for greater activity or be treated as the mass, target receivers of the pre–prepared services that do not require the high level of involvement. Moreover, one of the very interesting elements is the dynamics of social networks in context of the individuals positions. A very promising field of research is the investigation of the influence of adding to or removing from social network actors with high/low position on the topology of the whole structure. These various opportunities of applications are the main motivations for a development of a the new method of node position estimation.

The main goal of this thesis is to develop a new method of Node Position (*NP*) estimation in a network of Internet users that takes into account following facts:

- **Node Position changes over time**
- **Node Position dependence on the positions of the other nodes**
- **Node Position dependence on the strength of the relations between users**
- **Node Position for one– and multirelational complex networks**

Thus the thesis of this work is as follows: **it is possible to create the ranking of nodes of one– and multirelational network of Internet users based on their position. High node position of a node means that many nodes notice a given node as important one and in consequence stay in the relationship with it**.

Another motivation to develop the new method of assessing the user centrality is that the existing methods tend to be very inefficient when applied to the complex social networks, such as these existing in the Internet. Note that, efficiency is one of the most important factors that must be taken into account during analyzing networks with a large number of nodes and connections. On the other hand, the developed method should provide the sufficient accuracy of calculations. Since better accuracy requires more resources the trade–off between performance and accuracy of the computations has to be addressed. Thus, the additional **criterion that must be met by the developed 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**.

In order to achieve the defined goals the following objectives were established and the realization of which is the main contribution to the development of the research area that is called complex networked systems:

1. To classify and define the types of networks of users that can be extracted from the Internet;
2. To define the following terms: Internet identity (*IID*) and Internet relationship (*R*);
3. To prepare the state–of–the–art of the existing methods of user position assessment;
4. **To develop the new node position method that takes into account both the position of other nodes as well as the strength of relationships**;

5. **To propose the methods of relationship strength (called from now on commitment function) evaluation for both static data and the data that changes in time as well as one– and multirelational social networks**.

The conducted research, the outcome of which is presented in this thesis, combines different fields of research such as social networks, Web mining and graph theory into a new interdisciplinary area.

Outline of the Thesis

The thesis is on the social networks in the Internet and the new method of node position assessment. The whole dissertation consists of **four major parts**.

The first part presents the state–of–the–art of the social networks research domain. The issues that are addressed in this part concern both the regular social networks and the networks existing in the Internet (Chapter 2), the main aspects of social network analysis (Section 3.1) and the existing methods of user position assessing. Additionally, the comparison of the existed methods is provided (Chapter 3). In this stage the definitions required in later chapters are introduced. The crucial described concepts are: Internet Identity (*IID*), Internet Relationship *R*, Homogeneous Social Network *HSN*, System–based Social Network *SSN*, and Internet Multisystem Social Network *ISN*. Moreover, the author also proposes the classification of social networks existing in the Internet and presents examples for each of the created classes of social networks (Chapter 2).

The second part is devoted to a detailed description and analysis of the new method of the node position evaluation within the network of Internet users (Chapter 4). The Node Position $NP(x)$ of user x respects both node position value of user x acquaintances as well as the strength of the relationships that other users maintain with the user x. Not only all elements of the method but also a simple example that vividly presents the concept of node position calculation is presented. Three different algorithms of Node Position assessment are proposed in the thesis. Moreover, the formal analysis of the proposed method which includes the complexity analysis of the proposed algorithms as well as the theorems and proofs regarding the Node Position characteristics are described (Chapter 5).

The next, third part is devoted to the experiments that were conducted (Chapter 6). The elements that were investigated are as follows:

1. General characteristic of the method, such as maximum and minimum value, mean value, standard deviation, etc.

2. The influence of the method parameters: ε, τ on the outcomes of the methods.

3. Comparison of Node Position characteristics with features of other centrality indices, e.g. distribution of values, number of duplicates, etc.

4. Efficiency tests

   — the efficiency of three developed algorithms is compared,

   — the efficiency of the proposed method of Node Position assessment with other measures is investigated, In the last, fourth part the conclusions that were drawn during the performed research and the possible applications of proposed the node position method are presented (Chapter 7).

Chapter 2

Social Networks

The goal of this chapter is to introduce the basic concept of a regular social network and its representation as well as the taxonomies of social networks. Moreover, the aim of this part is to present the research area from which the social networks of Internet users originate. Finally, the concept and types of networks of Internet users are described.

2.1 Regular Social Network

The social networks (*SN*s), which in this thesis are also called regular social networks (*RSN*s) have recently become a very actively researched area and are regarded as an important element of information society [1], [55]. This is due to a huge variety of existing social networks and many possible areas where they can be applied. Since the relationships from the network, their maintenance and quality reflect social behavior of individuals, the research on them can be helpful when carrying out the quantitative and qualitative assessment of human relationships. The concept of *SN* is utilized to describe the relationships between friends, co–workers, members of a particular society, relatives in the family, etc. Not only the character of the relationships can be analyzed, but also their strength and direction. Although social network analysis (*SNA*) emphasizes the connections between people, the results of *SNA* provide also much information about individuals themselves. There are many, different kinds of social networks and the taxonomy of the social networks is not established. Research in a number of scientific fields have demonstrated that social networks emerge on many levels, from families up to the level of nations, and play a critical role in determining the way in which problems are solved, organizations are run, and the degree to which individuals achieve their goals.

The concept of social network, first coined in 1954 by J. A. Barnes in [6], has been in a field of study of modern sociology, anthropology, geography, social psychology and organizational studies for last the few decades.

The person who built the modern social network theory was Stanley Milgram. He studied the small–world phenomenon, which states that even if persons x and y do not know each other directly, they can share a mutual relationship that is another person who knows them both [91], [114]. The theoretical model of this small–world phenomenon was created by Pool and Kochen [91] and served as the basis for Milgram's research that was purely pictorial. Stanley Milgram conducted two experiments — Kansas Study and Nebraska Study — in which he asked many people from one city to forward a letter to a chosen person in another city. The only stipulation was that a sender could only forward this letter to a person whom he or she knew on a first—name basis. Afterward Milgram analyzed the results of the experiment and concluded that people in the USA create the social network and they are connected within this network with "six degrees of separation". It means that a message in such a network would be delivered on average through the usage of five intermediaries [91]. Kochen confirmed that this value is relatively stable even if the starter selection criteria is changed [33]. Howard claims that six degrees of separation may by true offline while less than three degrees is more likely in an online case [61].

Since 1967 social networks have become one of the research areas where scientists from different fields are looking for inspiration. The social network analysis supported by computer science gives the opportunity to develop and expand other branches of knowledge.

The concept of social network has been studied in many different contexts, e.g. corporate partnership networks (law partnership) [81], scientist collaboration networks [98], [37], movie-actor networks, friendship network of students [5], a set of business leaders who cooperate with one another [84], [106], sexual contact networks [95], customers networks [126], [67], [49], labours market [93], public health [27], psychology [101], etc.

The general concept of society can be considered as the background for the social network definition. A society is not merely a simple aggregation of individuals; it is rather the sum of relationships that connect these individuals to one another [87].

The main idea of social network is simple. It is the set of actors i.e. group of people or organizations, which are the nodes of the network, and ties that link the nodes [1], [119], [55] (Figure 2.1). Social network indicates the ways in which actors are related. The tie between actors can be maintained according to either one or several relations [46] that can be directed or undirected, weighted or unweighted. Moreover, the network gives egos (focal actors) access not only to their alters (people that are directly connected with ego), but also to alters of their alters [46] (also called "friends of my friends").

The nodes of the social network are not independent beings. Some of the characteristics that describe members of the network can be defined (e.g. demographic and interest data about people). However, none of *SNA* methods samples the individuals independently. The actors are connected via relationships, which are characterized by content, direction and strength [46]. The content indicates the resource that is exchanged, e.g. in computer—mediated communication (*CMC*) the information can be treated as the resource [46].

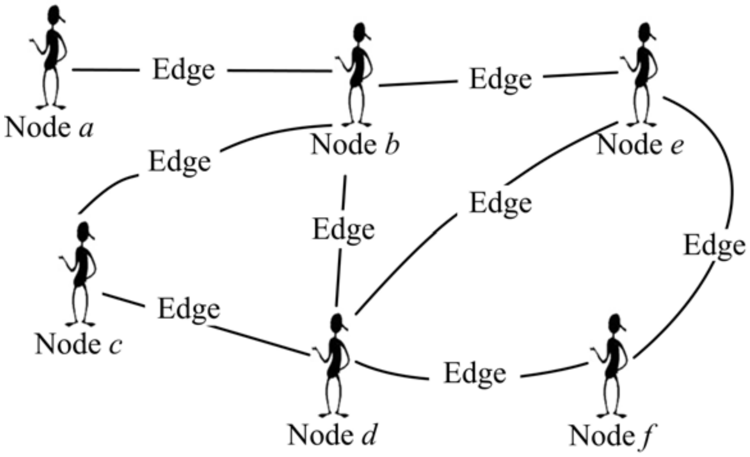


Figure 2.1: A regular social network

The direction determines if the relationship is directed or undirected. The relationship between employees and their supervisor is directed. The former works for a supervisor and this is a relationship between an employee and the boss. The latter pays wages or a salary to the employees and this is another directed relation between the boss and employees. A friendship is usually undirected, nevertheless it can be unbalanced. It means that one person can define the friendship with another person as strong, whereas the other person can claim that this friendship is weak [46]. The last of the enumerated characteristics of the relationship is its strength. There are many ways to determine whether a relationship is strong or weak [86], [123], e.g. through specifying the frequency with which actors communicate with each other, importance of exchanged information, and the amount of social capital sent from one actor to another [46].

Although the concept of social network seems to be quite obvious, every researcher defines the social network in a slightly different way. Some of them define the social network in a very formal way, e.g. Yang, Dia, Cheng, and Lin [126] while others prefer more sociological approach [119], [58]. More insight into the problem of the social network definition is presented in the Table 2.1. The listing shows how the concept of an actor, a relation, and a social network is described by different researchers. It appears that definitions from Table 2.1 are the most representative although there are many other scientists that have investigated the concept of social network [33], [26], [110]. However, other definitions are in fact a mixture of the presented ones.

Table 2.1: Existing definitions of an actor, a relation, and a regular social network

Author	**Actor**	**Relation**	**RSN**	**Examples of RSN**
Wasserman and Faust [119]	An actor is a discrete individual, corporate or collective social units	A set of ties of a specific type; a tie is a linkage between a pair of actors	The finite set or sets of actors and one or more relations defined on them	Friendship among children in a classroom; all nations in the world and the formal diplomatic connections between them
Hanneman and Riddle [55]	Actors are also called points, nodes or agents	Relationships, edges or ties; one or more kinds of relations between pairs of actors	A set of actors that may have relationships with one another	Family; co-workers in a company; the network of neighbors; friendship among students in a classroom
Garton, Haythornthwaite, and Wellman [46]	People, organizations or other social entities	Relationships, such as friendship, co-working or information exchange	A set of social entities connected by a set of social relationships	Friendship among people; co-workers in a company; people who communicate with one another via computer

Author	**Actor**	**Relation**	**RSN**	**Examples of RSN**
Hatala [58]	Actors are people or groups of people	Patterns of interaction or ties between actors	A set of actors with some patterns of interaction or ”ties” between them; represented by graphs or diagrams illustrating the dynamics of the various connections and relationships within the group	Co-workers within a company
Liben-Nowell and Kleinberg [84]	People or other entities embedded in the social context	Edges represent interaction, collaboration, or influence between entities	Structures whose nodes represent entities embedded in the social context, and whose edges represent interaction, collaboration, or influence between entities	Co-authors of the scientific papers in a particular discipline; project groups in a large company; business leaders who have served together on a corporate board of directors
Yang, Dia, Cheng, and Lin [126]	A node in a graph; each node represents a customer	The undirected, unweighted edges in the graph; each edge represents the connectedness between two nodes	An undirected, unweighted graph	Customer’s social network which is derived from customer’s interaction data

Many scientists tried to classify social networks and create taxonomies [120], [46]. Nevertheless, they considered only some specific subset of existing social networks, e.g. Barry Wellman described computer–supported social networks (*CSSN*) [120]. It is very hard and complex task to build one coherent and complete classification of the regular social network and there is no established one.

To put the different kinds of regular social networks in order, they can be classified based on the type of the relationship that connects two persons. In this case, business and social connections can be distinguished (Figure 2.2). The former ones contain social networks that consist of people who are linked with each other because of things they do together but simultaneously they do not share their private lives. Those can be called professional networks, e.g. in a company — employees create the social network of co–workers. Also, people who organize together e.g. a conference or other event, create social network of co–organizers. These people are connected because they work together and their cooperation usually brings some outcome, e.g. an article, a conference, a book, etc.

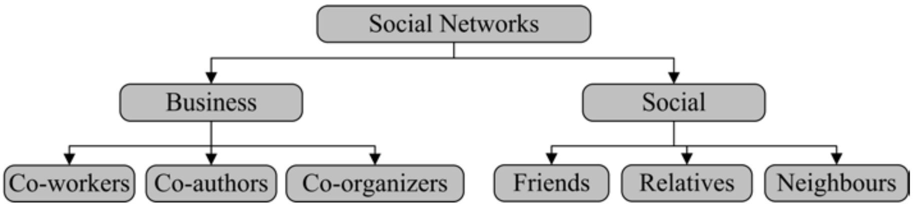


Figure 2.2: The division of social networks based on the type of the relationship

On the other hand, the social relationships indicate the connections with emotional background. Relatives are the group of people that we are family with; nevertheless usually people are not in touch with every member of their family.

The thing that should be emphasized is that the tie between two persons is usually the combination of many different kinds of relationships, which can differ in strength. Figure 2.3 presents a theoretical situation of people who are employed in one company. They are not only co–workers, but also other relationships exist between them. For example, although person z and v work in the same company, they are not co–workers but friends. The fact that two people are employed in the same organization does not mean that they work together.

Moreover, each of the relationships can differ in direction and strength. For example, person y can claim that he/she is a really good friend of x, whereas x can admit that y is a friend but not so close as y thinks.

Additionally, the classification of social networks can be based not only on the type of relations that occur in the network, but also on the type of the communication channel between members that serves to exchange resources i.e. they can be either in person or device supported (virtual, via computer, phone, snail mail, etc.). This is illustrated in Figure 2.4.

Note that social networks existing in the real world are much more tangi-

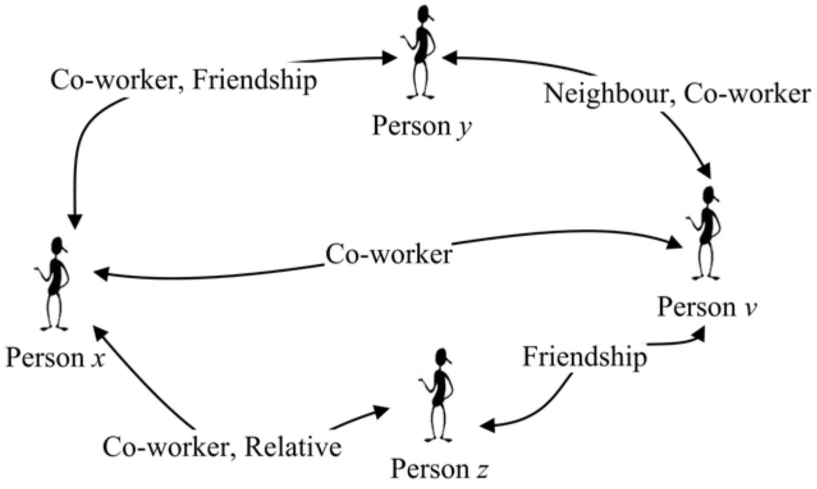


Figure 2.3: An example of relationships between people working in a company

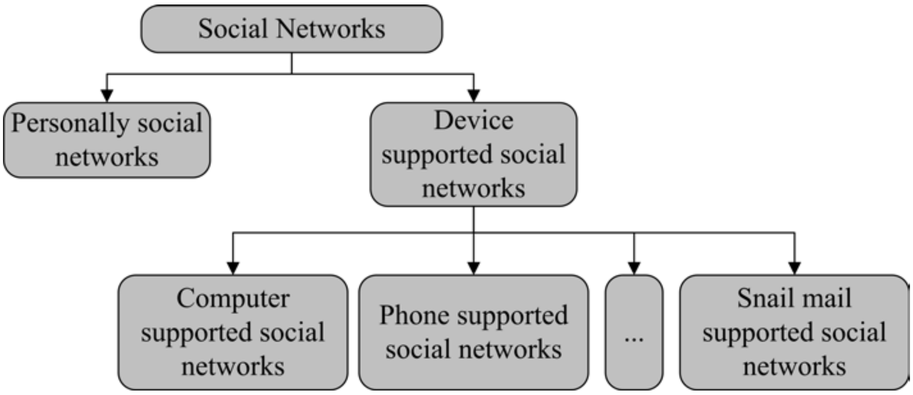

Social Networks
Personally social networks
Device supported social networks
Computer supported social networks
Phone supported social networks
...
Snail mail supported social networks

Figure 2.4: The division of social networks based on the type of the communication channel

ble while device supported social networks (*DSSN*) suffer from limited social presence [120]. In the "in person" networks not only the words and information are important, but also verbal physical context, nonverbal cues, and observable information about social characteristics. These elements do not occur in *DSSN* [120]. On contrary to them "in person" *SN*, *DSSN* enable communication between people who are in different places on the globe. This taxonomy similarly to the previous one illustrated in Figure 2.2 does not exclude the situation in which two persons communicate in more than one way, e.g. two people can both write e–mails to each other and meet together personally.

The proposed above classifications are not the only possible ones. However, the social networks are very complex systems and can be analyzed from different points of view. Thus, the researchers have the opportunity to use the most suitable taxonomy for their experiments.

2.2 Network of Internet Users

The continuously increasing popularity of the Internet resulted in greater availability of various types of services over the computer network. People who use these services have created a new kind of virtual societies usually called social networks of Internet users. These are also often referred to as online social networks [29], [46], [61], [82], web-based social networks [49], [48], computer–supported social networks [120] or virtual communities [1].

The main features that distinguish social networks on the Internet from the regular ones are as follows:

1. Lack of physical, personal contact — only by distance, even very long distances;

2. In many cases the lack of unambiguous and reliable correlation between member's identity in the virtual community, i.e. internet identity (see Section 2.3) and their identity in the real world;

3. The possibility of multimodal communication; simultaneously with many members but also the possibility of easy switching between different communication channels, especially online and offline, e.g. online VoIP and offline text communication;

4. The simplicity of a break up and suspension of contacts or relationships;

5. The relatively high ease of gathering the data about communication or common activities and its further processing. Mining of Internet–based social networks is easier and social dimensions are more definable when compared to standard social networks;

6. Reduced reliability of the data about users and their activities available on the Internet. Users of internet services relatively frequently provide fake personal data due to privacy concerns.

In the literature, the name web communities was firstly used to describe the set of web pages that deal with the same topic [47], [41]. Adamic and Adar [1] 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 an email communication system [1], [31], [112]. On the other hand, a computer–supported social network introduced in [46], [120] appears when a computer network connects people or organizations. Finally, Golbeck affirms that a web–based social network must fulfill the following criteria: users must explicitly establish their relationships with others, the system must have explicit support for making connections, and relationships must be visible and browsable [48]. Mainly social networking sites like LinkedIn [34] or MySpace [28] meet these criteria.

Based on the kind of service people use, many examples of the social networks in the Internet can be enumerated. To the most commonly known belong: a set of people who date using an online dating system [16], a group of people who are linked to one another by hyperlinks on their homepages [1], customers who buy similar stuffs in the same e-commerce [126], the company staff that communicates with one another via email [2], [68], [112], [31], [127], people who share information by utilizing shared bookmarking systems [92] such as del.icio.us. Yet another multirelational social network can be established within the multimedia sharing system like Flickr [97] or YouTube.

2.2.1 Concept of the Network of Internet Users

Since many different types of social networks can be distinguished on the Internet, let us try to specify some basic definitions of basic kinds of Network of Internet Users (*NIU*), beginning from the simplest **homogeneous social network**, through the **system–based social network** to the most complex **Internet multisystem social network** (Figure 2.5 and Figure 2.6). More detailed insight into various kinds of social networks can be found in Section 2.5.

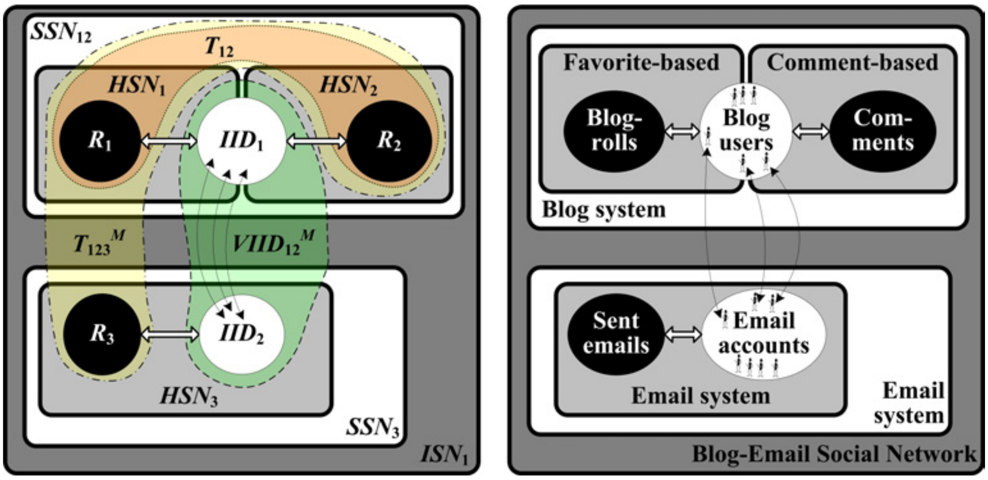


Figure 2.5: Homogeneous *HSN*, system–based (*SSN*), and internet multi-modal social network (*ISN*)

Definition 2.2.1 *A homogeneous social network on the Internet HSN=(IID,R) exists within a single internet–based system S. It consists of the finite set of internet identities IID — registered, non–anonymous users of the internet system S, and the set of all internet relationships R of the same kind that join pairs of IID members:*
$R = \{(iid_i, iid_j) : iid_i \in IID, iid_j \in IID\}$. *HSN is also called the single layer social network.*

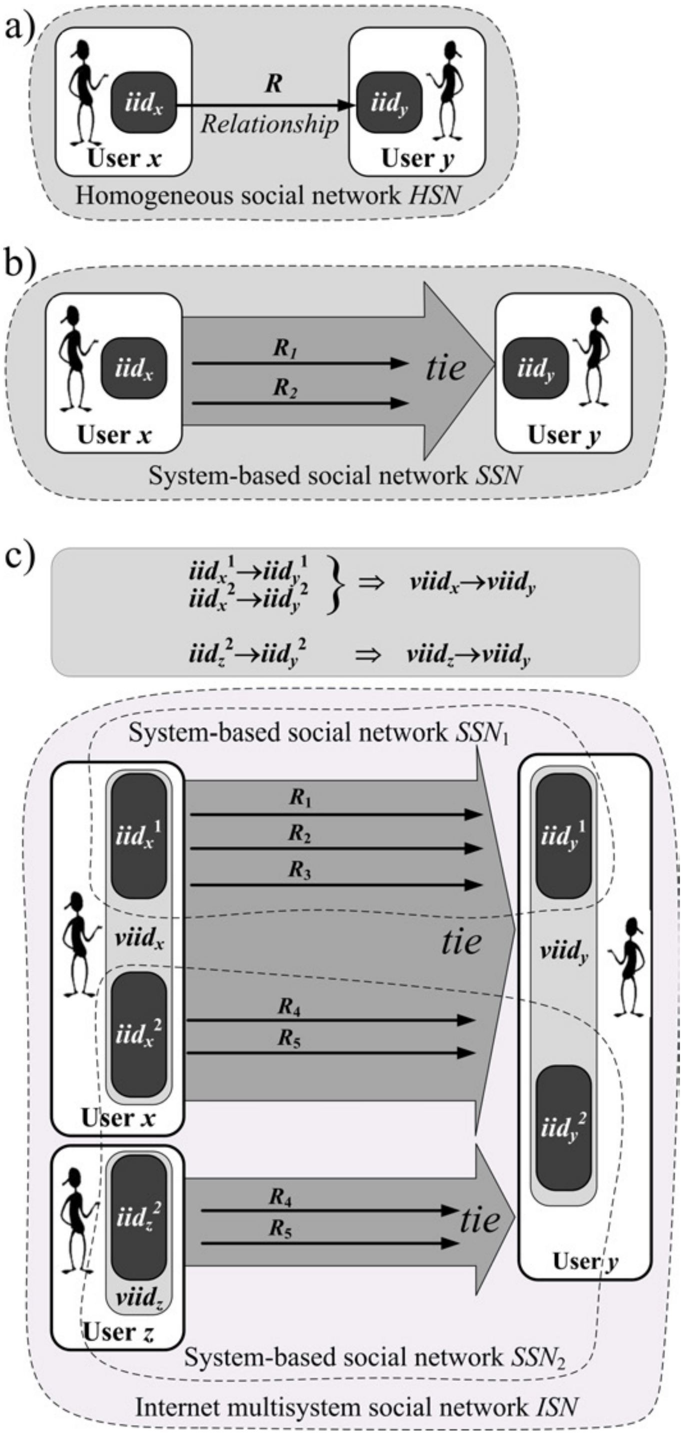

a)
iid_x
R
Relationship
iid_y
User x
User y
Homogeneous social network HSN
b)
iid_x
R_1
R_2
tie
iid_y
User x
User y
System-based social network SSN
c)
$iid_x^1 \rightarrow iid_y^1$
$iid_x^2 \rightarrow iid_y^2$
$\Rightarrow viid_x \rightarrow viid_y$
$iid_z^2 \rightarrow iid_y^2 \Rightarrow viid_z \rightarrow viid_y$
System-based social network SSN_1
iid_x^1
R_1
R_2
R_3
iid_y^1
$viid_x$
tie
$viid_y$
iid_x^2
R_4
R_5
User x
iid_y^2
iid_z^2
R_4
R_5
tie
$viid_z$
User z
User y
System-based social network SSN_2
Internet multisystem social network ISN

Figure 2.6: Examples of homogeneous *HSN*, system–based (*SSN*), and internet multimodal social network (*ISN*)

The single system on the Internet is the homogeneous system maintained or operated by the same subject (company, group of companies) usually using common interface and/or protocol. The main indicator of the system is the availability of data. From this point of view, two email systems operated by two separate companies for example Microsoft (Hotmail) and Google (Gmail) are two different systems on the Internet unless they exchange internal data about user communication. Thus, we are unable to create any relationship between two Gmail users based on their email exchange, having solely data from the Hotmail server. In consequence, two separate *HSN*s have to be created, one from the data available for Microsoft and one based on Google's records. On the other hand, a social network built upon the personal homepages connected with one another with the help of HTML hyperlinks can be treated as a single system since the information about mutual relationships is public even though it is scattered.

The same type of relationships means that two *HSN*s' members iid_i and iid_j share the same activity, e.g. they communicate with each other using emails or comment posts in the WordPress blogging system. Note that sometimes many different *HSN*s can be recognized within a single internet system. For example, based on the shared tagging lists to pictures, links to favorites and contact lists, three separate homogeneous social networks can be distinguished.

An internet identity *IID* is a digital, authenticable and permanent representation of a person, organization or organizational unit, group of people, or other social entity like family or group of interest (see Section 2.3). Some examples of internet identities are email addresses, logins to a specialized system such as blog logins (WordPress), instant messenger logins, an account's name in online social network system (Facebook, LinkedIn, Orkut, MySpace, etc.), logins to a multimedia sharing system (Flickr, YouTube) as well as URLs of personal homepages. On the other hand, a dynamic ID assigned to a single web session, email ID or ID of an anonymous user in e–commerce or search engine are not valid internet identities. Session or email IDs do not represent humans but their particular activities whereas anonymous or temporary users are neither persistent nor authenticable.

In general, internet relationships R can be either directed (as in Definition 2.2.1) or undirected. In the latter case, the definition would have to be modified: $R = \{iid_i, iid_j : iid_i \in IID, iid_j \in IID\}$. Besides, relationships can be either weighted ($R \rightarrow \Re$) or unweighted (binary) — all edges are considered equivalent.

Due to social character of *HSN*, it is usually reasonable to ensure only irreflexive relationships, i.e. $(iid_i, iid_j) \in R \Rightarrow i \neq j$. In other words, self-choices relationships [119] are not considered.

Definition 2.2.2 *A system–based social network SSN=(IID,T)is a multirelational network that consists of a finite set of internet identities IID and a finite set of system ties T linking pairs of internet identities. Set T is built from all relationships* $R_1, R_2, \cdots, R_N$ *existing in the system, where N is the number of homogeneous social networks HSN uncovered in the system ; i.e.* $T = \{(iid_i, iid_j, k_1, k_2, \cdots, k_N) : iid_i \in IID, iid_j \in IID, k_l = 1 \iff (iid_i, iid_j) \in R_l \text{ or } k_l = 0 \text{ otherwise}\}$.

The *SSN*s are also called multirelational (multilayered) social network. The examples of internet systems, from which multirelational social networks can be extracted, are: blog systems (WordPress, Blogger), multimedia sharing systems (Flickr, YouTube), complex instant messengers (Skype, ICQ), Gmail email system extended with personalized searching by utilizing Google search engine. Each of these internet systems contains one or more *HSN*s, which form single *SSN*s.

In WordPress, users can both maintain their blogs as well as tag them with the keywords that usually provide the information about the content of their diaries. Based on tags used and shared by users, we can create a homogeneous social network HSN_1. All people who exploit the same tags get into mutual relationships R_1. On the other hand, WordPress users can also maintain a blogroll i.e. a list of links to other blogs they like the most. These connections are the basis to create the second relationship type R_2 and another HSN_2 which can be called favorite–based homogeneous social network (Figure 2.5). Similarly to tags, the third R_3 and HSN_3 can be extracted from opinions that concern the same blog posts. People who comment the same blogs are in the mutual relationship and there is a high probability that users are interested in similar topics or prefer the same authors. Note that the relationships within different HSNs have different characteristics.

A homogeneous social network can in fact be simultaneously the system–based social network. Such case occurs when the set of ties T in the system–based social network is based only on one homogeneous relationships R_1, e.g. homogeneous network HSN_3 extracted from email communication is at the same time, the system–based (email–based) social network SSN_3 (Figure 2.5). However, it is valid only if we do not exploit relationships derived from contacts in address books. In yet another example, personal web pages connected with hyperlinks form both the homogeneous and system–based network of internet users.

Some separate system–based social networks can be merged in one complex internet multisystem social network.

Definition 2.2.3 *An internet multisystem social network ISN for the set of m system-based social networks* $SSN_i = (IID_i, T_i), i = 1, \cdots, m$ *is the tuple* $(VIID^M, T^M)$*, where* $VIID^M$ *is the set of virtual internet identities that merge all internet identities derived from all component system–based social networks* SSN_i *related to the same social entity . The set of ties* T^M *in turn aggregates relationships derived from the component* SSN_i*, i.e.* $T^M = \{(viid_x, viid_y, k_{11}, \cdots, k_{1N_1}, k_{21}, \cdots, k_{1N_2}, \cdots, k_{m1}, \cdots, k_{mN_m}) :$ $viid_x \in VIID^M, viid_y \in VIID^M, k_{ij} = 1 \iff (iidi_x, iidi_y) \in R_{ij}$*,* $viid_x$ *and* $viid_y$ *correspond to* $iidi_x$ *and* $iidi_y$ *in* SSN_i*, respectively, or* $k_{ij} = 0$ *otherwise*$\}$*,* R_{ij} *is the* j*th relationship set from all* N_i *relationships existing in* SSN_i*.*

An example of internet multisystem social network can be Blogger (SSN_{12} in Figure 2.5) that enables to log into the system using either its own user names (IID_1) or external Gmail accounts (IID_2). Since both systems have some common user identities it is possible to merge two system–based networks into one internet multisystem social network.

Internet multisystem social networks can be extracted from component system–based networks by merging their internet identity sets.

2.2.2 Taxonomy of the Networks of Internet Users

The aim of the previous section was to propose the definitions of network of users in the Internet whereas the goal of the following section is to classify the existing networks according to their features. These various kinds of social networks exist within the Internet. All of them can be named device supported social networks (*DSSN*) [120]. On contrary to "in person" social networks, which are much more tangible, *DSSN* suffer from limited social presence [120].

Social networks on the Internet can be divided into several groups using different criteria. They can be: dedicated *SN* (e.g. dating or business networks, networks of friends, graduates, fun clubs), indirect *SN* (instant messengers, address books, e–mails), common activities *SN* (e.g. co–authors of scientific papers, co–organizers of events), hyperlink networks (links between homepages), etc.

To put these different kinds of networks in order, they can be classified with respect to the following criteria:

1. The character of the relationship that connects two *IID*s (for more information see Section 2.4);
2. The type of the internet identities that build the social network (for more information see Section 2.3);
3. The type of the communication channel between members that is used to exchange resources;

4. Real time or non real time networks;

5. The type of the access to the network, (open/restricted access);

6. The level of the awareness of the members;

7. Dedicated– or common–service based networks.

In the classification based on the character of the relationship that connects two persons two basic kinds of relations can be distinguished: business and social connections (similarly to regular social networks). The former ones contain social networks that consist of people who are linked with each other due to common professional activities but simultaneously they do not share their private lives [37]. Those can be called professional networks. On the other hand, the social relationships indicate the connections with emotional background.

Moreover, the classification of social networks can be also made based on the types of internet identities that are the elements of the particular social networks. Overall, three types of such networks exist, i.e. these that consist of individual identities, group identities, or both of them. The most common are the networks containing mixture of both of these types of identities.

Another classification of networks of internet users can be based on the type of the communication channel between members that is used to exchange resources i.e. email, instant messengers, VoIP systems, video conferencing, etc.

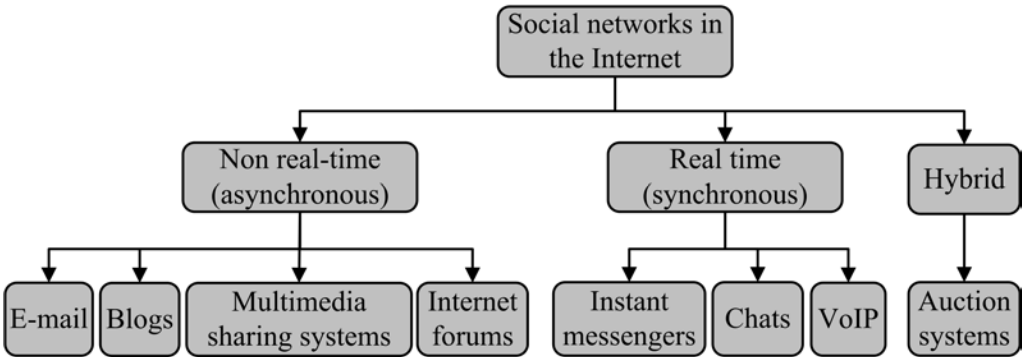


Figure 2.7: Real time vs. non real time networks

In general, the networks of Internet users can be divided into non–real time and real–time online social networks (Figure 2.7). The former enable asynchronous communication between two persons or from one person to a group of people [120]. Its example can be the electronic mail system. When person x sends an email to person y, the relationship between these people comes into existence. On contrary to the email system that supports the communication between either two persons or small selected groups of people, Internet forums, blogospheres and multimedia sharing systems enable all users from a given social network to read all messages submitted by every single member of the network. Their functionality is similar to bulletin board from the real world.

Chats, instant messengers, and VoIP systems create the second group of social networks that are supported by computer networks. Here the communication between users is synchronous, for example in an online chat (e.g. Internet Relay Chat) the user has to be online. Chats enable to submit messages that will be seen by all people who participate in it and who will have an opportunity to respond to these messages. The instant messengers (e.g. ICQ) serve to exchange information between two persons or limited group of people. The development of the Internet resulted in not only text messages being exchanged, but also voice and video streams. These media are used by VoIP systems, e.g. Skype or Ventrillo, increasing social presence [120].

There are also some hybrid systems that provide both synchronous and asynchronous communication like auction systems. Online users of such service can observe results of their activities immediately, but they can also be informed about other activities e.g. via email.

Another classification of social networks can be done based on the type of the access to the social network. The networks can be with an open or restricted access. In the former ones everybody can join them, e.g. Facebook, MySpace, ICQ, etc. while in the latter if one wants to become a member then somebody who has already been a member must invite this person, e.g. LinkedIn. There also exist networks with the restricted access, which are limited only to people who belong to the specific group or company.

The proposed above classifications are not the only possible ones. However, they highlight the fact that there exist many possible taxonomies of networks of Internet users.

2.3 Internet Identity — Node of the Network of Internet Users

Each network of Internet users consists of nodes — network members, called in this thesis internet identities, and relationships that connect these nodes. In this section the concept and types of internet identities will be presented as well as a proposal of *IID*s integration process is described.

2.3.1 Concept of the Internet Identities

Each concrete physical individual or a group of people who are the users of internet–based services can possess an internet identity. This internet identity (*iid*) is the short digital representation that has to fulfill several conditions. It must be verified, permanent, and authenticable. Moreover, internet identities are objects that can be unambiguously ascribed to one social entity, i.e. a person (individual identity), a group of people or an organization (group identity). Thus, the task of internet identity is to transfer the physical entity from the real to the virtual world (Figure 2.8). The concept of internet identities was considered in [104], [105], [75], [117]. Internet identity can also be called online identity [38], [64], [124]. However, it suggests that online identities are restricted to only online, synchronized services and for example email addresses could not be covered by online identities.

This mapping enables to define the connections between social entities based on the connections between their internet identities (see Section 2.4). Since we are not able to study relationships between physical social entities in the Internet, the only possible social network analysis is the analysis of internet identities.

Definition 2.3.1 *An internet identity iid is a short digital, authenticable, unambiguous and permanent representation of a physical social entity — a concrete human or a group of people, who are conscious users of the given internet–based system.*

Based on the conducted research on the existing on the Internet social networks seven basic features of internet identities can be enumerated:

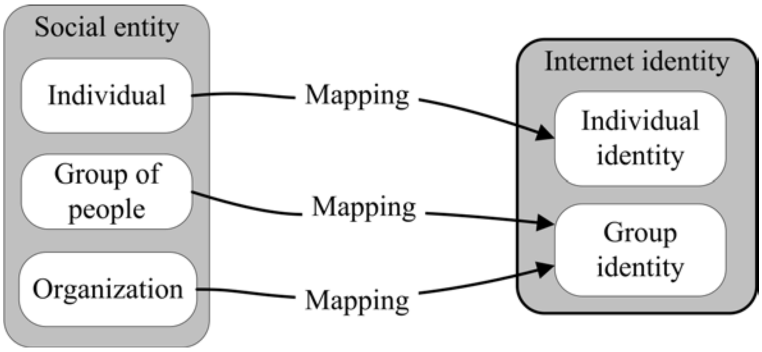


Figure 2.8: Mapping of social entities into internet identities

1. Succinctness;

2. Authentication;

3. Uniqueness;

4. Durability;

5. User's awareness;

6. Correspondence to concrete humans;

7. Extraction from the Internet services.

An internet identity is a short digital representation of physical entity. Hence, concatenation of the name and the postal address does not fulfill this condition — it is too verbose. Moreover, only authenticable, verifiable users are considered, so they at least have to be registered in the system. No other action is necessary and the users do not have to use the service any more. For instance, one can register in the e–commerce system and get internet identity and after that never utilize this account to buy any products. Nevertheless, due to a lack of relationships, such internet identity would probably be isolated in a network. There is another similar example: people who send emails to the new, just registered user x, automatically get into relationship with x. Although this new user x may not have read these emails and not sent any emails yet, x possesses his/her own internet identity (the registered email address) and even some relationships with the email senders; everything with no x's involvement since registration.

The registration to the service must be done knowingly. Thus, users created by the system administrator should not be considered as the members of the social network unless they are aware of their registration. It may happen that fulfilling of this requirement is hard to achieve and we would need to assume, especially during automatic data processing, that all registered accounts are valid internet identities.

Uniqueness of *iid* has to be ensured by the system itself. There should not be two identical email addresses on the Internet or two identical user names in the blogging system.

Furthermore, the internet identity must not be temporary. For instance, it cannot be dedicated only to one single user visit in the system and different from other sessions of the same user.

Several typical examples of internet identities can be mentioned:

- Email address,
- Login to social network sites (Facebook, Friendster, LinkedIn, Orkut, MySpace, Classmates),
- Login, identifier, nickname or user name in a specialized system. In this case, *iid* is usually a tuple (login, system):
  - Registered user name in an online blogging system (WordPress, Blogger),
  - Instant messenger or VoIP communicator nickname (Skype, ICQ, MSN, AIM, Yahoo! Messenger, GTalk),
  - Login to multimedia sharing systems (Flickr, YouTube),
  - Login to social services like social bookmarking (del.icio.us), social travel network (TripUp), social searching (Technorati),
  - Account in an e–commerce (Amazon, iTunes Store),
  - User name in an auction system (eBay),

  - Login to a web–based financial service (PayPal, WebMoney, ebanks, ebrokers),
  - Registered user in a personalized web portal, especially news service (My Yahoo!, CNN), online journal (The New York Times),
  - Account in a specialized service available on the Internet, for example: online library: ACM or IEEE Web Account with access to ACM Digital Library or IEEE Computer Society Digital Library respectively,
  - X.509 certificates used to authenticate SSL clients while logging into web sites with restricted access,

- URL to the personal home web page,
- Login to a comprehensive identity system (OpenID [105]).

There are also some examples that are NOT the internet identities:

- ID of a single web session – it corresponds to the activities of humans rather than the social entity itself and it is temporal,
- ID of searching session, ditto,
- ID assigned to the exchanged objects, e.g. email ID, ditto
- First and second name of an individual published on their personal web page as it can be ambiguous,
- Temporal ID assigned to an anonymous user in an e–commerce system, usually used only for one visit as it is neither authenticable nor permanent,
- Anonymous commentator of posts in blogging system, ditto,
- Company profile published in the web site as it is not a short digital representation,
- Postal address published on the contact web page, ditto,
- X.509 certificate or its serial number issued to an SLL web server as the server is not a conscious user,
- Authors of scientific papers gathered in the online bibliographical DBLP database[1]. Although, it contains data about co-authorship and in consequence their mutual relationships, the authors' names do not reflect internet identities. Besides, the authors are not conscious user of any internet service,

[1] http://www.informatik.uni-trier.de/ ley/db/

- Guest account in an internet service, e.g. one "student" account common for all anonymous users in an online e-learning system as it does not correspond to a tangible social entity,
- Anonymous account to FTP servers as it is not authenticable,
- Trial account in an internet service unless the trial period is long enough as it is not permanent,
- Accounts transferred from another system by a system provider without user awareness as these new users are not aware of this change unless they accept this operation.

Nevertheless, the thing to remember is that people try to be as anonymous as possible on the Internet. This is often the reason why people have multiple internet identities. Additionally, people may want to separate their private and corporate activities (profiles) [36]. As a result, one physical social entity can possess many internet identities in one system. For example user *z* possesses one account in the blogging system (iid_5) and two separate email accounts (iid_4 and iid_7) as illustrated in Figure 2.9. All these *z*'s internet identities can be merged into virtual identity that represents all internet identities of one social entity: virtual *ID z* aggregates iid_4, iid_7, and iid_5. On the other hand, one internet identity is connected with only one social entity. In other words, the only restriction for the internet identity is that it has to refer to exactly one physical social entity — an individual or a group of people.

Definition 2.3.2 *Virtual internet identities aggregate distributed internet identities exisiting in different internet–based systems. A virtual internet identity viid corresponds to all internet identities iid related to a single physical social entity. Simultanously, each internet identity is related to only one virtual identity.*

Note that some users of internet services may correspond to the same social entity in the real world, e.g. users *u* and *z* refer in fact to the same single person denoted *u–z* (Figure 2.9). In some cases, we are able to identify that two different virtual identities belong to one physical entity, e.g. based on the data provided by users in their registration forms. Then, we can join virtual *ID z* and virtual *ID u* into another combined virtual *ID u–z*. The consequence of this kind of merging is the removal of the data about the reciprocal communication between the identities that are merged into one account. Note that this internal communication usually results from the way in which people organize their contacts with others. For example, one can posses two different email accounts — one for private communication and one for professional contacts but emails sent to the private account are usually forwarded to the company mailbox. A similar situation can also occur when

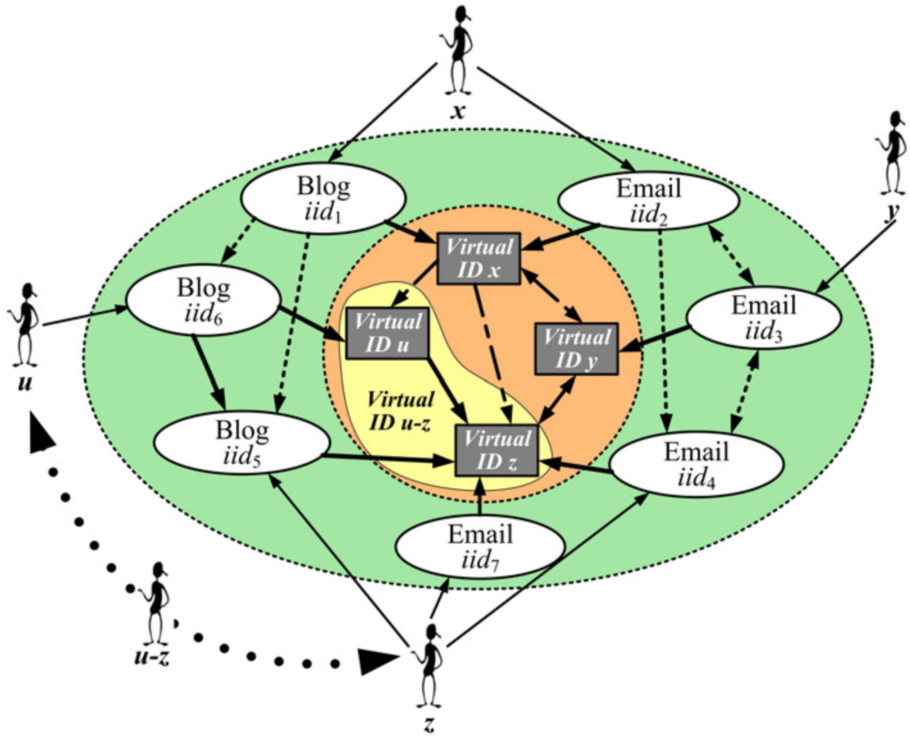


Figure 2.9: The concept of internet identities merging

the person makes a mistake while using the specific internet service, e.g. one registers to the system many times because he/she forgets the password or login.

In practice, it is usually difficult to obtain virtual identities, i.e. merge internet identities related to the same person in an automatic way. However there are some specialized systems like OpenID or eBuddy that enable to achieve it with the assistance of users themselves.

2.3.2 Individual and Group Internet Identity

An internet identity is the user identifier valid for one or more internet–based services that unambiguously distinguishes users of these services (Definition 2.3.1). There are either individual or group internet identities (Figure 2.8). An individual internet identity belongs to an individual — a single person, whereas a group identity corresponds to a group of people, e.g. a family that uses only one login to the blog or to an organization or all employees of the service department who use one common email account service@company.com to answer customers' requests.

Group identities can by identified by content analyses. If we study the signatures in the emails and we recognize more than one name there then it would mean that more than one person sends these emails. Moreover, sometimes the name of the internet identity can be directly matched with the name of the company or its department.

The interaction between group identities reflects the relations between two groups of people, e.g. two companies, two departments within one organization or two families. On contrary to the individual identities, the group identities are not restricted by social limits of single humans. According to Dunbar's studies, the maximum number of steady relationships that one can effectively maintain is about 150; it is also called the Dunbar's number [60], [39].

Furthermore, the behavior of people represented by group identities seems to be more stable over time than individual ones, e.g. when an individual goes on leave then the account is usually not used during this time whereas in the case of group identity even if some members are currently not available then the others take these users' duties over. Of course, it depends on the number of people who use this account as well as the type of the group identity. Probably, the greater the number of real, social entities related to a single group identity the more stable the behavior of this identity is. For instance, the general company email account used to contact its clients is likely to be steadier than the identity used by a single family.

Several different types of individual as well as group identities can be identified. The individual identities examples include:

- Private identity, e.g. instant messenger nickname to private account, private email address;
- Professional identity;
- Activity/interest–based identity — the login to the fanclub site;
- Consumer identity — login to the customer account to the web site of the telecommunication company or e–commerce;

The following types of group identities can be distinguished:

- Interest-based identity — special interest groups;
- Family–based identity — a wife and husband can use the same account in the e–banking system;
- Task–based — the common account for the project team;
- Position–based identity — many people who occupy the same position share the common account, e.g. all PhD students at the university use the same login to the academic intranet;
- Company–based identity — the homepage where the company provides the information about itself;
- Unit–based identity — the email address of the individual department in the company.

Only one kind of internet identity is usually retained in the single internet service or at least one kind significantly dominates. For that reason, most system–based social networks contain the generally homogeneous sets of *IID*. Note that some people can simultaneously maintain two or more types of identities. If we have enough information these separate identities can be merged into one virtual internet identity (Figure 2.9, users *u* and *z*). On the other hand, one internet identity can capture several types of identities corresponding to various activities of the person. For example, some people use one common email address in both private and professional life (one *iid* of two types) whereas the others utilize separate addresses for both these involvements (two *iid*s for one person). The same may be valid for social networking sites [36].

2.3.3 Internet Identities Integration

As it was presented in Seciton 2.3.2, various kinds of internet identities can be distinguished. Nevertheless, nowadays, it becomes more and more popular to merge two or more internet identities into single one in order to enable people to access different services with only one login. Thus, the single sign–on concept (*SSO*) extends also to the Internet. It is achieved by internal integration of two or more services delivered by the single provider or even the cooperation between independent providers.

The example of such integration can be found within Google services. The single email address (iid) enables the user to login into both the blog service — Blogger and the email service — Gmail. Of course, it is also possible for a person who does not use Gmail service to maintain a separate blog account.

Another integration system — OpenID allows to create a single common account that facilitates to login to nearly ten–thousand websites with this identity. It eliminates the necessity for creation of multiple usernames across different websites. OpenID concept is used among others within FOAF format to identify internet users

Yet another example can be eBuddy that is a free web–based messenger. This system provides the interface and engine which supports the communication via many other services including Windows Live Messenger, Yahoo, MySpace, Google Talk (GTalk), and others. Hence, it integrates many internet identities derived from separate systems into one eBuddy *ID*.

Generally, two or more social networks can be integrated based on matching and merging the internet identities existing within all of them. To achieve it, we ought to possess or gain the knowledge about real users and their internet identities within merged networks that are being integrated. For instance, if two system–based social networks, e.g. VoIP–based social network (1) and the network derived from personal homepages (2) (Figure 2.10) are supposed to be merged then for each social entity the set of the internet identities that a given person possesses in both networks need to be identified (user *a* has both homepage address and login to Skype system, whereas user *u* is only

Skype user). Thus, we are able to discover internet identities of the same

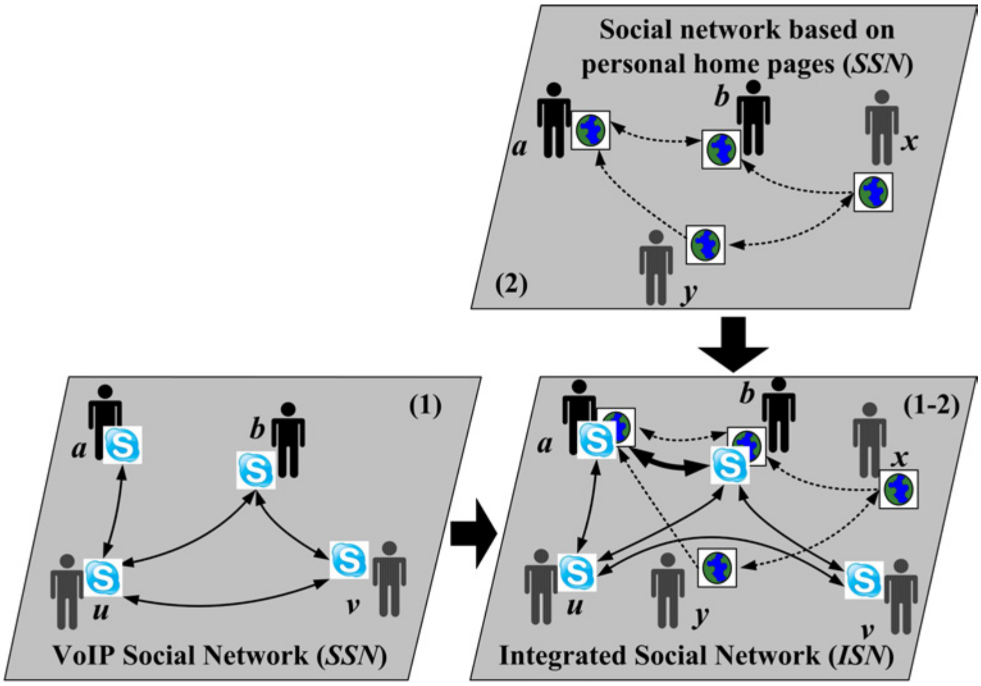

Social network based on personal home pages (SSN)
a
b
x
y
(2)
(1)
a
b
u
v
VoIP Social Network (SSN)
(1-2)
a
b
x
u
y
v
Integrated Social Network (ISN)

Figure 2.10: Integration of two system–based social networks by means of internet identity merging

users in both networks using our external knowledge (e.g. data from the paper contact), matching mechanisms (e.g. by email address) as well as information provided directly by network members (users of the VoIP network can deliver URLs to their homepages at registration time to this system or publish their account name to the VoIP system on their homepages). Additionally, the relationships between users from both systems can be utilized in the final, integrated social network (*ISN*) as (1-2) in Figure 2.10. The integration can provide additional extension possibilities for the merged networks. For example, users *a* and *b* in the VoIP system can be suggested and encouraged to communicate with each other based on the hyperlinks connecting their homepages; the thick solid arrow between *a* and *b* in the network 1-2, in Figure 2.10. A similar mechanism of merging two networks: a telecom social network and an internet–based network can be used by the telecommunication company to create an additional service for its customers: "call the acquaintances you do not talk to". In this case, the recommended people would be extracted from the internet–based social network.

The integration can also be performed based on the user profile matching. For example, if two internet identities have in the demographic profile the same name and address then there is a high probability that they both belong to one social entity.

Overall, integration of the internet identities may be a new trend within the web service development.

2.4 Internet Relationships — Edges of the Network of Internet Users

Apart from internet identities, the second crucial component of every network of Interent users are their relationships represented by links connecting pairs of nodes. The concept and types of both internet relationships and compound ties are presented in this section.

2.4.1 Concept of the Internet Relationship

Definition 2.4.1 *An internet relationship r, in the homogenous social network* $HSN = (IID, R)$ *is the directed connection* $r \in R$ *from one internet identity* iid_i *to another* iid_j*. Both internet identities* iid_i *and* iid_j *are of the same type, i.e.* $iid_i \in IID$ *and* $iid_j \in IID$*.*

Note that in the system–based social network *SSN*=(*IID*, *R*) a single tie $t \in T$ may contain up to N internet–based relationships described in the Definition 2.2.2.

A relationship in the social network is the connection from one member to another that reflects their acquaintance, private or professional relation or even high similarity of their inclinations or activities. The maintenance or even only creation of the relationship usually requires member's trust, commitment, emotion, or dedication of time and effort.

Several significant social properties can characterize a human relationship, in particular [55], [119]:

- Mutuality;
- Durability;
- Intensity;
- Intentions;
- Culture conditionings;
- Emotional level;
- Strength.

A relationship does not have to be symmetrical, e.g. Tom could be friend of John but John might not see Tom as his friend. Nevertheless, if a relationship is symmetrical then it is usually more durable. Moreover, a relationship may be durable for a certain period; afterwards it could significantly weaken or even diminish. Thus, a relationship is either more persistent or more temporal and the time factor emerges to be very important. If Tom sent John

20 emails over two weeks, but five years ago, then John would have most probably forgotten Tom by now. However, John would remember and feel a kind of durable relationship with Bill who has regularly sent John one email every quarter for the last five years. The number of emails is the same in both cases (20) but the latter appears to be much stronger right now. Each human relationship requires periodic support and refreshment. Furthermore, the longer the acquaintance lasts the more durable it is likely to be in the future.

The importance of contact intensity and communication features on the strength of the relationship may result from the culture both participants live in. Ten emails sent by people from one country may have greater significance than the same number of emails exchanged between individuals from another, more spontaneous nations. Many phone calls made late at night or in one's time off reflect a stronger relationship than the same calls made in regular working hours.

The strength of a relationship can depend on its basis, especially the type of communication or mutual activity. The meeting of commentators of the same blog or even hyperlinks between homepages generally connect people much less than the co–authorship of a scientific paper.

Some unusual factors may also be the sign of stronger relationships. An intensive correspondence in Polish is the evidence for stronger relationship between foreigners in Japan rather than the same communication in Japanese between natives. Nevertheless, the opposite meaning would be true but in Poland.

In some environments like the worldwide Internet, that is multicultural in its nature, the detection of some differences can be very difficult. Moreover, some features of human relationships may either require complicated content processing like extraction of the emotion level or even be very hard to discover like in the case of intentions.

Note that Definition 2.4.1 assumes that a single relationship binds only two internet identities. In more general approach, we can use hyperedges and a hypergraph as the representation of the social network [9]. A hyperedge connects any number of network nodes (but at least two). This can be useful especially in case of relationships derived from common activities or interest as well as based on profile matching. Comments on a single blog post involve all participating commentators; single interest can be simultaneously shared by many people; many members can have profiles similar to each other, etc.

2.4.2 Types of the Internet Relationships

The relationships existing in the Internet can be classified in many different ways and based on differen characteristics (Fig. 2.11):

- Active subject that is responsible for creation of new relationships (user, system);

- Awareness of the users that they are involved in relationships;
- Mutuality of the connection between users (asymmetrical, symmetrical, reflexive);
- General relationship sources (external world, the Internet);
- Data type used by the system for relationship creation (communication, common activity, user profiles, direct connections);
- Nature of relationships (professional, family, friendship, acquaintance, common interest, customer-based);
- Directness of relationship grounds (direct, quasi-direct, indirect), see Section 2.4.3;
- Visibility of relationships for the users.

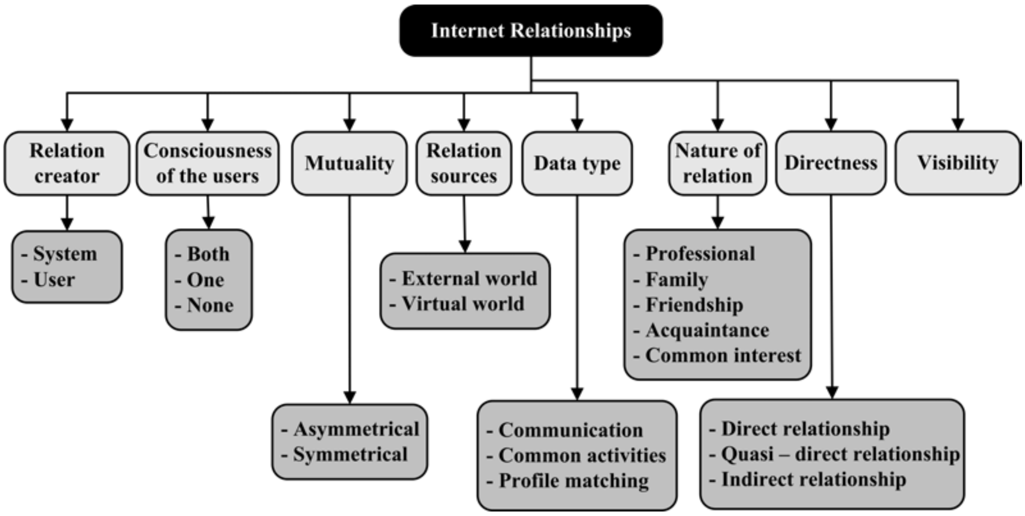


Figure 2.11: Taxonomy of internet relationships

In the first classification, the relationships can be divided in three main types: (i) created by the users, (ii) established by the system and (iii) the mixture of the two. In the first type, user x can set up a relationship with another person y by adding y's email address to x's private contact list or linking to y's homepage at x's private page. These kinds of relations are directed and the person y whose iid_y is added to x's contact list does not have to be aware of this fact. However, also the situation, in which both sides are aware of the relationship creation can appear. For example a new connection is established when two people exchange emails or one of the users sends an invitation to another within the social networking site (like Friendster, MySpace, or LinkedIn) and the other person accepts this invitation. Nevertheless, the relations can be initiated and created also by the system itself, for example when the profile matching is performed and in such a situation none of the users is aware of the just established connection. The last but not least common situation is when in the process of relationship creation both a system and a user are involved. For instance when the system recommends other users to the specific one then it initiates the relation but the user has to confirm that he/she is interested in such a relation by approving the suggestion generated by the system. Only when the user accepts the recommendation the connection is created.

When the awareness of the users that are involved in the relationship is considered then three kinds of connections can be distinguished. The first type occurs, when both internet identities participating in the relation are aware of this fact, e.g. two users communicating with the instant messenger or exchanging emails. The second situation happens when only one side of the relationship is aware. The example for this can be adding by single user x another person y's email address to x's private contact list or link to y's homepage at x's web page. Person y is usually not aware of these user's x activities. In the third type, we have relationships in which none of the participants is conscious of the connection existing between them, e.g. when the relationship is created by the system based on the profile matching also known as demographic filtering.

The next features of internet relationships is the direction and mutuality of the connection between users (Figure 2.12).

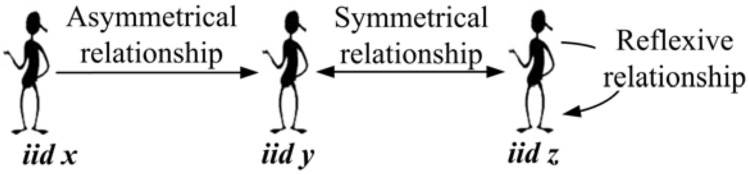


Figure 2.12: The direction of relationships [33]

The relationship can be asymmetrical, i.e. internet identity iid_x is in the relationship with internet identity iid_y but there is no reverse connection from iid_y to iid_x (Figure 2.12). The example of such relationship can be: if user x adds user's y blog to the favorites ones but user y does not do the same. On the other hand the symmetrical relationships exist when there is a mutual communication between users or when people share common activities, e.g. exchange emails or comment the same photo in the multimedia sharing system such as Flickr. Due to social and collective profile of social networks all reflexive relationships are usually excluded from consideration.

The connection between two internet identities can be also investigated based on its source, i.e. where does it origin from? The acquaintance can come from the external world, e.g. two network members know each other personally and they have exchanged their email addresses. When they start sending emails to one another then the relationship is set up in the virtual world. However, a relationship can also exist only in the virtual world. This situation appears, e.g. when one user sends the invitation to another one within the social networking site such as MySpace and additionally these users previously did not know each other in the real world. Finally, there are also relationships, in which two people do not know each other and the system itself creates the connection between these users based on the profile matching.

Another approach to relations classification is to split them according to the type of data used by the system for relationship creation. In consequence, an acquaintance can be created based on the data about mutual communication (email exchange), common activities (commenting the same multimedia objects, using the same commercial internet service), data derived from the profile matching, or data from users' contact lists (e.g. contact lists from instant messengers).

One of the most interesting taxonomies is the classification of the relationships according to their nature. Hence, among many types of relationships the following can be distinguished: professional, family, friendship, acquaintance, common interest, customer relationship (online consulting, e–learning, usage of specific internet service or its features), etc. Nevertheless, the process of specifying the character of the relation is a very complex task because it is hard to identify in the virtual world what kind of relation exists between two users unless they state openly the character of their connection. Another method that can serve to recognize the character of the relationships is the investigation of the parameters of the communication between two users (in particular its time and frequency) or common activities.

The relationships can be also classified on the basis of their visibility for other users. It especially concerns the social networking sites like Facebook or MySpace where people can directly define who can browse their profiles and relations. The number and specification of the visibility levels depend on the system, e.g. in Friendster, users decide whether their relationships can be viewed either only by the nearest friends, also by friends of a friend or maybe by whole community.

2.4.3 Directness of the Internet Relationship

There is also another, specific taxonomy of relationships with respect to their directness. We can distinguish three kinds of relationships: direct, quasi–direct, and indirect relations.

Social entities related to the internet identities can be more or less aware of the relationships they are involved in and this partly depends on the basis where relationships are derived from. For that reason, three kinds of internet relationships can be enumerated:

- Direct relationship — is a relationship that connects two internet identities with a direct connector, Figure 2.13. The direct connector is an object that is addressed to the specific internet identity and is usually related to the specific feature (communication, activity) existing in the system. For example, an internet identity establishes and supports a direct relationship while sending an email to another internet identity. Thus, the direct connector can be derived from an email, a phone call (VoIP), message sent by means of instant messengers, hyperlink binding one home web page with another one, an item in somebody's contact list, a connection in the social networking site.

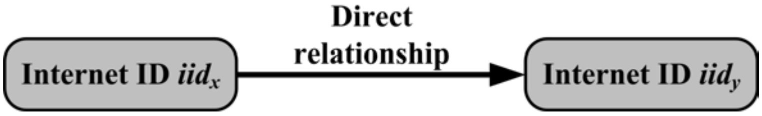


Figure 2.13: The direct internet relationship in the social network on the Internet

- Quasi–direct relationship — two internet identities are in the relationship but it is not required that they maintain the relationship themselves, e.g. people who comment on the same blog or participate in the common business meeting. There is always a meeting object, which serves as the communication medium between users, Figure 2.14. The roles of both internet identities, which are in this kind of relationship, in relation to the meeting object can be either the same or different.
  - Quasi–direct relation with equal roles r_{xy} means that internet identities iid_x and iid_y meet each other through the meeting object and their role in relation to this object is the same. In other words, they participate in common activity related to a certain object with the same role a, e.g. two users comment the same picture, both of them add the same object to their favorites or both use the same tags as metadata to describe their photos (Figure 2.15a).

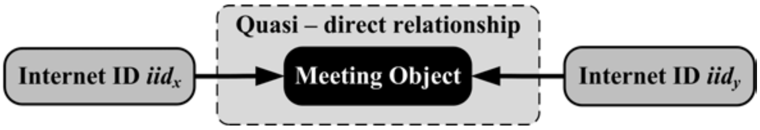


Figure 2.14: The quasi–direct internet realtionship in the social network on the Internet

  - Quasi–direct relation with different roles r^{ab}_{xy}, r^{ba}_{yx} — is the relation between two internet identities iid_x and iid_y that are connected through the meeting object (multimedia object or its additional features like tag) in the way they participate in common activity but their roles a and b towards the meeting object are different, e.g. iid_x comments a photo (role a — commentator) that was published by iid_y (role b — author) (Figure 2.15b). The non zero relation r^{ab}_{xy} entails the non zero relation r^{ba}_{yx}.

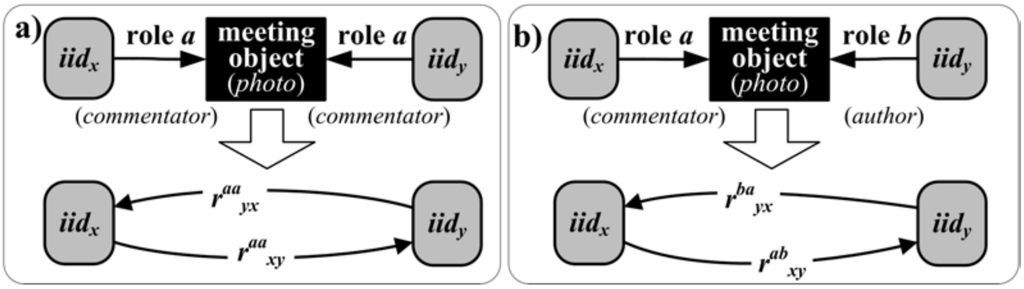


Figure 2.15: The object–based internet relationship with equal roles: commentator (a), and different roles: commentator and author (b)

- Indirect relationship — this kind of relationship exists when the internet identity is not aware of the fact that it is similar to another internet identity. Two internet identities are connected by indirect link when their profiles are similar (Figure 2.16). If these relationships are discovered and analyzed in a right manner then such knowledge can be used to change the hidden relationships into direct ones.

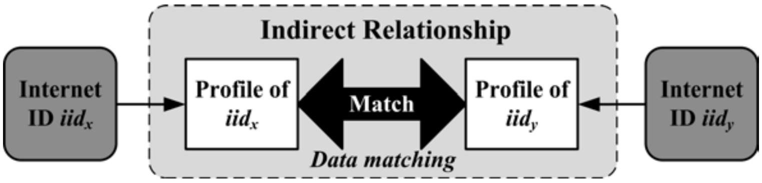


Figure 2.16: The indirect internet relationship in the social network on the Internet

It is worth noticing that the direct relationships can be supported and developed by utilizing the knowledge derived from the characteristic of indirect relationships, e.g. the recommendation systems can use the demographic filtering to suggest movies liked by people with a similar taste and/or with a similar profile.

2.4.4 Ties

The issue that is tightly connected with the concept of a relationship is a tie. A tie is the set of all relationships that exist between two internet identities. In other words, a tie between two internet identities aggregates all types of the relationships that exist between these two internet identities. The types of relationships, which create a single tie, can reflect different communication channels used to exchange information. For instance, two users who send emails to each other, use SMS and VoIP services for mutual communication maintain three types of relationships. In such case, the tie that exists between them consists of three separate relationships (Figure 2.17). An analogous situation appears when the complex character of mutual relationship is analyzed. Two people can be in the relationship of a friendship and in the same time be co–workers. In consequence, they maintain two types of connections and the set of them is called a tie.

Note that in a *HSN* a tie is the synonym of the relationship, because there is only one kind of the relationship in *HSN* allowed. The different types of relationships (e.g. friendship, family, professional, etc.) can be grouped into layers. A layer of the social network is, in fact, the single *HSN*. During research that was conducted on Flickr dataset [97] nine types of relationships were identified: relations created based on contact lists, tags used by more than one user, user groups, multimedia objects (pictures) added by users to their favourites, and opinions about pictures created by others. Relations based on contact lists represent direct intentional relations. Tag–based, group–based, favourite–favourite, and opinion–opinion relations are instances of object–based relations with equal roles, whereas favourite–author, author–favourite, opinion–author, and author–opinion are object–based relations with different roles. All these relations formed the basis for creation of nine separate layers in the social network. In consequence, each of the layers creates the separate *HSN*.

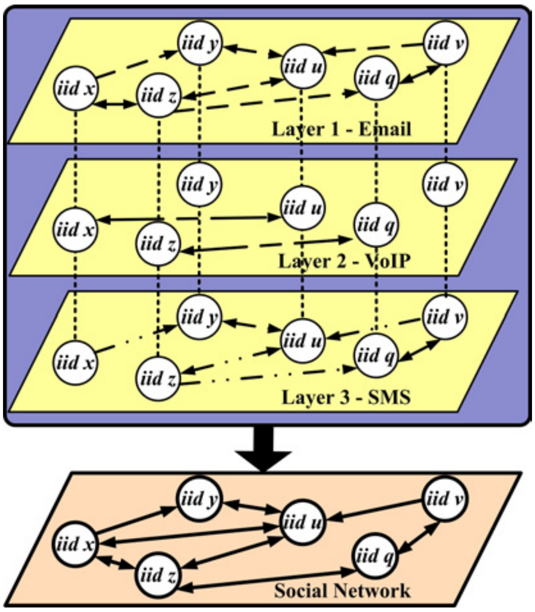


Figure 2.17: The tie concept in the social network

2.5 Examples of the Networks of Internet Users

In order to present the variety of social networks that exist within the Internet some examples of them are presented and compared with one another below.

2.5.1 Electronic Mail Services

Email systems are the bidirectional and asynchronous way of communication, which enable their users, who are in different places and on different schedules, to communicate with one another by exchanging messages [120]. This is the basis to form a social network, in which email addresses represent physical social entities. The email addresses are internet identities whereas the relationship in an email–based social network can be derived both from the communication from senders and recipients (two internet identities) as well as from the address books maintained by users. The registered email addresses and information about communication (logs of SMTP servers) as well as some information about private address books stored on the server can be acquired from separate, distributed mail servers, e.g. Gmail, Yahoo! Mail, MSN Hotmail, AOL Mail, etc. On the other hand, many email users utilize their own local email transfer agents (MTA) and maintain their address books only on the local storage. Obviously, this data is unavailable for external processing. Address books and communication (logs of exchanged emails) are two main sources to create, analyze and explore email–based human relationships. They can be treated either as the components of separate layers within the *SSN* or as a part of one coherent *HSN*. Since address books are often hard to obtain, the second approach is more common.

2.5.2 Instant Messengers

The instant messengers (IM) such as ICQ, Skype, Windows Live Messenger (former MSN Messenger), AOL Instant Messenger (AIM), Yahoo! Messenger, Google Talk (GTalk) serve to exchange information between two persons or limited group of people. The rapid development of high speed internet connections resulted in not only text messages but also voice and video streams being easily transferred online. These features are incorporated into many VoIP systems, e.g. Skype or Ventrillo, increasing social presence [120]. Nowadays, most of the instant messengers support also other kinds of communication channels. However, their primary goals are quite precisely defined – exchange textual information. The communication within instant messengers is synchronous. In contrary to email systems, an instant messenger provides easier way of collaboration because it offers a real time communication. Additionally, it is usually visible for the user whether other people from their contact list are available or not because there is the possibility to see the user status, e.g. online, away, not available. Since most data related to individual users is stored locally on their computers, the acquisition of communication data necessary to build the social network from instant messengers may be very difficult. Nevertheless, some IM operators provide the opportunity to transfer and retain some local data on the central server.

2.5.3 Blogs Services

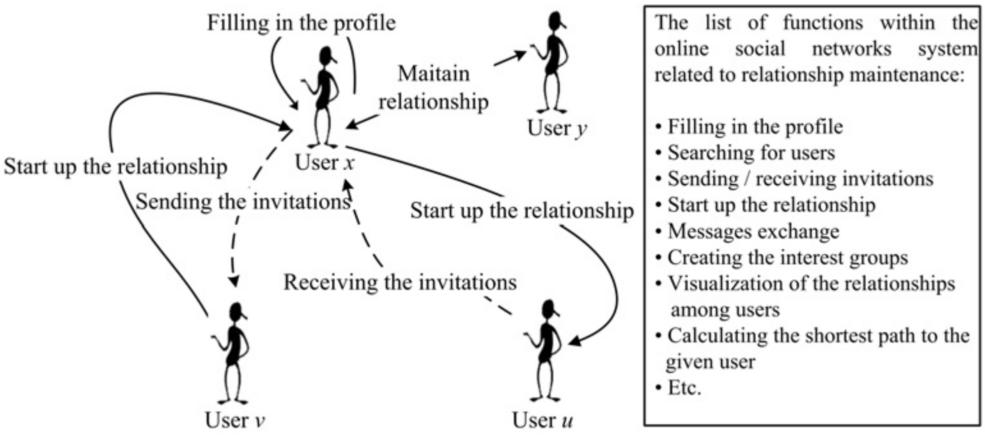
The blogs services like WordPress, Blogger, LiveJournal or Windows Live Spaces are not only the online diaries but they can also be treated as the system–based social network. In this case the login to the system is the single internet identity and the relation between two *iid*s can be created based on the list of favorites, tags commonly used, or comments made on the blog and additionally all of these connections are quasi–direct relationships. In the first situation, the favorite–favorite relationship denotes the connection between two people who added the same blog to their favorites, whereas relationships of the type favorite–author and author–favorite reflect the acquaintance between the person who has added the blog to their favorites and the author of this blog. Similar distinction can be made in the case of the relationships based on comments added to posts on a particular blog. Tag–based relationship exists if there are two users that have used the same tags to describe the content of their blogs.

2.5.4 Social Networking Sites

In the past few years the popularity of social networking sites (*SNS*) [30], [17], [53], [20], [99] has rapidly increased. They can also be called virtual communities, social network services [30], online social networks [53], online networking sites [53], social web sites [59] or social networking portals [99]. Their main goal is to create, maintain and present the social network to their users as well as match their users with each other. To achieve it, they make use of some additional communication services like emails, chats, instant messaging. Recently, the concept of social networking sites together with publishing and blog services has been commonly named as social networking [37], [122] and the common term for the systems is social websites [76].

Typical examples of social networking sites are: Facebook [29], [36], [78], Friendster [16], MySpace [28], Orkut, Tribe, Ecademy or LinkedIn [34], [82]. They are created and maintained by commercial companies. The main features of social networking sites are: self–expression (maintenance of personal profiles), including presentation of personal achievements, striking up relationships with others and mutual communication. There are several ways of communication between users within these online networking sites, which vary depending on the functionality of the portal: email, chat, forum, blog, comments, testimonials, photo/movie album, etc. The more communication channels are served by the network the better. It provides wider opportunity to create new and maintain existed relationships within the system.

In order to understand the concept of *SNS*, their main functions are described. Usually, at the registration stage, each user should fill in the profile (Fig. 2.18) that contains for example their demographic data, information about their hobbies, professional experience and general profile of people that they are interested in. After that, user x sends or receives invitations from other users of the network. If either user x replies to invitations or other participants reply to users' x's invitations, then the relationship is established (Fig. 2.18). However, not only the initiation of the relationship is important, but also its maintenance is one of the crucial parts of every *SNS*.



The list of functions within the online social networks system related to relationship maintenance:

- Filling in the profile
- Searching for users
- Sending / receiving invitations
- Start up the relationship
- Messages exchange
- Creating the interest groups
- Visualization of the relationships among users
- Calculating the shortest path to the given user
- Etc.

Figure 2.18: Main functions of the social networking site related to relationship maintenance

The deeper analyses of these kind of system was made e.g. in [17] and [99]. In the former one the definition and history of *SNS* was presented whereas in the latter one the authors classified the *SNS* according to the following criteria: whether they are registration or connection based; whether user profiles are social or professionally oriented as well as if explicit relationships can be defined; and if sites are not–for–profit or profit–based. The sites that were compared in [99] were: Orkut, Friendster, Tribe, Tickle, LinkedIn, Spoke, Ecademy, Ryze, Meetup.

2.5.5 Multimedia Sharing Systems

Systems like Break.com, Google Video, Metacafe, OneWorldTV enable a user to upload and manage their own multimedia contents such as photos, videos, animations which are commonly called multimedia objects (MOs). Each of the multimedia objects can be tagged by the author. In other words, a user can describe their MOs with one or more short phrases that usually denote the content of this element. These tags used by the members can be the basis for creation of a social network based on tagging, in which a relationship between two members exists if both of them have used at least one common tag to describe their multimedia objects [89]. Simultaneously, users have the opportunity to interact, collaborate and influence one another in different ways. Hence, they can not only tag the items they have published but also comment MOs added by others, include them to their favourites, etc. Additionally, users have the opportunity to set up new, direct relationships with other system users.

People who cooperate with one another or share common activities via publishing system can be seen as a specific social community. The members of this community, represented by their internet identities, can be related either directly or indirectly. Direct relationships are derived from connections explicitly provided by users who, for example, place other users into their contact lists. Nevertheless, two or more internet identities can also be related indirectly through an external object like a group or tag they share or an item they both comment. Users even do not need to be aware of the indirect relationships they are involved in.

The Flickr system is an example of such a multimedia sharing system. In this photo sharing system, nine miscellaneous relationship layers can be identified from the data about user activities, i.e. contact lists, tags, groups of items, favourite pictures, and comments to photos [71], [72]. Some of them like favourites and opinions were split into three separate layers, e.g. author—commentator, commentator-–author, commentator-–commentator. Flickr with its layers is a typical *SSN*, where a single set *IID* of internet identities exists and nine different types of relationships can be distinguished: R^c — contact based, R^t — tag–based, R^g — group–based, R^{ff} — favorite-–favorite, R^{fa} — favorite-–author, R^{af} — author–favorite, R^{oo} — opinion–opinion, R^{oa} — opinion–author, and R^{ao} — author–opinion. Similar relationships can be recognized in every multimedia sharing system.

2.5.6 Auction Systems

The main goal of the auction systems is to enable people to sell and buy different products to and from other users. The examples of such systems are eBay or OnSale, in which people as well as businesses can buy and sell their goods and services worldwide. Obviously, every person who wants to use such a system must register with a unique name that becomes the user internet identity *IID*. After the log in to the system, the members can create new auctions and sell things as well as buy different items so the natural relationships between buyers and sellers emerge. Nevertheless, this is not the only type of relations that can be identified in the systems of this kind. Potential buyers can ask the seller a question referring to products they offer using additional system functionalities. Usually, sellers may remove some bids provided by unreliable users establishing in this way a kind of negative relationship. Moreover, users have the possibility to directly invite some selected members to participate in the auctions they manage. Once the auction is completed, both the winner and seller can comment on the quality of the service. Additionally, indirect relationships between buyers or sellers can be extracted due to similar items bought or sold. The auction systems with their functionality (especially ratings) provide the opportunity to analyze not only the existence of the relationships but also their intensity and dynamics.

2.5.7 Social Search Engines

A social search engine is a type of search engine that generates the answer to user queries and evaluates its relevance based on the interactions or contributions of other users. Before the social search engine will be able to provide this type of answer, the appropriate information about user preferences must be gathered. This can be done for example by social bookmarking or the system can ask the user whether the answer to the query is relevant or not. Every user of this kind of service must possess their own account (the internet identity) that enables the personalization to be permanent. People can exchange the information about their preferences so that we can create a social network connecting people with similar interests. Many forms of social search may be distinguished, from the simple shared bookmarking or tagging of the content to more sophisticated approaches that combine human intelligence with computer Information Retrieval algorithms.

On contrary to machine–based searching, e.g. using Google's PageRank, the social approach gives the opportunity for more personalized and in consequence probably more relevant answers to queries asked by the specific individuals.

There are some start–up portals for social searching like Wikia Search, Mahalo.com.

2.5.8 Social Bookmarking and Cataloging

Social bookmarking enables users to store, organize, search and last but not least share with other users bookmarks of web pages [50]. Some popular sites serving as social bookmarking are: del.icio.us [82], Furl, Google Bookmarks, Diigo.

The bookmarks depending on the features of the given service can be saved privately, shared only with some chosen individuals, groups or only inside a certain network. Most social bookmarking services enable users to organize their bookmarks with the shared tags and/or folders. They also enable viewing bookmarks associated with the given tag. Most of the bookmarking services provide also additional features such as the possibility of rating and commenting on bookmarks, the ability to import and export bookmarks from web browsers, emailing of bookmarks, web annotation, and building groups, etc.

All above features enable to extract social networks within social bookmarking sites, in which user logins are the internet identities and separate relationships can be derived from different shared meeting objects like bookmarks, tags, folders, groups, etc.

Social cataloging is a concept similar to social bookmarking. Its main aim is to provide users the opportunity to catalog things they possess, e.g. books, music, films, etc. Each user creates and shares with other members the description of items they want to catalog.

Some popular sites serving as social cataloguing are: (i) for books – LibraryThing, Shelfari, Goodreads, Anobii, Books iRead, (ii) for music – Discogs, Rate Your Music, Last.fm, (iii) for movies – Flixster; (iv) for scholary citations – Bibster, CiteULike, Connotea.

User share the metadata about the items as well as interact and cooperate with each other by improving their descriptions. The social network can be created based on the description of the item that can be treated as an object that connects users who participate in its creation and maintenance.

An example of social bookmarking and simultaneously social cataloging service is CiteULike, which facilitates sharing scientific references among researchers. It also supports import of bibliographical descriptions directly from some most common sites such as Amazon.com, SpringerLink or ScienceDirect.

2.5.9 Homepages

A single homepage is the web site that contains and provides information about a specific person. Homepages are usually maintained by users to whom these pages belong; they can add and update information about their life, work and interests. Users can also incorporate some hyperlinks to others' homepages into their HTML contents. These external, linked homepages can belong to their friends, family members, partners they cooperate with or even other people being considered as interesting by the creator. The URL address of the homepage can be treated as the internet identity of the person this web site belongs to. Moreover, all links to others homepages are signs of direct relationships from the given internet identity to all others it links to. Hence, the system of homepages is an example of the (*HSN*). Note that the relationships are asymmetrical; it means that the target homepages do not have to contain the reverse links. Besides, in such a network it is not possible to assess the strength of relations, so there are only two states: a relation either exists or not.

2.5.10 Knowledge Sharing Systems

Knowledge markets (Experts-Exchange, Mahalo Answers, Yahoo! Answers [82], Knowledge Search, ChaCha.com, Answerly.com) are examples of social networks that enable users, on one hand, to post a request and set a virtual price for the relevant answer while, on the other hand, to answer the questions that others have asked and get reward for the correct answers.

Moreover, users have the opportunity to rate and comment the answers they have received from others. Based on each type of these activities the separate layer in *SSN* can be created. The knowledge in such systems is treated as regular, tangible goods. The currency that is used to pay for the most relevant answers are points as in Experts-Exchange or virtual currency as in the case of Mahalo Answers where the binding currency is Mahalo Dollar. Nevertheless, none of the enumerated systems enables to pay using real money for the valid information and the only award for the correct answers is the high prestige among other network members. ChaCha and Answerly are the examples of the systems, in which the experts are paid for their answers but people can still use the system for free.

2.5.11 Virtual Worlds and Multiplayer Online Games

The virtual worlds and multiplayer online games (Second Life, Sims, World of Warcraft) are the examples of systems, in which users maintain their own avatars that represent them in the virtual world. People can create not only their avatar but also the whole neighborhood they want to live in. This leads to the situation that social entities create for themselves the second life that is parallel to the real one. In online games, users can cooperate with other players by attending common missions. Sometimes, there is even a situation that a task cannot be accomplished by a single person. Thus, it is inevitable that users merge into groups.

These systems somehow map the real world to the virtual one. That means that the virtual world social networks can correspond to the real world social networks. The internet identities will be the avatars and any cooperation between them is the basis for creation of the relationship between the network members.

2.5.12 Collaborative Authoring Systems — Wikis

Wikis are yet another example of social networks on the Internet where users represented by their internet identities collaborate in order to create the common content. For example, in Wikipedia one user initiates the work on the specific term and other users can contribute by changing and improving the term description. Such cooperation, while creating the content, provides the opportunity to obtain the outcome, the quality of which is higher than in the case of a single author. An article or a term description in collaborative authoring systems plays a similar role to the one that plays description of an item in social cataloging systems, i.e. it is a meeting object that connects people who are involved in the process of creating it.

2.5.13 Friend Of A Friend Project

The general purpose of Friend Of A Friend (FOAF) project is quite simple: to build such representation of users, their activities, and acquaintances that can be processed by a computer. In order to achieve this goal the appropriate FOAF machine–readable ontology was developed. These comprehensive users' profiles, which also include the links to their friends, create a homogenous social network *HSN* [21], [90]. From the technical point of view, FOAF files are defined using Web Ontology Language (OWL) being an extension of Resource Description Framework (RDF). The usage to FOAF is free so every internet user can exploit it to create their personal profile and to define the relationships maintained by this person. People are mostly using the FOAF format to put their personal data into an RDF file and to publish it on their homepages. Next, web crawlers gather and aggregate the information, for example SECO [57]. Moreover, every participant possesses a unique identity – OpenID that is used while processing the relationships defined by this user. This enables computers to find people who are similar to each other or who maintain similar relationships. Recently, many social networking sites have started to support the FOAF format to exchange user profile information [51].

2.5.14 Complex Communications Systems

The popularity and diversity of the instant messengers was the inspiration to create some integrated services such as eBuddy or Miranda that enable to join together separate user accounts from different communication systems. For example, eBuddy, which is web and mobile messenger, supports multiple instant messaging services such as Windows Live Messenger, Yahoo!, AIM, Google Talk (GTalk), Facebook and MySpace IM, ICQ within one interface. The eBuddy system utilizes its own eBuddy ID (a joint internet identity) to authenticate its users.

2.5.15 Comparison

The enumerated and described above categories of social networks on the Internet can be compared in many separate aspects (see Table 2.2 and Table 2.3). The layers within ties, see Sec. 5.4, which can be distinguished within each of the network classes are presented in Table 2.2. These layers are derived from different types of user activities within the given category of social networks, including direct mutual communication between users via different communication channels, similar and shared activities towards a certain meeting object (e.g. common usage of tags, commenting or adding to favorites of the same objects, being a part of the same group, etc), contact lists, or even similarities between users' profiles they maintain.

Based on the analyses of the characterized categories of social networks on the Internet their key features were identified and the comparison of these characteristics between different social network classes is presented in Table 2.3.

Table 2.2: Layers in system-based social networks on the Internet

Category	**Layers in the social network (directness of relationships)**
Email service	a) Communication: sent/received emails (direct) b) Address book
Instant messengers	a) Communication: a separate layer for each communication channel, e.g. text messages, VoIP, SMS, video conference, etc (direct) b) Address book (direct) c) Profile-based similarities (indirect)
Blog service	a) References to other blogs (blogrolls[2]) (direct) b) Comments, a separate layer for commentators and commentator–author (quasi–direct, meeting object: a post) c) Common usage of tags/keywords/categories (quasi–direct, meeting object: a tag, keyword, category) d) Profile-based similarities (indirect)
Social networking sites	a) Communication, separate layer for each communication channel, e.g. sent/received emails, video conference, etc (direct) b) Contact list (direct) c) Groups of interest, school classes (quasi–direct, meeting object: a group) d) Profile-based similarities (indirect)
Multimedia sharing systems	a) Contact list (direct) b) Comments and favorites: a separate layer for commentators and commentator–author of the shared objects as well as for favorite–favorite and favorite–author (quasi–direct, meeting object: a multimedia object) c) Common usage of tags (quasi–direct, meeting object: a tag) d) Common groups (quasi–direct, meeting object: a group)

[2] Blogrolls can be seen as the address books in case of the email service or instant messengers

Category	**Layers in the social network**
Auction systems	a) Auction: a separate layer for seller–bidder, seller–buyer/commentator and bidder–bidder (quasi–direct, meeting object: an auction, a bid, transaction, or comment to the auction) b) Invitation to the restricted auction: inviting-invited, invited-invited (quasi–direct, meeting object: an auction) c) Communication: questions and answers (emails) referring the auction (direct) d) Removal of unwanted bids by the seller, a negative relationship (direct) e) Profile–based similarities (indirect)
Social search engines	a) Profile–based similarities (indirect)
Social bookmarking and cataloging	a) Shared bookmarks or item descriptions (quasi–direct, meeting object: bookmark, bibliographical description) b) Profile–based similarities (indirect)
Homepages	a) References (hyperlinks) to other homepages (direct)
Knowledge sharing systems	a) Answers and questions: a separate layer for users who answer the same question (answer–answer) and question–answer that consists of relations between author of the question and people who answer this question
Virtual worlds	a) Communication: chats between users (direct) b) Common participation in missions (quasi–direct, meeting object: mission)
Wikis	a) Creating of common articles (quasi–direct, meeting object: term)

Table 2.3: Features of the system–based social networks

No.	Feature	Email service	Instant messengers	Blog service	Social networking sites	Multimedia sharing system	Auction system	Social search engine	Homepages	Social bookmaring	Virtual worlds
1	Public availability of data about social network	No	No	Yes	Medium[a]	Yes	Medium[b]	No	Yes	Yes	No
2	The build-in function for searching for members to get into new relationships with	No	Yes	Yes	Yes	Yes	Yes	No	No	No	No
3	Visibility of own incoming relationships within the GUI system[c]	No	No	No	Yes	No	No	No	No	No	No
4	Visibility of own outgoing relationships within the GUI system[d]	Partly[e]	Partly[f]	Partly[g]	Yes	Partly[h]	Yes[i]	No	Yes/No[j]	No	No

[a] The access to the data about the whole network is usually restricted, i.e. there is only access to the relationships of the recently searched people

[b] Data about some user activities are available only for the system operator, e.g. removal of unwanted bids, some historical auctions

[c] It means that there exists a function in the system that directly displays the list of the connections from other users to the given one

[d] It means that there exists a function in the system that directly displays the list of the connections from the given user to others

[e] Relationships derived from address books are visible whereas those based on outgoing emails are not

[f] Relationships derived from contact lists are visible whereas those created upon communication only are not

[g] In case of direct reference layer – yes (blogrolls), in case of the comments – no

[h] Relationships derived from contact lists are visible whereas those created upon communication only are not

[i] Most relationship layers are visible in the form of a sort of lists, however, there is no list of bidder-bidder relations aggregated from all auctions

[j] Outgoing hyperlinks are visible but separate lists of links to other homepages may exist in the form of "see also" section.

No.	Feature	Email service	Instant messengers	Blog service	Social networking sites	Multimedia sharing system	Auction system	Social search engine	Homepages	Social bookmaring	Virtual worlds
5	The possibility of defining the type of relationship	No	No	No	Yes	No	No	N/A	No	N/A	No
6	The awareness level of being in the incoming relation	High/None[a]	High/None[b]	Medium[c]	High	Medium[d]	High	None	None[e]	None	High
7	The awareness level of being in the outgoing relation	High	High	High	High	High	High	None	High	None	High
8	Does the user have to directly state that he/she is with someone in the relationship?	No	No	Yes/No	Yes	Yes/No	Yes/No	No	Yes	No	No

[a] Users are conscious of emails they receive but the content of address books maintained by others is usually unknown

[b] Users are conscious of incoming communication but the content of address books maintained by others is usually unknown

[c] There is information in the system that other user refers to the given person as an acquaintance, however this information is not directly given to this user

[d] There is information in the system that other user refers to the given person as an acquaintance, however this information is not directly given to this user

[e] To obtain information about homepages referring to the given one the usage of search engine is necessary – the obtain results are not necessarily complete nor actual

No.	Feature	Email service	Instant messengers	Blog service	Social networking sites	Multimedia sharing system	Auction system	Social search engine	Homepages	Social bookmaring	Virtual worlds
9	Dedicated tools for relationships maintenance	Yes/No[a]	Yes/No[b]	No	Yes[c]	Yes/No[d]	No	No	Yes/No[e]	No	Yes/No
10	Do groups of *IID*s exist?	Yes[f]	Yes[g]	Yes[h]	No	No	Yes/No[i]	No	Yes[j]	No	No
11	Is the data about relations centrally maintained?	Yes[k]	Yes/No[l]	Yes[m]	Yes	Yes	Yes	No	No[n]	No	Yes

[a] "Yes" refers only address books

[b] "Yes" refers only address books

[c] E.g. the information that reminds users about their friends birthdays

[d] "Yes" refers only contact lists

[e] There are some content management systems (CMS) with buit–in reference management mechanisms

[f] E.g. the group alias that serves to customer relationship management purposes

[g] E.g. the group alias that serves to customer relationship management purposes

[h] E.g. two persons can be the coauthors of one blog and in consequence both of them use the same *iid*

[i] There are some stores, even large ones, that do their businesses via auction systems

[j] Homepages can belong to organizations

[k] Yes, but the data can be centrally maintained only within one email server/domain, e.g. @google.com)

[l] It depends on the provider of the service, e.g. in the case of Skype system the data about contact lists are centrally maintained whereas the data about communication are not stored

[m] Yes, but only within a single blog service, e.g. WordPress

[n] The data about relations is centrally maintained only in the case when all the homepages are localed on the one server)

Chapter 3

User Position in Social Network

3.1 Social Network Analysis

This chapter is devoted to the social network analysis (*SNA*). The goals and methods used in the network analysis are presented and described in details.

In social networks some typical phenomena such as small world effect [91], [121], clustering [32], strong and weak ties [52] and many others may be observed. Various human features, extracted from user profiles, which can have more or less significant influence on the process of formation of a relationship, can be also discovered. In order to identify these phenomena the appropriate *SNA* method ought to be applied.

Social network analysis stems from traditional social analysis used by sociologists and anthropologists in the first half of the 20th century. After introducing mathematical interpretation of social networks scientists started developing social network analysis.

SNA can be defined as "*the disciplined inquiry into the patterning of relations among social actors, as well as the patterning of relationships among actors at different levels of analysis (such as persons and groups)*" [19]. Another definition of *SNA* was proposed by Valdis Krebs: "*Social network analysis (SNA) is the mapping and measuring of relationships and flows between people, groups, organizations, computers, web sites, and other information/knowledge processing entities. The nodes in the network are the people and groups while the links show relationships or flows between the nodes. SNA provides both a visual and a mathematical analysis of human relationships.*" [79].

It should be emphasized that the regular social data (Table 3.1) is quite different than social network data (Table 3.2). Traditional social data describes actors whereas social network data can contain social data but mainly describes connections between actors rather than actors themselves [55]. In other words, network data analysis puts emphasis not on the individuals themselves, but on the relationships among people [55]. Because of the fact that the social network analysis focuses on investigation of connections it does not mean that *SNA* is not interested in actors. After drawing conclusions network analysis may focus on actors to retrieve additional information and to better understand the network, however it is not its primary goal.

Table 3.1: Example of simple social data

Name	**Surname**	**Gender**	**Age**	**Marital status**
Kate	Davis	Female	29	Single
Frank	Martin	Male	37	Divorced
Jason	Smith	Male	56	Married
Ann	Jones	Female	25	Married
Carol	Damon	Female	43	Single

Table 3.2: Example of social network data. 0 means that person A does not know person B, 1 means that person A knows person B

Name A/B	**Kate**	**Frank**	**Jason**	**Ann**	**Carol**
Kate	—	1	0	1	1
Frank	1	—	0	0	1
Jason	0	0	—	0	1
Ann	1	0	0	—	1
Carol	1	1	1	1	—

Four main steps in *SNA* can be distinguished [46]:

— Selecting a sample

— Collecting data

— Choosing and applying the method of *SNA*

— Drawing conclusions

In order to identify and investigate the patterns that occur within the network, first the selection of a group of people should be done. The possibility of analyzing every node of the network (especially these huge and heterogeneous) is usually limited by the available resources and because of that the representative group of actors ought to be chosen for further analysis. This group of actors is called population [55] or sample [46]. After that, the data is collected. Many methods of gathering data such as questionnaires, interviews, observation, and artefacts exist [46]. However, most of researches agree that the best method is the hybrid one that copes with the shortcomings of the enumerated methods and combines all of them [107]. Additionally, the network data differs from the conventional sociological data [55]. Network data in contrast to the traditional data, which consists of rectangular array of measurements, consists of a square array of measurements. The columns and rows of the array are subjects or cases and each cell of the array describes the relationship between these subjects [55]. The *SNA* research has identified three types of data also called units of analysis, which should be and are investigated: relations, ties [46], and actors.

The next step in SNA is to choose the most suitable method of analysis. The crucial methods, which are currently used to identify the structure of a social network, the phenomena, and their intensity are (Figure 3.1):

— Full network method

— Snowball method

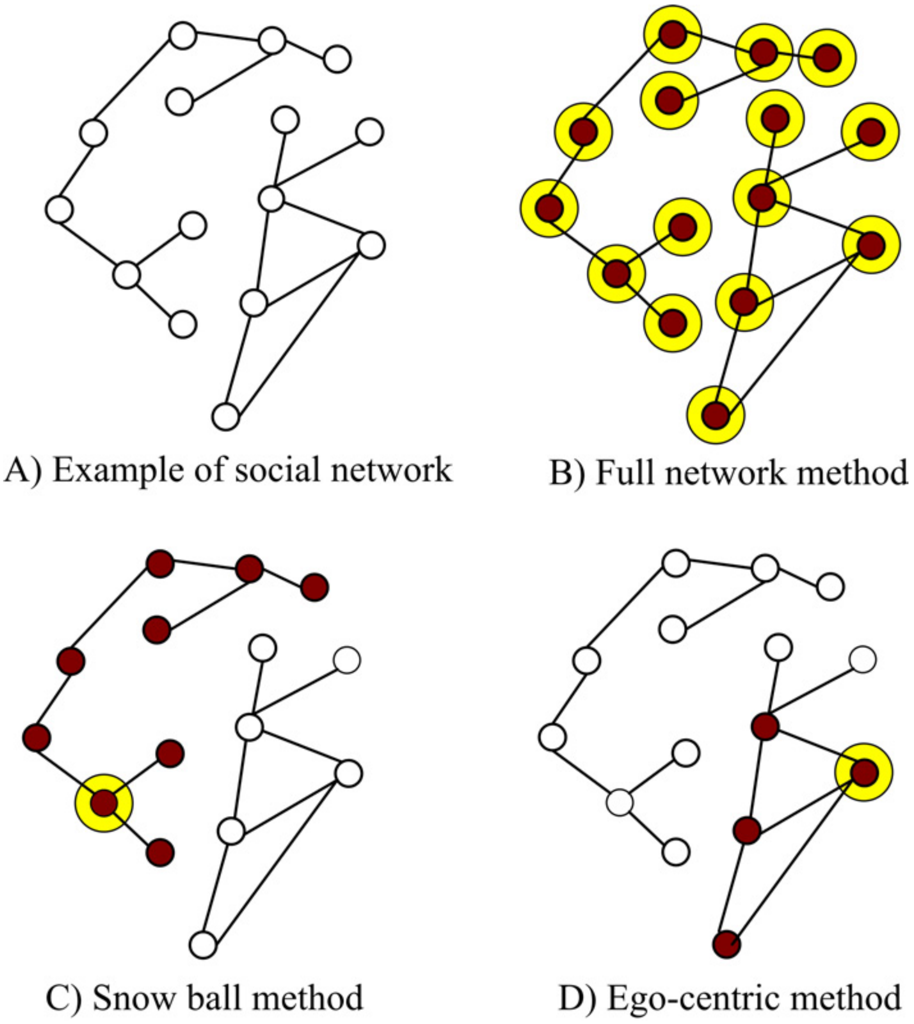
— Ego–centric method ”with alter connections”

— Ego–centric method ”ego–only” [55], [46]

The full network method is the most complex one, because all members of the network and all their possible connections are taken into consideration [55]. To analyze the whole network not only the complete list of connections between people is created, but also the links to external environment [46]. This is resource consuming process; nevertheless, the biggest advantage of this approach is that it provides one full and integrated view on all ties within the network. On the other hand, it is really hard to create such description, because it demands resources and is time consuming. Additionally there is always the possibility that some of the connections will be missed, especially in a case of an extensive network with many ties.

An alternative strategy, which is less complex, is the snowball method [55]. Firstly, we define a group of actors (nodes) who describe their connections to other people. Next the same task, i.e. an identification of all outgoing connections, is done for the actors that have been identified in the first step. This recurrence is executed until all ties have been defined or we have decided to stop creating new ties due to time limits. The biggest shortcoming of this method is the strong possibility that not all connections and not all actors, particularly isolated ones, will be identified.

If there is no need to identify all connections in the network, the ego–centric method can be used [55]. It focuses on a single individual rather than on groups or pairs. In the first step, one ”ego” is chosen. The information about this ego connections is retrieved, together with their target actors and relationships among them. As a result, a sub–network is created that helps to understand the possibilities and constraints of the given individual. In this approach, we consider the ”ego” and their alter direct connections [55]. However the ”ego only” approach can be also exploited. In this case, we are not interested in the connections between the various alters but we only concentrate on a single ego and their first level connections [55].



A) Example of social network

B) Full network method

C) Snow ball method

D) Ego-centric method

Figure 3.1: Methods of social network analysis

The last step that enables to identify the existing within the particular social network patterns is to draw the conclusion from the investigation. The issue that has to be emphasized is that collecting network data and picking the right method of analysis is an extremely challenging task.

In this dissertation the full network method was introduced and utilized because in the node position measure we assume that the position of an individual recursively depends on the positions of all network members.

3.2 User Position

Measures (also called metrics) are used in social network analysis to describe the actors' or ties' characteristic features 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. However, only the measures that can be used in directed graphs are considered because the new proposed measure is dedicated to be applied in directed graphs. Directional relationships result in two types of position measures, i.e. prestige and centrality. Both of them consider how prominent the user is within a community. Prominent means that a given actor is particularly visible to other members of the network [77], [62], [45]. Note that centrality is based on the "choices" made by a user whereas prestige depends on "choices" that a given user received from others [119]. 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 [18]. The term "centrality" or "prestige" is by no means clearly defined [18] and can be interpreted as influence, prestige or control.

3.2.1 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 [119]. The idea of prestige was first introduced by Moreno and he called this concept as status [94].

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 [119], [26]. In other words, more prominent people are those who received more nominations from members of the community [4]:

$$IDC(x) = \frac{i(x)}{m-1} \tag{3.1}$$

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.

Proximity Prestige

Proximity prestige $PP(x)$, in contrast to closeness centrality, reflects how close all members are within the social community to member x [119]. This measure depends on the geodesic distances e.g. the length of the shortest paths[1] from all members to x:

$$PP(x) = \frac{\frac{p(x)}{m-1}}{\frac{1}{p(x)}\sum_{i=1}^{p(x)} d(x,y)} = \frac{p(x)^2}{(m-1)\cdot\sum_{i=1}^{p(x)} d(x,y)} \tag{3.2}$$

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

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 [66]. ”It’s not what you know, but whom you know” [119]. In this method, first the initial centrality of each member is established by utilization of one of the existing measures e.g. degree, betweeness 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, the eigenvector–like measure of 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 [13]. As an extension of this method, called alpha-centrality, Bonacich and Lloyd [13] 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

[1] Algorithms that can be used in the process of calculating the shortest paths are presented in Appendix A.

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 eigenvector prestige takes into account not only the direct connections but also the indirect ones. Moreover, each indirect path is assigned with an appropriate weight [12].

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 [80] and this enables similar node measures to be applied both to social networks and hypertext or web-based systems [15]. PageRank was introduced by Brin and Page to assess the value and importance of web pages [22], [10], [23]. 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. Note that all links in PageRank have the same weight and importance.

3.2.2 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 [119]. 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 [7].

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 [103], [111]:

$$ODC(x) = \frac{o(x)}{m-1} \tag{3.3}$$

where:
$o(x)$ — the number of the first level neighbors 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 [119]. 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.

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 [18]:

$$EC(x) = \frac{1}{max\{d(x,y) : y \in M\}} \tag{3.4}$$

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

Closeness Centrality

The closeness centrality pinpoints how close a member is to all the others within the social network [7]. 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 [8]. A similar idea was studied for hypertext systems [15]. 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 [108] and is calculated as follows:

$$CC(x) = \frac{m-1}{\sum_{y \neq x, y \in M} c(x,y)} \tag{3.5}$$

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 [26], [33], [79].

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 [26], [33], [43], [44], [110]:

$$BC(x) = \frac{\sum_{i \neq x \neq j, i,j \in M} b_{ij}(x)}{b_{ij}} \tag{3.6}$$

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.

3.3 Comparison of the User Position Indices

The existing user position measures can be divided based on the way in which they are calculated into three main groups: degree, shortest paths and rank group (Table 3.3). 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 take into consideration 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 proposed in this thesis node position.

Table 3.3: Groups of the user position indices

Group Name	**Name of User Position Indices**	**Complexity**
Node Degree	Indegree Centrality, Outdegree Centrality	Local measure — considers only the number of the first–level relationships [18] – Does not take into account the position of other nodes
Shortest Paths	Eccentricity Centrality, Closeness Centrality, Betweenness Centrality, Proximity Prestige	– Global measure — calculates the length of the shortest paths to all nodes – Does not take into account the positions of other nodes
Rank of Actor	Rank Prestige	– Global measure — takes into account the positions of other nodes

Each of the described measure can be applied in different situations and all of the posses different characteristics. The features and examples of application are presented in Table 3.4.

The existing centrality and prestige measures suffer from many shortcomings (Table 3.5) that result in the need for developing a new method that cope with these disadvantages. The most important ones are enumerated

Table 3.4: Comparison of User Position Indices

Name	**Properties**	**Example of Application**
IDC	– 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) [18]
PP	– 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
RP	– Global measure – Consider the positions of other nodes	– Identifying members who are perceived by others as important users
ODC	– 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
EC	– Determine a node that minimizes the maximum distance to any other node in the network [18] – Global measure	– Facility location problem that uses minmax criterion [54], e.g. determining the location for an emergency facility
CC	– Node with high *CC* can be very productive in communicating information to others – Actors with high *CC* need not rely on others to receive information [7] – It is inversely related to distance between nodes	– Whenever we are interested in people who can spread the information about e.g. special offers – Minisum location problem [54] 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
BC	– Express the extent to which a given node controls the communication between two nodes	– Finding users who control the information transfer between other members

below together with the short notion about how the new developed node position method will overcome 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 [119]. However, this problem can be no more neglected as most of the networks in the Internet consist of more than one type of relationship (see Chapter 2). The proposed in this thesis method copes with this problem and this is undoubtedly added value in the field of node position methods (see Appendix B).

- **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 proposed in this thesis 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 Networs** — all methods based on the shortest paths and in consequence rank prestige 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 proposed method in this thesis 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 to 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 (see Chapter 6.4). Only the degree–based measures are efficient. However, these methods tend to diversify users only to limited extend. The performed experiments have revealed that over 95% users obtain the same degree prestige and centrality (see Chapter 6). 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 proposed in the thesis 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 (see Appendix C).

Table 3.5: Advantages and Disadvantages of User Position Indices

Name	**Advantages**	**Disadvantages**
IDC	– Simple and easy to compute – Quite informative in many applications [119]	– Big number of Duplicates (Section 6.2 and 6.3) – Local measure – takes into account only firs–level neighborhood [18] – Lack of applicability in multirelational weighted networks
PP	– Global measure – Consider the whole topological structure of a network	– The disconnected graph results in value 0 of *PP* 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
RP	– Global measure – Consider the position of other nodes	– 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
ODC	– Simple and easy to compute – Quite informative in many applications [119]	– Big number of Duplicates (Section 6.2 and 6.3) – Local measure – takes into account only firs–level neighborhood [18] – Lack of applicability in multirelational weighted networks
EC	– Global measure – Consider the whole topological structure of a network	– The disconnected graph results in value 0 of *EC* 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

Name	Advantages	Disadvantages
CC	– Global measure – Consider the whole topological structure of a network	– The disconnected graph results in value 0 of *CC* 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 *CC* is not defined for a given pair of nodes [18] – Very complex and inefficient in large networks – Lack of applicability in multirelational weighted networks
BC	– Global measure – Consider the whole topological structure of a network	– The disconnected graph results in value 0 of *BC* 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 [25] – Very complex and inefficient in large networks – Lack of applicability in multirelational weighted networks

Chapter 4

Node Position in Network of Internet Users

4.1 General Concept

On the Web, in the era of Web 2.0, there is a great need to assess not only the significance of web pages created by people and published in web services [22], but also the importance of people within virtual social networks. The new method proposed and studied in this thesis is called **Node Position** *NP* and it enables to estimate how valuable the particular individual within the network of Internet users *NIU* is.

The importance of the node in the weighted and directed *NIU*, expressed by the node position function, tightly depends on the strength of the relationships that other members of the network maintain with the given node as well as on the node positions of these members — called acquaintances. In other words, the member's node position is inherited from others but the level of inheritance depends on the activity of the members directed to this person, i.e. intensity of common interaction, cooperation or communication. The activity contribution of one user absorbed by another is called commitment and is presented in Figure 4.1 as weights of edges.

Node position function $NP(x)$ of a member x in the social network of Internet users, respects both the value of node positions of all other network members as well as the level of their activities in relation to x [68], [69], [70], [73]:

$$NP(x) = (1 - \varepsilon) + \varepsilon \cdot \sum_{i=1}^{m_x} (NP(y_i) \cdot C(y_i, x)) \tag{4.1}$$

where:
ε – the constant coefficient from the range $(0; 1]$;
y_i — x's acquaintances, i.e. the members who are in direct relationship to x: $C(y_i, x) > 0$;
$C(y_1, x)$,...,$C(y_m, x)$ – the commitment function that denotes the contribution in activity of y_1,...,y_m directed to x.

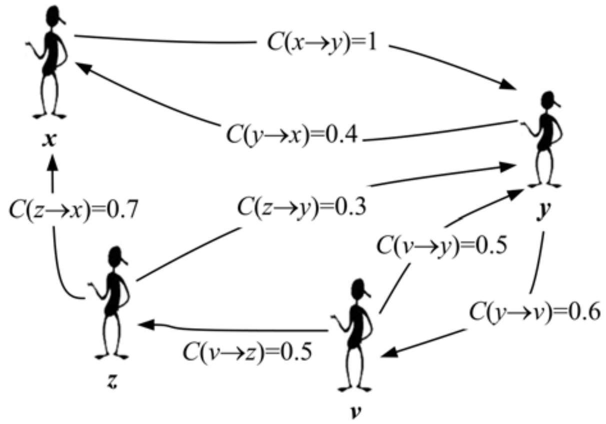


Figure 4.1: Example of the network of Internet users *NIU* with the assigned commitment values

m_x — the number of x's acquaintances.

The value of ε denotes the dependence of node position measure on external influences: to what extent x's node position is 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 greater 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 people, i.e. with the highest 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 involve their acquaintances. The node position of the 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.

There are five important constraints regarding commitment function derived from the relationships $C(y,x)$ in $NIU(IID,R)$ [73]:

1. Commitment function $C(y,x)$ reflects the strength of the relationship from y to x in $NIU(IID,R)$, $x,y \in IID$, $x \neq y$. If there exists the relationship $(y,x) \in R$ then $C(y,x) > 0$. If there is no relationship from y to x, i.e. $(y,x) \notin R$ then $C(y,x) = 0$, except in the case of condition 5.

2. The value of commitment is within range $[0;1] : \forall(x,y \in IID)C(y,x) \in [0;1]$.

3. Commitment function to itself equals 0: $\forall(y \in IID)C(y,y) = 0$.[1]

[1] In the case when the user e.g. sends emails to himself/herself then this communication is not taken into consideration and is excluded from the further analysis

4. The sum of all commitments has to equal 1, separately for each network member:
$$\forall(y \in IID) \sum_{x \in IID} C(y,x) = 1 \tag{4.2}$$

5. If a member y is not active to anybody but some other members x are active to y and since no isolated members are allowed in $NIU(IID, R)$, in this case, to satisfy condition 4 (Equation 4.2), the sum 1 is distributed equally among all y's acquaintances – x (Figure 4.2), i.e. all values of $C(y,x)$:

$$\begin{aligned} \sum_{z \in IID} C(y,z) = 0 \Rightarrow& \\ \Rightarrow \forall(x \in IID : C(x,y) > 0)& \\ C(y,x) = \frac{1}{card(\{x \in IID : C(x,y) > 0\})}& \end{aligned} \tag{4.3}$$

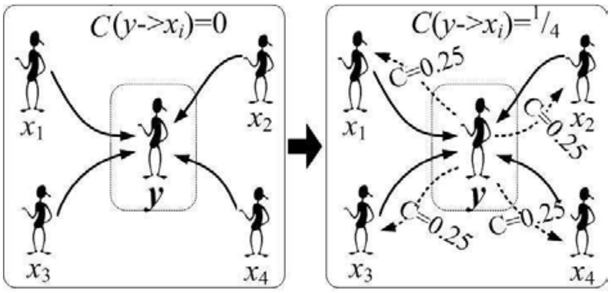


Figure 4.2: Distribution of the commitment for an inactive member y equally among all y's acquaintances

In other words, the value of commitment function $C(y,x)$ from y to x is usually obtained from raw data about direct activity of member y in relation to x or as the equal potential contribution in activity in case of the total lack of y's activity.

Member y from Figure 4.2 is not active to anybody within the network, but there are four members (x_1, x_2, x_3, x_4) who are active to user y. In this case the commitment function is equally distributed among all y's acquaintances.

Note that the network of internet users $NIU(IID, R)$ must not contain any isolated members. This restriction is derived from the lack of possibility to satisfy all enumerated above conditions for such members, especially condition 4 (Equation 4.2).

The consequence of the 4th constraint is that if member y is active to only one other member x, then $C(y,x) = 1$.

4.2 Commitment Evaluation

To assess the strength of the relationship between two individuals x and y within the network of Internet users the commitment function $C(y,x)$ is used. It denotes the amount of the member y's activity that person y passes to member x.

The commitment $C(y,x)$ of member y within activity of acquaintance x is directly evaluated from source data as the normalized sum of all contacts, cooperation, and communications from y to x in relation to all activities of y:

$$C(y,x) = \begin{cases} \dfrac{A(y,x)}{\sum\limits_{x \in IID} A(y,x)}, & \text{when} \sum\limits_{x \in IID} A(y,x) > 0 \\ 0, & \text{when} \sum\limits_{x \in IID} A(y,x) = 0 \end{cases} \tag{4.4}$$

where:
$A(y,x)$ — the function that denotes the activity of person y directed to member x, e.g. number of emails sent by y to x; $A(y,x) \geq 0$;
m — the number of nodes in *NIU*.

Note that according to requirement 3 for the commitment function we need to ensure that $A(y,y) = 0$, i.e. emails sent to themselves are excluded.

As it can be easily proved Equation 4.4 fulfills also all other requirements for relationship commitment function. Note that there may exist some inactive members y in the network, for which $\sum_{x \in IID} A(y,x) = 0$ and in consequence $\sum_{x \in IID} C(y,x) = 0$. In all such cases the process described in condition 5 (Equation 4.3) needs to be performed, in order to fulfill the fourth condition (Equation 4.2).

One of the activity types within the network of Internet users is communication via emails. In this case the commitment function $C(y,x)$ will be calculated as the number of emails sent by user y to x divided by the number of all emails sent by user y.

The time factor is not considered in the Equation 4.4. Similar approach is utilized by Valverde et al. where the strength of the relationships is established by the number of emails sent to a person in the group [116]. However, the authors do not respect the general activity of the given individual. This general, local activity exists in the form of denominator in Equation 4.4. The approach proposed by Valverde et al. suffers from a significant shortcoming. It means, if a user x sends ten emails to a user y and overall 1,000 emails and a user z sends ten emails to a user y and overall 15 emails then both of these relationships are of the same strength. However, the significance of a relation $x \rightarrow y$ is different than of the relation $z \rightarrow y$ as the x sends to y only 1/100 of all emails whereas z sends 2/3 of all emails to y. The results in a fact that relationship strength calculated in this way is not comparable.

In another version of relationship commitment function $C(y,x)$ all member's activities are considered with respect to their time. 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 in each period are considered separately for each individual:

$$C(y,x)=\begin{cases}\dfrac{\sum\limits_{i=0}^{k-1}(\lambda)^i\cdot A_i(y,x)}{\sum\limits_{x\in IID}\sum\limits_{i=0}^{k-1}(\lambda)^i\cdot A_i(y,x)} & \text{when } \sum\limits_{x\in IID}\sum\limits_{i=0}^{k-1}(\lambda)^i\cdot A_i(y,x)\geq 0\\ 0, & \text{when } \sum\limits_{x\in IID}\sum\limits_{i=0}^{k-1}(\lambda)^i\cdot A_i(y,x)=0\end{cases} \tag{4.5}$$

where:
i — the index of the period: for the most recent period $i = 0$, for the previous one: $i = 1, \cdots$, for the earliest one $i = k$–1;
$A_i(y, x)$ — the function that denotes the activity level of person y directed to member x in the ith time period, e.g. number of emails sent by y to 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 person y 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. For example, if one year's data set is processed and a period is a month then $k = 12$. For $\lambda = 0.9$, the data from January is considered with the factor $0.9^{11} = 0.31$, for February we have $0.9^{10} = 0.35, \cdots$, for October $0.9^2 = 0.81$, for November — 0.9 and finally for December $0.9^0 = 1$. This in a sense is similar to an idea which was used in the personalized systems to weaken older activities of recent users.

One of the concepts that can be also utilized in the time analysis is the sliding time window [40], [65]. The basic idea of the sliding time frame algorithm is to establish the time window (e.g. the period of one month) and then move this window forward by a specific period of time (e.g. day by day). In such a case the values of commitment function are calculated for each time window separately and after that the dynamic of relationship strength and not only can be analyzed.

One of the activity types is the communication via chat. In this case, $A_i(y, x)$ is the number of chats that are common for x and y in the particular period i; and $\sum_{x\in IID} A_i(y, x)$ is the number of all chats in which y took part in the ith period. If person y had many common chats with x in comparison to the number of all y's chats, then x has greater commitment within activities of y, i.e. $C(y, x)$ will have greater value and in consequence the node position of member x will grow. Note that $C(y, x)$ will have value 1 when member x is the only interlocutor of person y.

However, not all of the elements can be calculated in such a simple way. Other activities are much more complex, e.g. comments on forums or blogs. Each forum consists of many threads where people can submit their comments. In this case, $A_i(y, x)$ is the number of user y's comments in the threads in which x has also commented, in period i, whereas sum $\sum_{x \in IID} A_i(y, x)$ is the total number of comments that have been made by all x who are y's friends on these threads, at the same time.

Additionally the calculation of the commitment function within the multirelation networks of Internet users. *SSN* consist of layers and the commitment function is calculated separately for each layer and after that the aggregated commitment function is assessed. The example of commitment function evaluation for the *SSN* extracted from the data gathered from the Flickr system is presented in Appendix B.

4.3 Node Position Calculation

The node position is calculated in the iterative way that means that the left side of Equation 4.1 is the result of iteration while the right side is the input:

$$NP_{n+1}(x) = (1 - \varepsilon) + \varepsilon \cdot \sum_{y \in IID} NP_n(y_m) \cdot C(y, x) \tag{4.6}$$

where:
$NP_{n+1}(x)$ and $NP_n(x)$ — the node position of member x after the $(n + 1)$th and nth iteration, respectively.

To perform the first iteration, we also need to have an initial value of node position $NP_0(x)$ for all $x \in IID$:

$$NP_1(x) = (1 - \varepsilon) + \varepsilon \cdot \sum_{y \in IID} NP_0(y_m) \cdot C(y, x) \tag{4.7}$$

Since the calculations are iterative, we also need to introduce a stop condition. For this purpose, a fixed precision coefficient τ is used. Thus, the calculation is stopped when the following criterion is met:

$$\forall(x \in IID)|NP_n(x) – NP_{n-1}(x)| \leq \tau \tag{4.8}$$

Obviously, another version of the stop condition can be also applied, e.g.:

$$|SNP_n – SNP_{n-1}| \leq \tau \tag{4.9}$$

where:
SNP_n and SNP_{n-1} — the sum of all node positions after the nth and $n - 1$th iteration, respectively.

Based on Equation 4.6 the PIN algorithm (**P**osition **I**n the **N**etwork) was developed [96]. Three versions of this algorithm are proposed in this dissertation, i.e. PIN^{nodes}, PIN^{hybrid}, and PIN^{edges}. These algorithms differ in the implementation and in consequence their efficiency varies.

All algorithms require the same set of input data and provide as the output the node position values for each network member and their ranking position regarding its node position as well as the number of iterations and time that was required to meet the stop condition that is one of the input parameters. Other input data that must be provided in order to evaluate the node position are: the list C that consists the commitment value for each ordered pair $(x_1, x_2) \in IID$, the initial node position for each member of the network, ε coefficient from range $(0; 1]$.

4.3.1 PIN Nodes

The first proposed algorithm PIN^{nodes} is the direct implementation of the node position concept. It has been completed without any optimization techniques. The name of the algorithm comes from the fact that all calculations are made from so called "node perspective", i.e. the node position is calculated one by one for each network node — member.

First, two lists NP_{prev} and NP that contain the node position values are created. NP_{prev} stores node positions from the previous iteration whereas in NP the final values calculated in the current iteration are saved. At the beginning, the initial node positions values NP_0 are assigned to the elements from NP_{prev}.

After that for each member x from IID its NP is set to $(1 - \varepsilon)$. Next, for each member y from IID the value of commitment function $C(y, x)$ is multiplied by $NP_{prev}[y]$ and by ε. The result of this operation is added to the current value of x's node position that is stored in $NP[x]$. Finally, the values from NP are assigned to NP_{prev} and the iteration finishes. The next iteration is performed unless the stop condition is met.

The PIN Nodes

Input:
IID, R - set of members and their relationships,
C - list that consists the commitment values, one for each ordered pair $(x_1, x_2) \in$ IID,
$NP_0 =< NP_0(x_1), NP_0(x_2), \cdots, NP_0(x_m) >$ - vector of initial node positions, $m = card(IID)$,
ε - coefficient from Equation 4.6, $\varepsilon \in (0;1]$,
τ - stop condition (precision coefficient), e.g. $\tau := 0.00001$.
Output:
$NP =< NP(x_1), NP(x_2), \cdots, NP(x_m) >$ - the vector of final node positions, $m = card(IID)$,
Ran - the ranking of individuals from IID,
n - the number of iterations,
t - processing time.

```
begin
 $n := 0$;
 $t := 0$;
 $NP_{prev} := NP_0$
 repeat
  begin
   for (each member x from IID) do
    begin
     $NP[x] := (1 - \varepsilon)$;
       for (each member y from IID) do
         $NP[x] := NP[x] + \varepsilon \cdot NP_{prev}[y] \cdot C[y, x]$;
      end;
     $NP_{prev} := NP$;
     $n := n + 1$;
    end;
   until stop condition 4.8 is fulfilled for all members;
   create ranking list Ran based on NP;
   t :=processing time;
  end.
```

4.3.2 PIN Edges

The second developed algorithm is called PIN^{edges} and its name comes from the fact that all calculations are made from so called "edge perspective", i.e. that the node position is calculated rather by taking into the consideration the edges and their weights (commitment functions assigned to the edges) than evaluating node position one by one for each network node — member.

First, the list NP that contains the initial node position values is created by assigning NP_0 to NP, i.e. that initially all NPs equal 0 for each network member.

Afterwards for each edge $r(x, y)$ from the set R increase node position value of user y ($NP(y)$) node position of x ($NP(x)$) multiplied by the value of commitment function from user x to y ($C(x, y)$). Next for each member of IID multiply the obtained node position of the given user by ε and add the appropriate component $1 - \varepsilon$.

The next iteration is performed if the stop condition is not fulfilled. Otherwise the whole process is completed.

The PIN Edges

Input:
IID, R - set of members and their relationships,
C - list that consists the commitment value for each ordered pair $(x_1, x_2) \in$ IID,
$NP_0 =< NP_0(x_1), NP_0(x_2), \cdots, NP_0(x_m) >$ - the vector of initial node positions, $m = card(IID)$,
ε - coefficient from Equation 4.6, $\varepsilon \in (0; 1]$,
τ - stop condition, i.e. the precision coefficient, e.g. $\tau := 0.00001$.
Output:
$NP =< NP(x_1), NP(x_2), \cdots, NP(x_m) >$ - the vector of final node positions, $m = card(IID)$,
Ran - the ranking of individuals from IID,
n - the number of iterations,
t - processing time.

```
begin
 n := 0;
 t := 0;
 NP := $NP_0$
 repeat
  begin
   for (each edge r(x,y) from R) do
    NP[y] := NP[y] + NP[x] · C[x,y];
   for (each member x from IID) do
     NP[x] := $(1 - \varepsilon) + \varepsilon \cdot$ NP[x];
    n := n + 1;
   end;
  until stop condition 4.8 is fulfilled for all members;
  create ranking list Ran based on NP;
  t :=processing time;
 end.
```

4.3.3 PIN Hybrid

The third algorithm, named PIN^{hybrid} combines two previous approaches.

First, the list NP that contains the initial node position values is created by assigning NP_0 to NP, i.e. that initially all NPs equal 0 for each network member.

After that all nodes of the network are divided into b disjunctive subsets $\{s_1, s_2, \cdots, s_b\}$. Next steps are repeated until the stop condition is fulfilled. For each created subset s_k the following action is performed: for each edge $r(x, y)$ in which y belongs to subset s_k increase y's node position $(NP(y))$ with x's node position $(NP(x))$ multiplied by the value of commitment function from user x to y $(C(x, y))$. Next for each member of IID multiply the obtained node position of the given user by ε and add the component $1 - \varepsilon$.

If the stop condition is fulfilled then the whole process is completed.

The PIN Hybrid

Input:
IID, R - set of members and their relationships,
C - list that consists the commitment value for each ordered pair $(x_1, x_2) \in$ IID,
$NP_0 =< NP_0(x_1), NP_0(x_2), \cdots, NP_0(x_m) >$ - the vector of initial node positions, $m = card(IID)$,
ε - coefficient from Equation 4.6, $\varepsilon \in (0; 1]$,
τ - stop condition, i.e. the precision coefficient, e.g. $\tau := 0.00001$.
Output:
$NP =< NP(x_1), NP(x_2), \cdots, NP(x_m) >$ - the vector of final node positions, $m = card(IID)$,
Ran - the ranking of individuals from IID,
n - the number of iterations,
t - processing time.

```
begin
 $n := 0$;
 $t := 0$;
 $NP := NP_0$
 divide the set IID into b disjunctive subsets $\{s_1, s_2, \cdots, s_b\}$
 repeat
  begin
   for(each disjunctive subset $s_k$) do
    for(each edge $r(x, y)$ where y is member of $s_k$) do
       $NP[y] := NP[y] + NP[x] \cdot C[x, y]$;
     for(each member x from IID) do
      $NP[x] := (1 - \varepsilon) + \varepsilon \cdot NP[x]$;
     $n := n + 1$;
    end;
   until stop condition 4.8 is fulfilled for all members;
   create ranking list Ran based on NP;
   t :=processing time;
  end.
```

4.4 Example of Node Position Calculation

The following case study presents how to calculate the node position of individuals within the network of Internet users as well as the main characteristics of the proposed measure [69]. Let us assume that we have a network of email users as in Figure 4.3. The arc values indicate the values of commitment functions between a pair of members.

In order to calculate the node position the iterative algorithm described above is utilized (see Section 4.3). The calculation of node position is repeated until the defined stop condition is reached. In the presented case study the stop condition is defined as the precision to the fifth decimal place between two following iterations.

Firstly, the influence of ε coefficient's value on the final node position of an individual is investigated. The aim of the second part is to assess the number of necessary iterations in relation to the initial node position values of the members of the human community.

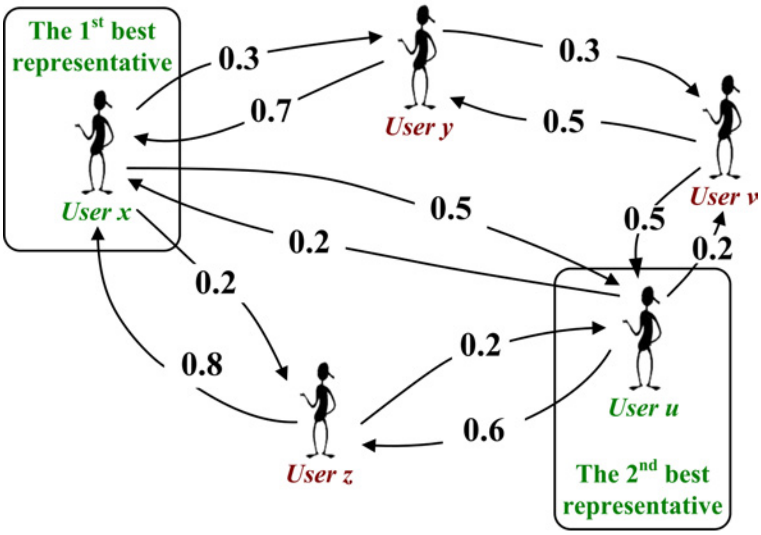


Figure 4.3: The network of Internet users that can be extracted from the email communication

4.4.1 The Influence of ε

In the first step the initial node positions for all members (Figure 4.3) are established as follows: $NP(x) = 0.2$, $NP(y) = 0.2$, $NP(z) = 0.2$, $NP(u) = 0.2$, $NP(v) = 0.2$. The following values of ε have been taken into account: $\varepsilon = 0.1$, $\varepsilon = 0.5$, and $\varepsilon = 0.9$. For all cases the stop condition is the same, i.e. no difference in node position values with precision of 5 digits after the decimal point for all the members in two following iterations. The number of necessary iterations as well as the node position distribution has been studied in relation to ε (Table 4.1, Table 4.2, Table 4.3).

Table 4.1: Node position calculation for the network from Figure 4.3; $\varepsilon = 0.1$

$\varepsilon = 0.1$						
Iteration No.	**1**	**2**	$\cdots$	**7**	**8**	**9**
$NP(x)$	0.2	0.934	$\cdots$	1.06758	**1.06758**	**1.06758**
$NP(y)$	0.2	0.916	$\cdots$	0.97951	**0.97951**	**0.97951**
$NP(z)$	0.2	0.916	$\cdots$	0.98258	**0.98258**	**0.98258**
$NP(u)$	0.2	0.924	$\cdots$	1.02051	**1.02052**	**1.02052**
$NP(v)$	0.2	0.910	$\cdots$	0.94979	**0.94979**	**0.94979**

Table 4.2: Node position calculation for the network from Figure 4.3; $\varepsilon = 0.5$

$\varepsilon = 0.5$						
Iteration No.	**1**	**2**	$\cdots$	**19**	**20**	**21**
$NP(x)$	0.2	0.67	$\cdots$	1.30447	**1.30447**	**1.30447**
$NP(y)$	0.2	0.58	$\cdots$	0.88142	**0.88142**	**0.88142**
$NP(z)$	0.2	0.58	$\cdots$	0.96289	**0.96289**	**0.96289**
$NP(u)$	0.2	0.62	$\cdots$	1.10816	**1.10816**	**1.10816**
$NP(v)$	0.2	0.55	$\cdots$	0.74302	**0.74303**	**0.74303**

Table 4.3: Node position calculation for the network from Figure 4.3; $\varepsilon = 0.9$

$\varepsilon = 0.9$						
Iteration No.	**1**	**2**	$\cdots$	**120**	**121**	**122**
$NP(x)$	0.2	0.406	$\cdots$	1.51933	**1.51933**	**1.51933**
$NP(y)$	0.2	0.244	$\cdots$	0.74265	**0.74265**	**0.74265**
$NP(z)$	0.2	0.244	$\cdots$	1.02148	**1.02148**	**1.02148**
$NP(u)$	0.2	0.316	$\cdots$	1.19999	**1.20000**	**1.20000**
$NP(v)$	0.2	0.190	$\cdots$	0.51651	**0.51651**	**0.51651**

The data from Table 4.1, Table 4.2, and Table 4.3 provides the information about the influence of the coefficient ε value on the number of iterations that ought to be performed before the stop condition is fulfilled. It can be easilly noticed that the greater ε is, the more iterations must be performed (Table 4.4).

Node positions values for each user from the email–based community differ depending on the value of the coefficient ε, and are presented on the charts (Figure 4.4). Note that user x's node position is always the highest while user v possesses the lowest node position. The highest or lowest values for all the members are reached when $\varepsilon = 0.9$. Moreover, the node position of the individual is nearly linearly dependent on the value of ε (see regression lines in the Figure 4.4).

Table 4.4: The number of iterations in relation to the value of ε

ε	Number of iterations
0.1	9
0.5	21
0.9	122

All the calculations of the node positions are gathered together in Figure 4.5. If we would need to extract the best representative of the email–based community from Figure 4.3, then we always would select member x and next in order — member u. Note that they both have the greatest number of acquaintances — three, as compared to all the others who possess only two (Figure 4.3).

Some additional information about the influence of the coefficient ε onto the members' node positions provides the average node position within the human community and the standard deviation of the node position's value (Figure 4.6). When ε is greater, the distance between the minimum and maximum node position within community increases. The next conclusion is that the average node position does not depend on the value of ε. In all cases, it equals around 1 (Figure 4.6, Figure 4.7). However, the standard deviation differs depending on the coefficient ε value. The greater ε, the bigger standard deviation is. Furthermore, the dependence between the value of the coefficient ε and standard deviation is linear (Figure 4.7).

As it was presented, ε influences the number of iterations and the value of the distance between the members' node positions. The coefficient should be picked out very carefully. On the one hand the large number of calculations would slow down the process due to a big number of iterations. On the other hand too few iterations may cause the values of all node positions to be too close to each other.

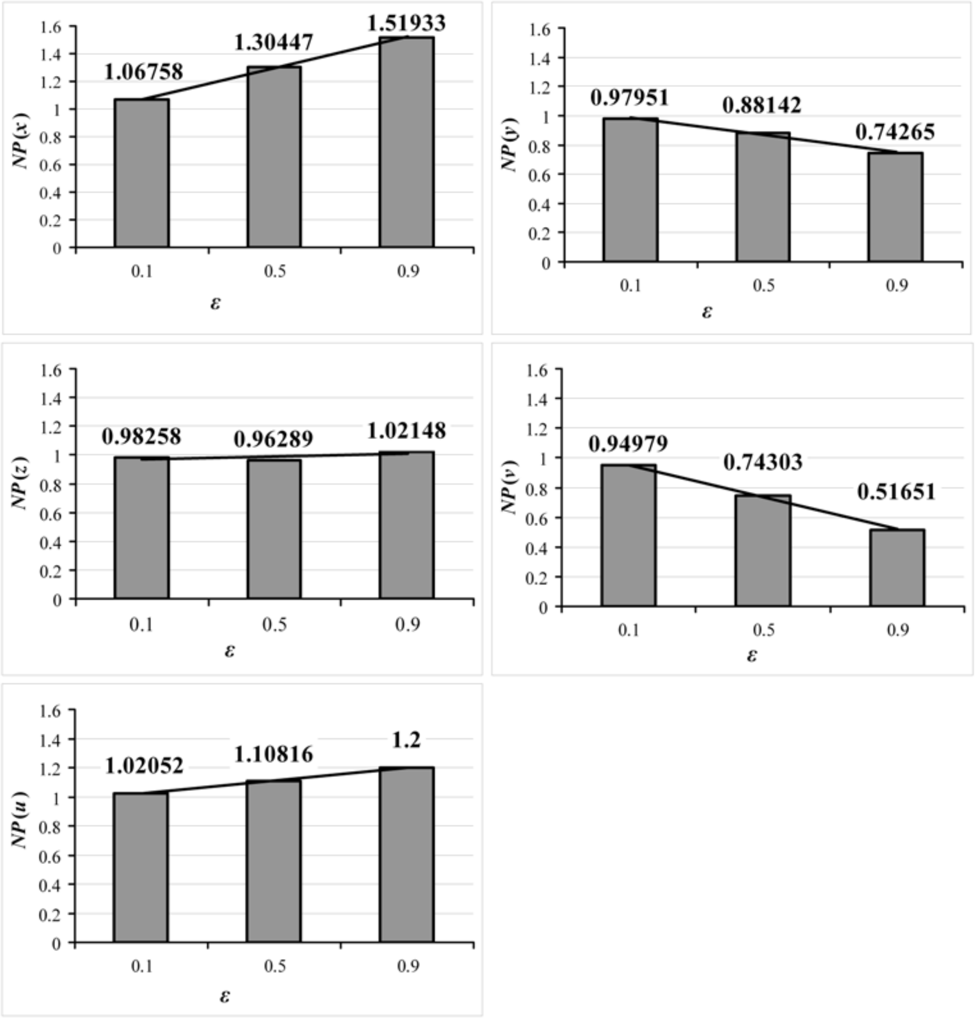


Figure 4.4: The values of members' node positions in relation to ε value

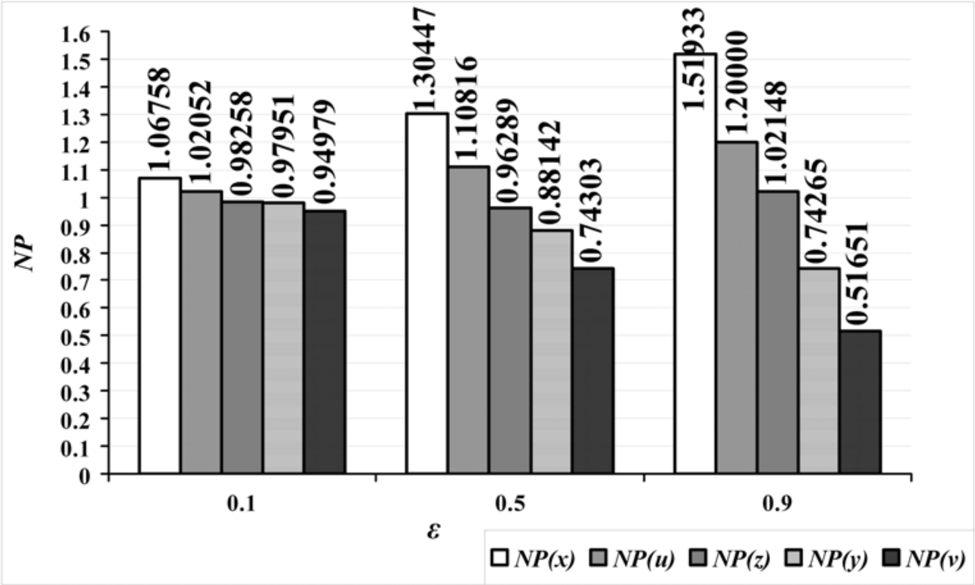


Figure 4.5: The value of social position in relation to ε

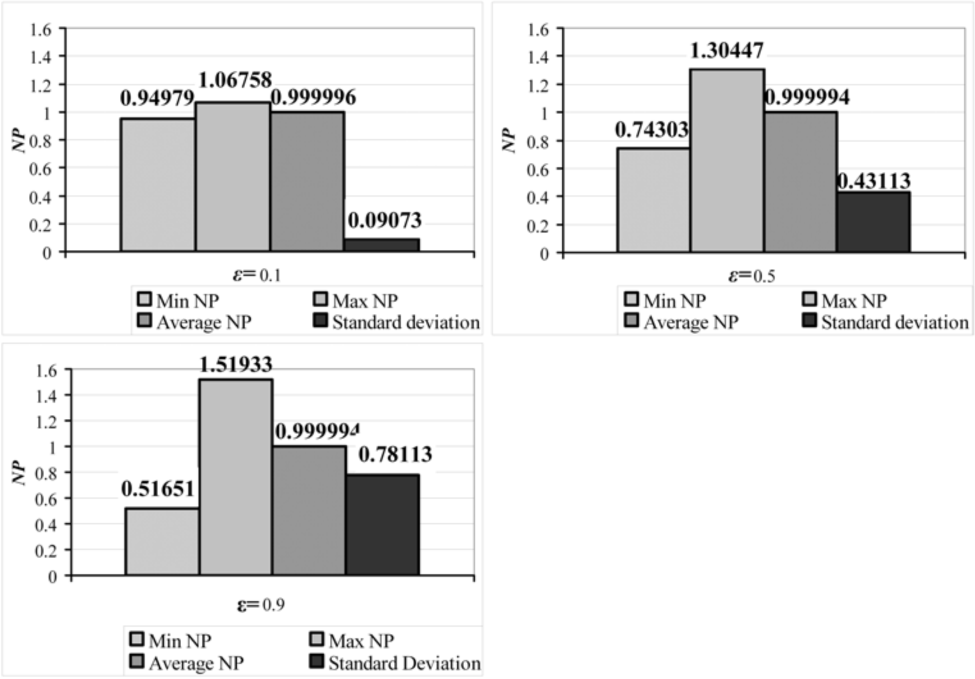


Figure 4.6: The minimum, maximum, average, and standard deviation of node position calculated for the same community but for different values of ε

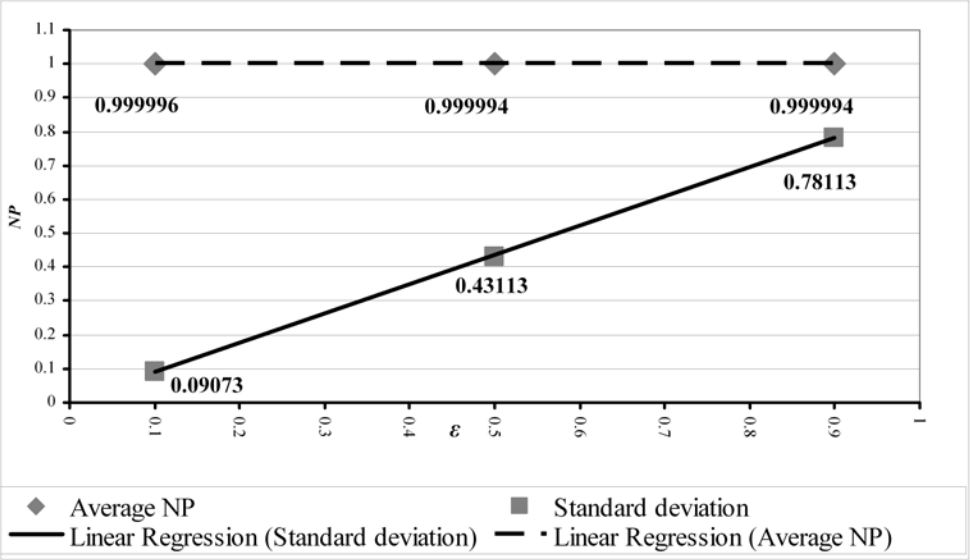


Figure 4.7: The linear regression for average node position and standard deviation

4.4.2 The Influence of Initial Node Positions

Another issue that is investigated is the influence of the initial node position values on the number of iterations that must be performed before the stop condition is met. Two groups of tests have been carried out for the network of Internet users in Figure 4.3 and for two different values of ε, i.e. $\varepsilon = 0.5$ and $\varepsilon = 0.9$. For each group, five different sets of initial values have been studied: $NP(x) = NP(y) = NP(z) = NP(u) = NP(v) = 0$ (case 1), $NP(x) = NP(y) = NP(z) = NP(u) = NP(v) = 0.2$ (case 3), $NP(x) = NP(y) = NP(z) = NP(u) = NP(v) = 1$ (case 4), $NP(x) = NP(y) = NP(z) = NP(u) = NP(v) = 3$ (case 5). In case 2, initial values have been assigned relatively close to the final ones (Figure 4.4): $NP(x) = 1.4$, $NP(y) = 0.8$, $NP(z) = 1$, $NP(u) = 1.1$, $NP(v) = 0.6$. The stop conditions are the same as in the previous calculations. This means that there is no difference in the node position values with precision to the 5th decimal place for all the members in two following iterations. The results of experiments are presented in Table 4.5, for $\varepsilon = 0.5$ and in Table 4.6, for $\varepsilon = 0.9$, respectively.

The first thing that should be emphasized is that the final value of the node position is not influenced by the initial values of node positions assignment. In all cases the final values of node positions are very similar. They are exactly the same with a precision to the 4th decimal place. In consequence, the representatives of the community from Figure 4.3, who will be selected based on their node position, would always be the same: member x and next member u; regardless of the initial values.

The smallest number of iterations is reached when the initial values of

Table 4.5: Node position calculation for different sets of their initial values ($\varepsilon = 0.5$)

Case	*NP(x)*	*NP(y)*	*NP(z)*	*NP(u)*	*NP(v)*	No. of iterations
Case 1	0	0	0	0	0	**21**
Case 2	1.4	0.8	1	1.1	0.6	**16**
Case 3	0.2	0.2	0.2	0.2	0.2	**21**
Case 4	1	1	1	1	1	**12**
Case 5	3	3	3	3	3	**20**

Table 4.6: Node position calculation for different sets of their initial values ($\varepsilon = 0.9$)

Case	*NP(x)*	*NP(y)*	*NP(z)*	*NP(u)*	*NP(v)*	No. of iterations
Case 1	0	0	0	0	0	**110**
Case 2	1.4	0.8	1	1.1	0.6	**65**
Case 3	0.2	0.2	0.2	0.2	0.2	**113**
Case 4	1	1	1	1	1	**21**
Case 5		3	3	3	3	**118**

node positions for every member equals 1 (case 4). Among all other cases the best result has been achieved when initial values are close to their final values (case 2). In other cases, the number of iterations is larger: between 20 and 21, for $\varepsilon = 0.5$ (Table 4.5) or at least 110, for $\varepsilon = 0.9$ (Table 4.6).

4.4.3 The Convergence of Node Position Method

The values of node position for every user after the following iterations, starting with 1 as the initial values of *NP* are shown in Figure 4.8. The calculations have been performed for $\varepsilon = 0.9$. The chart in Figure 4.8 reveals that the algorithm used for node position evaluation is convergent. Additionally, the algorithm tends to converge faster for smaller ε rather than for the greater ones (Figure 4.9).

The experiments on the network from Figure 4.3 show that the sum of all node positions is convergent to the number of nodes within the network, in this case — 5. This feature of the method was proved in Chapter 5, see Theorem 1. Two separate sets of initial node position values for the network from Figure 4.3 are presented in Table 4.7.

The pace of convergence both for individual members (Figure 4.9) and for their sum (Figure 4.10) tightly depends on ε value and it is greater for smaller ε.

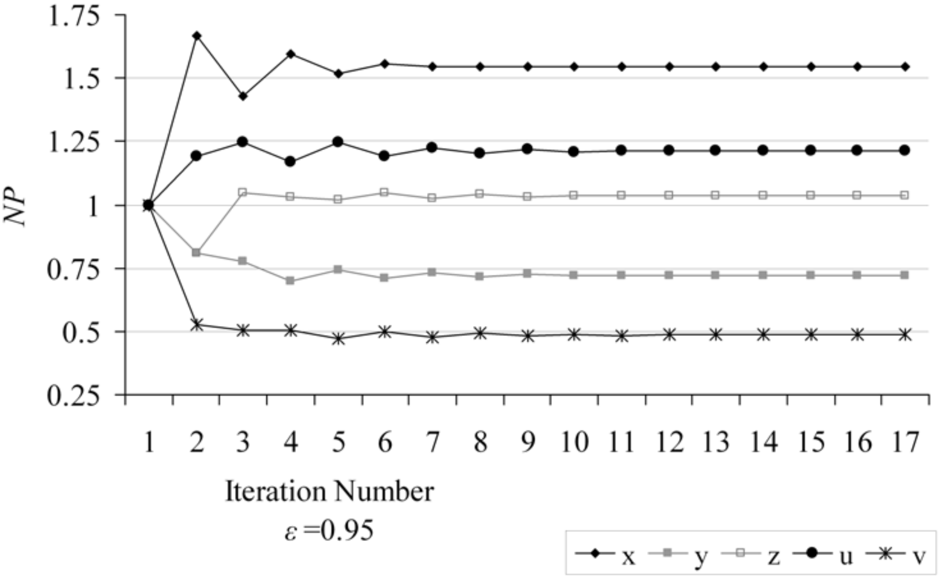


Figure 4.8: The values of node position after every iteration for $\varepsilon = 0.95$

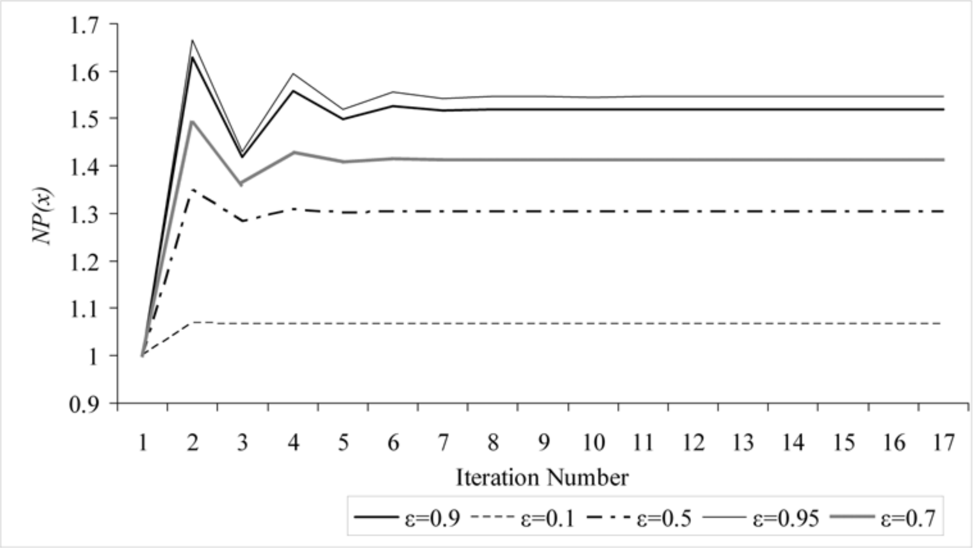


Figure 4.9: The values of the member x's node position after every iteration for various ε

Table 4.7: The initial node positions for set *NP_01* and *NP_02*

	NP_01	*NP_02*
$NP_0(x)$	0.5	2
$NP_0(y)$	0.1	2
$NP_0(z)$	0.9	2
$NP_0(u)$	0.5	2
$NP_0(v)$	0.4	2

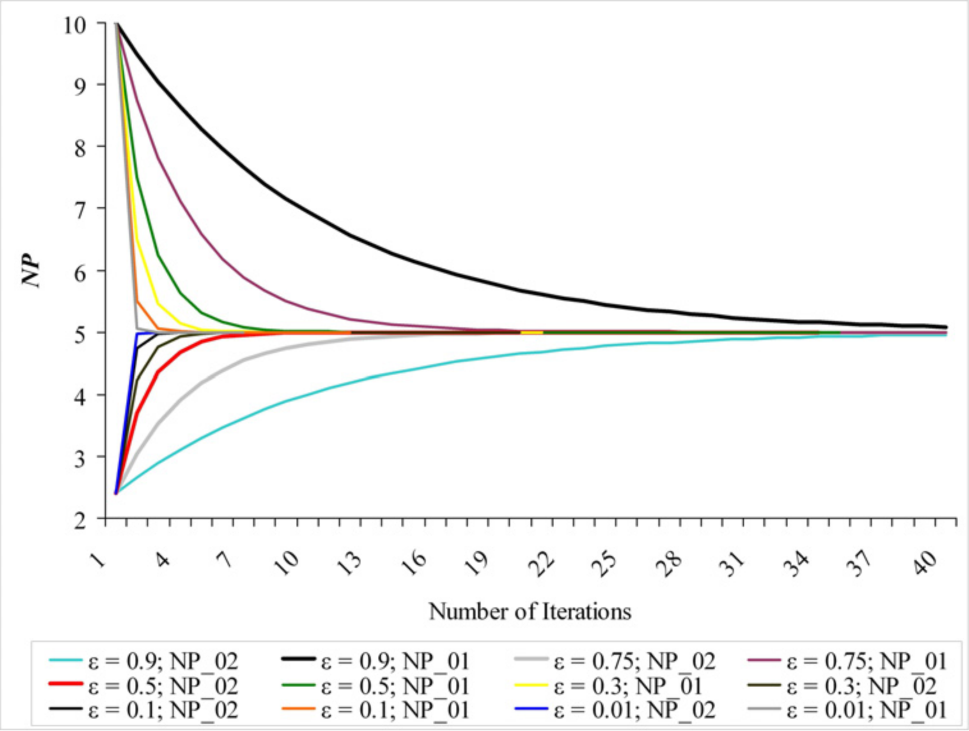


Figure 4.10: The convergence of the node position sum for various ε and various initial sums

4.4.4 Example — Conclusions

The goal of the analyzed example was to present the proposed node position method and its main features.

The research revealed that the method is convergent. The pace of convergence depends on ε and it is greater for smaller ε. This is formally proved in Chapter 5. Moreover, the dependence between pace of convergence and ε is best approximated by the exponential decay function (see Section 6.3). In consequence a value of ε influences number of iterations. For greater ε the required number of iterations increases. The dependence between ε value and the number of iterations is best approximated by the exponential growth function[2].

Another parameter that affects the number of iterations is a vector of initial node positions. The smallest number of iterations is needed when all initial positions equal 1. Additionally, a set of initial positions does not influence the final node position values.

Coefficient ε influences also the standard deviation of node position method. The greater ε, the bigger standard deviation and the dependence between these variables is linear. This results in rescaling the range of the node position values. For greater ε the wider range of node position values we have and this facilitates to distinguish the users.

[2] More deeper insight into this problem is presented in Section 6.3

Chapter 5

Formal Analysis of the Node Position Method

The node position is a new measure that possesses several interesting features. The most important ones, such as fixed limit of sum and fixed average value, convergence of node position iterative calculation as well as the main factors of this convergence [70], are presented in this chapter..

5.1 Total and Average Node Position

First, let us focus on the characteristic of the sum of all *NP*s within the network as well as the average value of node position measure within the entire network.

Lemma 1.
For every natural number n, we have:

$$SNP_{n+1} = m \cdot (1 - \varepsilon) + \varepsilon \cdot SNP_n,$$

where $SNP_n = \sum_{i=1}^{m} NP_n(x_i)$ is the sum of all node positions after the nth iteration; $m = card(IID)$.

Proof.
$SNP_{n+1} = \sum_{i=1}^{m} NP_{n+1}(x_i)$ and according to Equation 4.1:

$$SNP_{n+1} =$$

$$= \sum_{i=1}^{m}((1 - \varepsilon) + \varepsilon \cdot \sum_{j=1}^{m} NP_n(y_j) \cdot C(y_j, x_i)) =$$

$$= m \cdot (1 - \varepsilon) + \sum_{i=1}^{m} (\varepsilon \cdot \sum_{j=1}^{m} NP_n(y_j) \cdot C(y_j, x_i)) =$$

$$= m \cdot (1 - \varepsilon) + \varepsilon \cdot (\sum_{j=1}^{m} (NP_n(y_j) \cdot \sum_{i=1}^{m} C(y_j\, x_i)))$$

Note that:

$$\forall_{j \in m, j \neq 1} \sum_{i=1}^{m} C(y_j, x_i) = 1$$

This gives:

$$SNP_{n+1} = m \cdot (1 - \varepsilon) + \varepsilon \cdot (\sum_{j=1}^{m} (NP_n(y_j)))$$

Thus, $SNP_{n+1} = m \cdot (1 - \varepsilon) + \varepsilon \cdot SNP_n$. □

Lemma 2.
For every natural number n, we have:

$$SNP_n = m \cdot (1–\varepsilon^n) + \varepsilon^n \cdot SNP_0,$$

where $SNP_0 = \sum_{x \in IID} NP_0(x)$ is the sum of all initial node positions.

Proof.

i) For $n = 1$, we have $SNP_1 = m \cdot (1–\varepsilon) + \varepsilon \cdot SNP_0$ is true due to Lemma 1.

ii) Assume that the statement is true for $n = k$, i.e.:

$$SNP_k = m \cdot (1–\varepsilon^k) + \varepsilon^k \cdot SNP_0.$$

We want to prove it for $n = k + 1$, i.e.:

$$SNP_{k+1} = m \cdot (1–\varepsilon^{k+1}) + \varepsilon^{k+1} \cdot SNP_0.$$

Indeed, by Lemma 1:

$$SNP_{k+1} = m \cdot (1 - \varepsilon) + \varepsilon \cdot SNP_k =$$

$$= m \cdot (1–\varepsilon) + \varepsilon \cdot (m \cdot (1 - \varepsilon^k) + \varepsilon^k \cdot SNP_0) =$$

$$= m \cdot ((1–\varepsilon) + \varepsilon \cdot (1 - \varepsilon^k)) + \varepsilon \cdot \varepsilon^k \cdot SNP_0 =$$

$$= m \cdot (1 - \varepsilon^{k+1}) + \varepsilon^{k+1} \cdot SNP_0$$

According to mathematical induction the statement $SNP_n = m \cdot (1–\varepsilon^n) + \varepsilon^n \cdot SNP_0$ is true for all natural numbers n. □

Theorem 1.

i) For $\varepsilon \in (0;1)$, the sum of all node positions SNP in the social network of internet users $NIU = (IID, R)$ is convergent to the number of all members in the network: $\lim_{n\to\infty}(SNP_n) = m$. As a result, the average node position in the network is convergent to 1.

ii) For $\varepsilon = 1$ and all natural numbers n we have $SNP_n = SNP_0$, where $SNP_0 = \sum_{i=0}^{m} NP_0(x_i)$ is the sum of all initial node positions.

Proof.

i) From Lemma 2, for $\varepsilon \in (0;1)$, we have:

$$\lim_{n\to\infty}(SSP_n) = \lim_{n\to\infty}(m \cdot (1-\varepsilon^n) + \varepsilon^n \cdot SSP_0) = m$$

ii) From Lemma 2, for $\varepsilon = 1$, we have

$$SNP_n = m \cdot (1\text{–}\varepsilon^n) + \varepsilon^n \cdot SNP_0 = SNP_0$$

for every n. □

5.2 Convergence

Convergence is the essential feature of every iterative algorithm. In the considered approach, it regards both the sum of all node positions and the node position of each individual.

Theorem 2.
If $\varepsilon \in (0;1)$ and initial sum of node positions SNP_0 is different from limit m, i.e. $|SNP_0\text{–}m| > 0$, then the less the value of ε is the faster the sum of node positions is convergent to its limit $m = \mathit{card}(IID)$.

Proof.
The pace of convergence means: after how many iterations n the value of $|SNP_n\text{–}m|$ becomes less than the given error level τ, $0 < \tau < |SNP_0\text{–}m|$, i.e. when $|SNP_n\text{–}m| < \tau$.

From Lemma 2:

$$|SNP_n\text{–}m| =$$
$$= |m \cdot (1\text{–}\varepsilon^n) + \varepsilon^n \cdot SNP_0\text{–}m| =$$
$$= |\varepsilon^n \cdot SNP_0\text{–}m \cdot \varepsilon^n| =$$
$$= \varepsilon^n \cdot |SNP_0\text{–}m| < \tau.$$
$$\varepsilon^n < \frac{\tau}{|SNP_0 - m|} < 1.$$

$$n > \log_{\varepsilon} \frac{\tau}{|SNP_0 - m|} > 0.$$

The closer the ε is to 0, the smaller n is. □
Hence, we need more iterations, for larger ε, if we want more precise results (the lower error level τ). When the initial sum SNP_0 is further from its limit m, the number of required iterations also grows.

Lemma 3.
If the initial values of node positions are nonnegative, then node positions have a lower and upper limit after every iteration: $\forall_{(n>0)} \forall_x NP_n(x) \in [(1–\varepsilon); A]$, where $A = max(SNP_0, m)$.

Proof.

The lower limit
From Equation 4.1, by induction, we prove that $\forall_n NP_n(x) \geq 0$. Next, also from Equation 4.1 we have:

$$NP_{n+1}(x) = (1–\varepsilon) + \varepsilon \sum_{y \in IID} \underbrace{NP_n(y)}_{\geq 0} \cdot \underbrace{C(y,x)}_{\geq 0} \geq (1 - \varepsilon)$$

The upper limit
Since all initial values of node positions are nonnegative then $SNP_0 \geq 0$. From Lemma 2:

$$SNP_n = m \cdot (1–\varepsilon^n) + \varepsilon^n \cdot SNP_0 \leq$$
$$\leq max(m, SNP_0) \cdot (1–\varepsilon^n) + \varepsilon^n \cdot max(m, SNP_0) =$$
$$= max(m, SNP_0) = A.$$

$SNP_n \leq A$ so any from the nonnegative components of SNP_n, i.e. $NP_n(x)$, must not exceed A.
Hence, $NP_n(x) \in [(1–\varepsilon); A]$. □
Lemma 3 reveals that there is a fixed lower and upper limit for every node position value, after every iteration and these limits are independent of iteration n.

Lemma 4.
If initial values of node positions are nonnegative, then $|NP_{n+k+1}(x) - NP_{n+k}(x)| \leq \varepsilon^k \cdot A \cdot m$.

Proof.
Based on Equation 4.1:

$$|NP_{n+k+1}(x) - NP_{n+k}(x)| =$$

$$\varepsilon \cdot | \sum_{y_1 \in IID} (NP_{(n+k+1)-1}(y_1) - NP_{(n+k)-1}(y_1)) \cdot C(y_1, x)| \leq$$

$$\leq \varepsilon \cdot \sum_{y_1 \in IID} |(NP_{(n+k+1)-1}(y_1) - NP_{(n+k)-1}(y_1))| \cdot C(y_1, x) \leq$$

$$\leq \varepsilon^2 \cdot \sum_{y_1 \in IID} \sum_{y_2 \in IID} |NP_{(n+k+1)-2}(y_2) - NP_{(n+k)-2}(y_2)| \cdot C(y_1, x) \cdot C(y_2, y_1) \leq$$

$$\leq \cdots \leq$$

$$\leq \varepsilon^k \sum_{y_1 \in IID} \sum_{y_2 \in IID} \cdots \sum_{y_k \in IID} |NP_{(n+k+1)-k}(y_k) - NP_{(n+k)-k}(y_k)| \cdot$$

$$\cdot C(y_1, x) \cdot C(y_2, y_1) \cdot \cdots \cdot C(y_k, y_{k-1}) \leq$$

$$\leq \varepsilon^k \cdot \sum_{y_k \in IID} \underbrace{|NP_{(n+k+1)-k}(y_k) - NP_{(n+k)-k}(y_k)|}_{\leq A, Lemma3} \cdot \underbrace{C(y_1, x)}_{\leq 1} \leq$$

$$\leq \varepsilon^k \cdot A \cdot m.$$

The last but one inequality results from $\sum_{y_{k-1} \in IID} C(y_k, y_{k-1}) = 1$ that has been applied $k - 1$ times. □

Note that according to Lemma 4 the pace of convergence depends on both the value of ε and the number of network members. It means that for smaller ε and smaller networks the fixed difference between two consecutive iterations will be reached faster than for larger ε and larger networks. The similar conclusion can be drawn from Theorem 2 in relation to the sum of all node positions.

Theorem 3.
If initial values of node positions are nonnegative, then their calculation based on Equation 4.1 is convergent, that means $\exists NP(x) = lim_{n \to \infty}(NP_n(x))$.

Proof.
By Lemma 4.

$$|NP_{n+k+l}(x) - NP_{n+k}(x)| \leq$$

$$\leq |NP_{n+k+l}(x) - NP_{n+k+l-1}(x)| + |NP_{n+k+l-1}(x) - NP_{n+k+l-2}(x)| + \cdots +$$

$$+ \cdots + |NP_{n+k+1}(x) - NP_{n+k}(x)| \leq$$

$$\leq A \cdot m \cdot (\varepsilon^{k+l-1} + \varepsilon^{k+l-2} + \cdots + \varepsilon^k) =$$

$$= A \cdot m \cdot \varepsilon^k \cdot (1 + \varepsilon + \varepsilon^2 + \cdots + \varepsilon^{l-1}) =$$

$$= A \cdot m \cdot \varepsilon^k \cdot \frac{1 - \varepsilon^l}{1 - \varepsilon} \xrightarrow{k \to \infty} 0$$

Hence, $NP_n(x)$ is a Cauchy sequence, therefore it is convergent. □

Theorem 3 assumes nonnegative values of initial node position. However, similar reasoning can also be performed for negative values. In this case each value should be split into nonnegative and negative parts. From a practical point of view, the assignment of negative initial values to node position appears to be useless since the final node positions are positive and this can only increase the number of necessary iterations.

Another approach to the proof of convergence based on the concept of power series, to a similar problem, i.e PageRank was presented in [23].

5.3 Interval of Limit Values

Regardless of the initial node position values, their limit values have to be from the range of the specific interval.

Theorem 4.
The limit value of node position does not exceed half of the number of members:

$$\forall(x \in IID) \lim_{n\to\infty}(NP_n(x)) \leq \frac{m}{2}.$$

The member x will have the greatest node position if all other members pass the whole of their commitment to the person x, i.e. $\forall(y \in IID, y \neq x)C(y,x) = 1$, and the member x's commitment will be spread among all x's acquaintances, i.e. $\forall(y \in IID, y \neq x)C(x,y) > 0$ (Figure 5.1). Moreover, it is not important how the central member x's commitment is distributed. In other words, member x reaches the greatest node position if member x gathers all commitments from all members y in the social network, i.e. fully inherits node positions of all y.

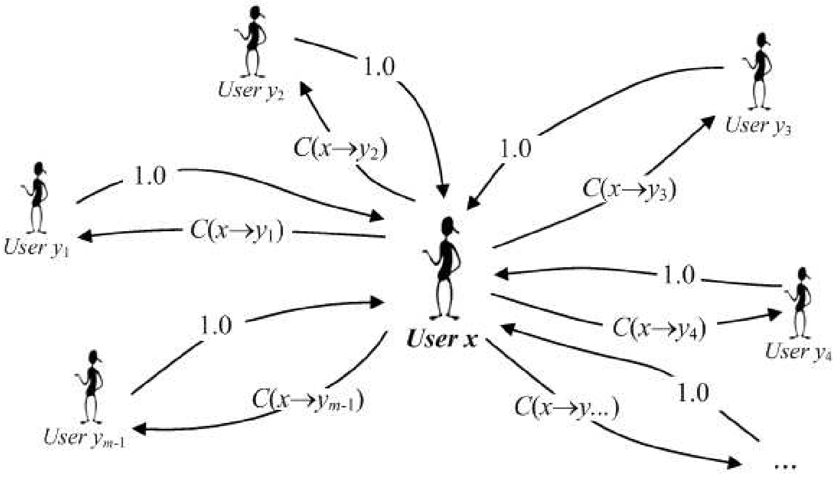


Figure 5.1: The community where individual x has the greatest node position

Proof.

$$NP_{max}(x) = (1-\varepsilon) + \varepsilon \cdot \sum_{i=1}^{m-1} NP(y_i) \cdot C(y_i, x),$$

where $NP_{max}(x)$ is the maximum value of limit $\lim_{n\to\infty}(NP_n(x))$.

$$NP_{max}(x) = (1-\varepsilon) + \varepsilon \cdot \sum_{i=1}^{m-1}((1-\varepsilon) + \varepsilon \cdot NP_{max}(x) \cdot C(x, y_i)) \cdot \underbrace{C(y_i, x)}_{1}.$$

$$NP_{max}(x) = (1-\varepsilon) + \varepsilon \cdot ((m-1) \cdot (1-\varepsilon) + \varepsilon \cdot NP_{max}(x) \cdot \underbrace{\sum_{i=1}^{m-1} C(x, y_i)}_{1}).$$

$$NP_{max}(x) = (1-\varepsilon) + \varepsilon \cdot ((m-1) \cdot (1-\varepsilon) + \varepsilon \cdot NP_{max}(x)).$$

$$NP_{max}(x) = (1-\varepsilon) + \varepsilon \cdot (m - m \cdot \varepsilon - 1 + \varepsilon + \varepsilon \cdot NP_{max}(x)).$$

$$NP_{max}(x) = 1 - \varepsilon + \varepsilon \cdot m - \varepsilon^2 \cdot m - \varepsilon + \varepsilon^2 + \varepsilon^2 \cdot NP_{max}(x).$$

$$(1-\varepsilon^2) \cdot NP_{max}(x) = (1-\varepsilon) + m \cdot \varepsilon \cdot (1-\varepsilon) - \varepsilon \cdot (1-\varepsilon).$$

$$(1-\varepsilon^2) \cdot NP_{max}(x) = (1-\varepsilon) \cdot (1 + m \cdot \varepsilon - \varepsilon).$$

$$NP_{max}(x) = \frac{(1-\varepsilon) \cdot (1 + m \cdot \varepsilon - \varepsilon)}{(1-\varepsilon) \cdot (1+\varepsilon)}.$$

$$NP_{max}(x) = \frac{1 - \varepsilon + m \cdot \varepsilon}{1+\varepsilon}$$

The node position is maximum when the function $f(\varepsilon) = \frac{1-\varepsilon+m\cdot\varepsilon}{1+\varepsilon}$ reaches its maximum value. The domain of this function is $\varepsilon = (0; 1]$ and $m \geq 2$. This is the constraint derived from the formula that serves to calculate the node position of the member of the community. First, the monotonicity of the function $f(\varepsilon)$ is studied. This function is non–decreasing, which is proved below.

A function $f(x)$ is said to be non–decreasing in an interval I if $f(b) \geq f(a)$ for all $b > a$, where a, $b \in I$. [63].

$$\varepsilon_2 - \varepsilon_1 > 0 \Rightarrow f(\varepsilon_2) - f(\varepsilon_1) \geq 0.$$

$$f(\varepsilon_2) - f(\varepsilon_1) = \frac{1 - \varepsilon_2 + \varepsilon_2 \cdot m}{1 + \varepsilon_2} - \frac{1 - \varepsilon_1 + \varepsilon_1 \cdot m}{1 + \varepsilon_1}.$$

$$f(\varepsilon_2) - f(\varepsilon_1) = \frac{(1 - \varepsilon_2 + \varepsilon_2 \cdot m) \cdot (1 + \varepsilon_1) - (1 - \varepsilon_1 + \varepsilon_1 \cdot m) \cdot (1 + \varepsilon_2)}{(1 + \varepsilon_2) \cdot (1 + \varepsilon_1)}.$$

$$f(\varepsilon_2) - f(\varepsilon_1) = \frac{1 - \varepsilon_1 \cdot \varepsilon_2 - \varepsilon_2 + \varepsilon_1 + \varepsilon_2 \cdot m + \varepsilon_1 \cdot \varepsilon_2 \cdot m}{(1 + \varepsilon_2) \cdot (1 + \varepsilon_1)} +$$

$$+ \frac{-1 - \varepsilon_2 + \varepsilon_1 + \varepsilon_1 \cdot \varepsilon_2 - \varepsilon_1 \cdot m - \varepsilon_1 \cdot \varepsilon_2 \cdot m}{(1 + \varepsilon_2) \cdot (1 + \varepsilon_1)}.$$

$$f(\varepsilon_2) - f(\varepsilon_1) = \frac{-2 \cdot \varepsilon_2 + 2 \cdot \varepsilon_1 + \varepsilon_2 \cdot m - \varepsilon_1 \cdot m}{(1 + \varepsilon_2) \cdot (1 + \varepsilon_1)}.$$

$$f(\varepsilon_2) - f(\varepsilon_1) = \frac{-\overbrace{(2 - m)}^{<0\forall(m\geq 2)} \cdot \overbrace{(\varepsilon_2 - \varepsilon_1)}^{>0}}{\underbrace{(1 + \varepsilon_2) \cdot (1 + \varepsilon_1)}_{>0\forall(\varepsilon>0)}}.$$

$$\forall_{\varepsilon\in[0;1)} f(\varepsilon_2) - f(\varepsilon_1) \geq 0.$$

This means that the function $f(\varepsilon)$ is non–decreasing, so it reaches the maximum value for $\varepsilon = 1$ and then $f(\varepsilon) = \frac{m}{2}$. This leads to the conclusion that $NP_{max}(x) = \frac{m}{2}$ (Figure 5.2). □

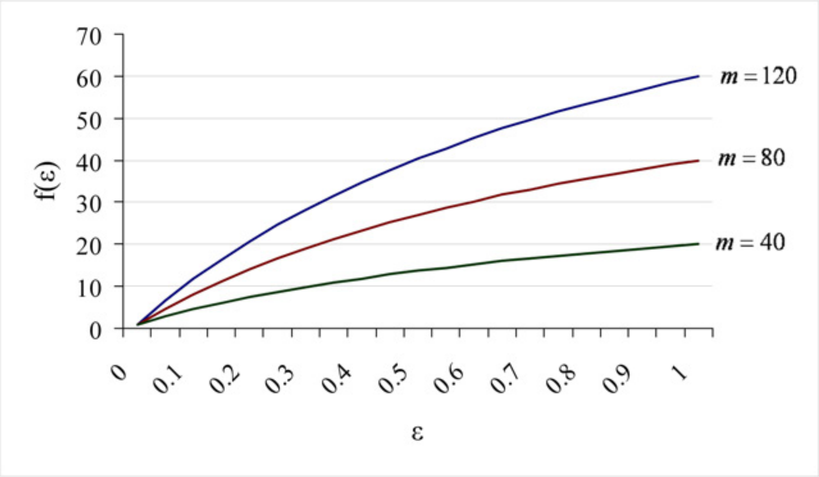


Figure 5.2: The chart of the function $f(\varepsilon)$ for the network that consists of m Internet users

The interval of node position depends on the number of members m within the network and the value of the coefficient ε (Figure 5.3). In general, the limit value of node position is from the range $[1 - \varepsilon, \frac{1-\varepsilon+\varepsilon \cdot m}{1+\varepsilon}]$, see also Lemma 3. The maximum value of the node position is reached for $\varepsilon = 1$ and in such cases node position equals $\frac{m}{2}$, where m is the number of members within the community.

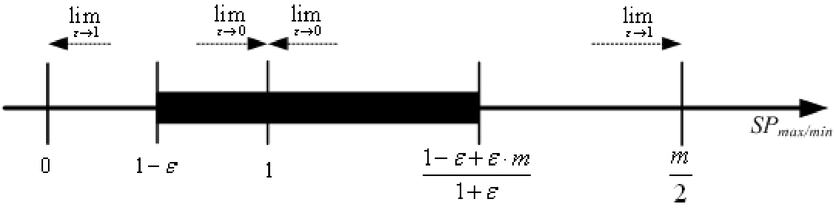

$\lim_{\varepsilon\to 1}$
$\lim_{\varepsilon\to 0}$
$\lim_{\varepsilon\to 0}$
$\lim_{\varepsilon\to 1}$
0
$1-\varepsilon$
1
$\frac{1-\varepsilon+\varepsilon\cdot m}{1+\varepsilon}$
$\frac{m}{2}$
$SP_{max/min}$

Figure 5.3: The range of the node position values

Chapter 6

Research

On one hand, the research was conducted in order to present the main features of the proposed node position method and on the other hand to compare it with other methods that are utilized to assess the centrality of a person within the network of Internet users that are one of the subsets of the complex network systems. Four datasets, which come from the real world, were used to investigate the proposed method:

- Thurman Network
- Enron Network
- Wroclaw University of Technology Network
- Telecommunication Network

The first dataset — Thurman Network — serves to present the method and what is more important to compare it with other centrality measures (Section 6.1).

The goal of the research on the Enron Network is to present the characteristic features of the proposed node position method, i.e. the distribution of node position values, its minimum, maximum as well as average values. Moreover, two different types of commitment function calculation have been proposed — the first one that does not take into consideration time and the second one that does. In consequence, the node position values without time factor and with time factor for all network nodes have been calculated and analyzed. (Section 6.2).

The network extracted from the email logs from the Wroclaw University of Technology email server is utilized to present the influence of method's parameters: ε coefficient and stop condition τ on the processing time, number of required iterations as well as on the number of distinct values of node position values (Section 6.3).

Finally, the telecommunication data serves to present the influence of ε coefficient on the processing time. (Section 6.4)

All datasets had to be cleansed and carefully prepared before the experiments were launched. After the cleansing process, from the obtained data the networks of users were extracted. More detailed description of networks creation process is provided in the next sections.

6.1 Thurman Network

The experiments on Thurman Network were carried out in order to present the method and compare its characteristics with other centrality measures [68], [70]. The measures that are utilized in the comparison process are outdegree centrality (*ODC*), closeness centrality (*CC*), indegree centrality (*IDC*), proximity prestige (*PP*), and two eigenvector methods i.e. *ODC Eigenvector — ODCE* and *PP Eigenvector — PPE*. The chosen methods represent each of the groups presented in Chapter 3.2.

6.1.1 Data Description and Preparation

The Thurman office social network is a non–symmetrical network of 15 people who worked in one company (Figure 6.1) [113]. The adjacency matrix for the Thurman network is presented in Table 6.1 where non–zero values represent the existence of the connection between two users and their values correspond to the values of commitment function i.e. strength of relationships. These non–zero values have equaled one in the original adjacency matrix.

In order to obtain the values of commitment function for each individual, value one — from the original matrix — is divided by the number of member's relationships, e.g. Emma communicates with nine members so her contribution of activity to each of her acquaintances equals $\frac{1}{9}$. The outcomes of these calculations are presented in Table 6.1.

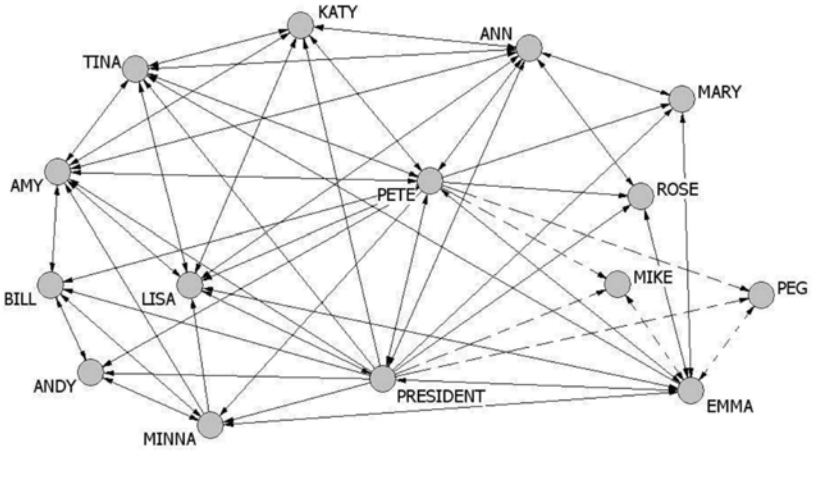


Figure 6.1: Graph representation of the classic Thurman network

Table 6.1: The values of commitment function within the Thurman network

Member	**12**	**1**	**5**	**2**	**8**	**6**	**3**	**10**	**4**	**9**	**7**	**11**	**13**	**14**	**15**
12.Emma	–	–	$\frac{1}{9}$	–	$\frac{1}{9}$	$\frac{1}{9}$	–	$\frac{1}{9}$	–	$\frac{1}{9}$	–	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$
1.Ann	–	–	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	–	–	$\frac{1}{8}$	–	$\frac{1}{8}$	$\frac{1}{8}$	–	–
5.Pete	$\frac{1}{14}$	$\frac{1}{14}$	–	$\frac{1}{14}$	$\frac{1}{14}$	$\frac{1}{14}$	$\frac{1}{14}$	$\frac{1}{14}$	$\frac{1}{14}$	$\frac{1}{14}$	$\frac{1}{14}$	$\frac{1}{14}$	$\frac{1}{14}$	$\frac{1}{14}$	$\frac{1}{14}$
2.Amy	–	$\frac{1}{6}$	$\frac{1}{6}$	–	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	–	$\frac{1}{6}$	–	–	–	–	–	–
8.Lisa	$\frac{1}{7}$	$\frac{1}{7}$	$\frac{1}{7}$	$\frac{1}{7}$	–	$\frac{1}{7}$	$\frac{1}{7}$	–	–	$\frac{1}{7}$	–	–	–	–	–
6.Tina	–	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	–	$\frac{1}{5}$	–	–	–	–	–	–	–	–
3.Katy	–	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	–	–	–	–	–	–	–	–	–
10.Minna	$\frac{1}{5}$	–	–	$\frac{1}{5}$	$\frac{1}{5}$	–	–	–	$\frac{1}{5}$	–	$\frac{1}{5}$	–	–	–	–
4.Bill	–	–	–	$\frac{1}{3}$	–	–	–	$\frac{1}{3}$	–	–	$\frac{1}{3}$	–	–	–	–
9.President	$\frac{1}{14}$	$\frac{1}{14}$	$\frac{1}{14}$	$\frac{1}{14}$	$\frac{1}{14}$	$\frac{1}{14}$	$\frac{1}{14}$	$\frac{1}{14}$	$\frac{1}{14}$	–	$\frac{1}{14}$	$\frac{1}{14}$	$\frac{1}{14}$	$\frac{1}{14}$	$\frac{1}{14}$
7.Andy	–	–	$\frac{1}{3}$	–	–	–	–	$\frac{1}{3}$	$\frac{1}{3}$	–	–	–	–	–	–
11.Mary	$\frac{1}{2}$	$\frac{1}{2}$	–	–	–	–	–	–	–	–	–	–	–	–	–
13.Rose	$\frac{1}{2}$	$\frac{1}{2}$	–	–	–	–	–	–	–	–	–	–	–	–	–
14.Mike	1	–	–	–	–	–	–	–	–	–	–	–	–	–	–
15. Peg	1	–	–	–	–	–	–	–	–	–	–	–	–	–	–

6.1.2 Comparison of Node Position with Other Centrality Methods

Before starting the experiments some assumptions were made. The initial node positions $NP_0(x) = 1$ are established for every member x in the network. The value of ε is 0.9 and the stoping condition is: no difference in node position values to the precision of 5 digits after the decimal point for all the members in two following iterations, i.e. $\tau = 0.00001$.

The other measures: outdegree centrality (ODC), closeness centrality (CC), indegree centrality (IDC), and proximity prestige (PP) have been calculated according to the appropriate formulas (see Chapter 3.2). They are compared with NP in Figure 6.2 and Figure 6.3. Centrality based on eigenvectors is calculated using two different initial, input centralities: outdegree centrality (ODC Eigenvector) and proximity prestige (PP Eigenvector). After that the eigenvector measure is compared to NP (Figure 6.4). The outcomes of the calculations were gathered and presented in Table 6.2

Table 6.2: The values of the analyzed measures in the Thurman office social network for each member of the network

Member	NP	ODC	CC	IDC	PP	$ODCE$	PPE
EMMA	1.90025	0.64286	0.73684	0.57143	0.70000	0.57143	1.94866
ANN	1.56732	0.57143	0.70000	0.57143	0.63636	0.57143	1.10606
PETE	1.48140	1.00000	1.00000	0.57143	0.70000	0.57143	0.82005
AMY	1.38236	0.42857	0.63636	0.57143	0.63636	0.57143	0.81412
LISA	1.36532	0.50000	0.66667	0.57143	0.70000	0.57143	0.72512
TINA	1.17424	0.35714	0.60870	0.50000	0.66667	0.50000	0.57005
KATY	1.01320	0.35714	0.60870	0.42857	0.58333	0.42857	0.50894
MINNA	0.86255	0.35714	0.60870	0.35714	0.60870	0.35714	0.50728
BILL	0.79626	0.21429	0.46667	0.35714	0.51852	0.35714	0.48447
PRESIDENT	0.73712	1.00000	1.00000	0.28571	0.56000	0.28571	0.30732
ANDY	0.63676	0.21429	0.56000	0.28571	0.50000	0.28571	0.38458
MARY	0.60897	0.14286	0.50000	0.28571	0.56000	0.28571	0.24732
ROSE	0.60897	0.14286	0.50000	0.28571	0.56000	0.28571	0.24732
MIKE	0.43264	0.07143	0.43750	0.21429	0.53846	0.21429	0.16778
PEG	0.43264	0.07143	0.43750	0.21429	0.53846	0.21429	0.16778
Characteristics							
Min.	0.43264	0.07143	0.43750	0.21429	0.50000	0.21429	0.16778
Max.	1.90025	1.00000	1.00000	0.57143	0.70000	0.57143	1.94866
Std. Dev.	0.45397	0.29779	0.17573	0.14200	0.06890	0.14200	0.46063
Number of different values	13	9	9	6	9	6	13
Duplicates [%]	13	40	40	60	40	60	13

Based on the obtained values, seven separate rankings have been created. The positions of each member in every ranking are presented in Table 6.3. Note that the order of people based on their node position varies a lot from

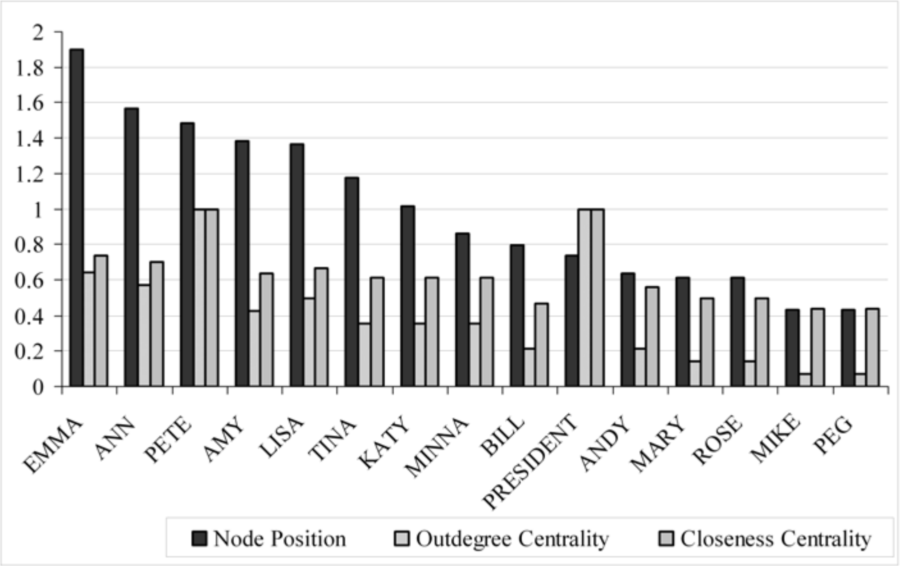


Figure 6.2: The comparison of centrality measures to node position in the Thurman network

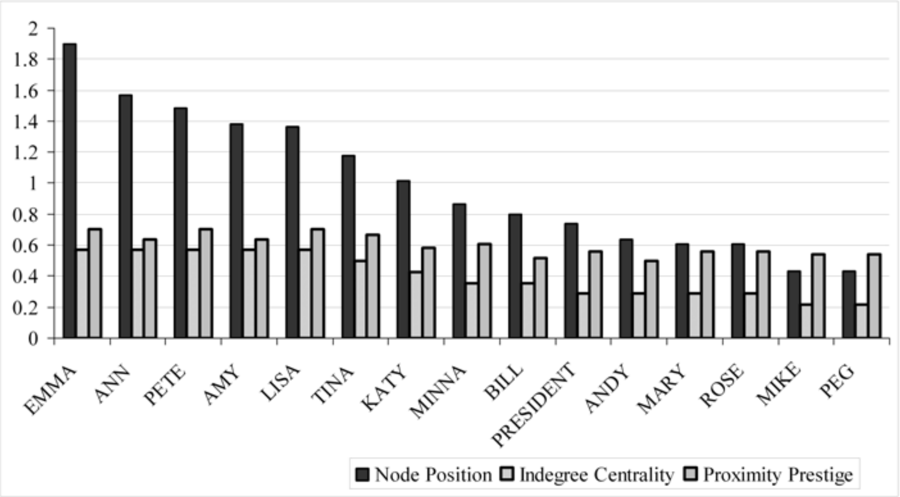


Figure 6.3: The comparison of prestige measures to node position in the Thurman network

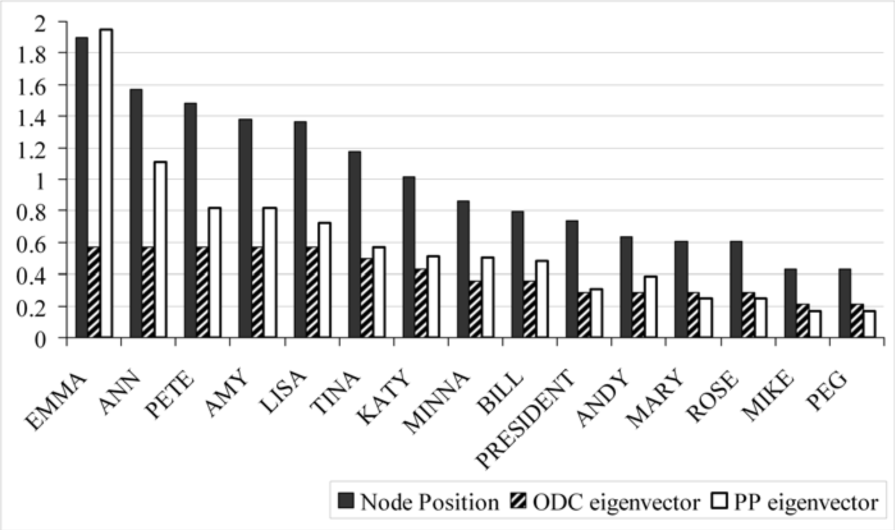


Figure 6.4: The comparison of eigenvector measures with node position in the Thurman network

the rankings obtained based on the outdegree centrality — *ODC* and closeness centrality — *CC*. On the other hand, the rankings of members based on their node position and indegree centrality are quite similar, even though the distribution of node position is greater. **Node position provides a better opportunity to distinguish individuals within the network as opposed to other centrality measures** (see % of duplicates in Table 6.2). The information that Emma, Ann, Pete, Amy, and Lisa have the same, greatest indegree centrality is insignificant since it results from the number of other members who are adjacent to these people. Only in the case of *PPE* measure the number of duplicates is the same as in *NP* measure. However, processing time for *PPE* method is longer than for *NP* because it requires first to calculate the *PP* for each node and after that application of the eigenvector method. *PP* is the measure based on the shortest paths problem and as it was presented in Section 6.4 in complex networks these methods are inefficient and resource consuming.

The node position measure $NP(x)$ takes into consideration not only the number of members who communicate to the evaluated person x but also their node positions and their contribution of activity directed to x. Based on these properties we can observe that Emma is the person with the highest node position in the network because Mike and Peg communicate only with Emma so they transfer their entire node positions to her. In other words, if Emma leaves the network, then it will split into smaller groups. This is the basis to claim that users with high *NP* tend to connect users with small number of connections i.e. these who are at the periphery of the network with the entire network. These users with high position can be treated as bridge that connects different subgroups of the network. The prestige measures do not respect these features and this appears to be critical in assessing the importance of an individual in the social network.

Table 6.3: The positions in rankings for the analyzed measures in the Thurman office social network, for $\varepsilon = 0.9$

Member	NP	ODC	CC	IDC	PP	$ODCE$	PPE
Emma	1	3	3	1	1	1	1
Ann	2	4	4	1	5	1	2
Pete	3	1	1	1	1	1	3
Amy	4	6	6	1	5	1	4
Lisa	5	5	5	1	1	1	5
Tina	6	7	7	6	4	6	6
Katy	7	7	7	7	8	7	7
Minna	8	7	7	8	7	8	8
Bill	9	10	13	8	15	8	9
President	10	1	1	10	9	10	11
Andy	11	10	10	10	12	10	10
Mary	12	12	11	10	9	10	12
Rose	12	12	11	10	9	10	12
Mike	14	14	14	14	13	14	14
Peg	14	14	14	14	13	14	14

Based on the comparison of the node position with both eigenvector measures we can observe that the rankings are very similar. However, in the case of *ODC* Eigenvector, five people occupy the first position but the structures of their connections differ a lot. For example Emma, Amy, Lisa and Ann occupy the same first position for *ODCE* but while analyzing the structure of their networks it can be noticed that Emma (the highest NP) plays important role in bridging between Mike and Peg and the rest of network whereas Amy, Lisa and Ann do not. Pete's network also differs in comparison to others' four members who occupy first position for *ODCE* measures. Although he communicates with each user, not all of the members communicate with him (8 people). It is the reason why he occupies the 3rd position in node position ranking, i.e. the node position measure promotes users with whom many users communicate, not these who just send messages to others and never receive the answer. Note that, in the eigenvector–like measures the influence on the user position have only the members who directly communicate with a given person. On the contrary, the NP value of an individual depends on NPs of all members in the network due to a recursive character of this measure. Moreover, the final eigenvector centrality varies depending on the method that was used to evaluate the initial centralities (see Figure 6.4), whereas the initial values of NP — according to the carried out experiments — do not influence their final, limit values. Furthermore, the node position measure also respects the strength of each relationship in the form of commitment in activity and it is an important element of its calculation. The commitment function is individualized for each relationship and it reflects the real contribution in activity directed from one person to another. In the eigenvector approach, we have only row normalization of the adjacency matrix and the individualized values of personal initial centralities.

Next part of the experiments was conducted in order to compare the rankings created upon different centrality indices. To compare rankings created upon different measures the Kendall's coefficient of concordance was used.

For each pair of the rankings from Table 6.3 Kendall's coefficient was calculated and the results are presented in Table 6.4 and Figure 6.5. Note that rankings based on the neighbors directly connected to the given one (incoming relationships), i.e. IDC, PP Eigenvector, and NP are in pairs very close; the range of Kendall's coefficient is from 0.828 up to 1 between IDC and ODC Eigenvector. Similarly, the rankings based on measures that take into account outgoing relationships ODC and CC are alike – Kendall's coefficient at the relatively high level of 0.895.

Table 6.4: Kendall's coefficient for each pair of rankings from Table 6.3

	NP	ODC	CC	IDC	PP	$ODCE$
ODC	0.724	-	-	-	-	-
CC	0.676	0.895	-	-	-	-
IDC	0.829	0.629	0.581	-	-	-
PP	0.657	0.571	0.619	0.657	-	-
$ODCE$	0.828	0.628	0.581	1	0.657	-
PPE	0.962	0.709	0.652	0.828	0.681	0.828

For that reason, the ranking based on node position (NP) is most similar to the rankings based on PP Eigenvector, ODC Eigenvector and indegree centrality (IDC). On the contrary, the node position ranking is least similar to the ranking of the proximity prestige (PP) and closeness centrality (CC) measures .

The rankings based on PP Eigenvector and ODC Eigenvector are more similar to other rankings rather than to each other. Kendall's coefficient for PP Eigenvector and ODC Eigenvector equals 0.828 whereas for ODC Eigenvector and IDC is greater and equals 1 and for PP Eigenvector and NP is 0.962. This leads to the conclusion that the initial centralities (ODC and PP) highly influence the final values based on eigenvectors.

6.2 Enron Network

6.2.1 Data Description and Preparation

The Enron dataset consists of the employees' mail boxes. Enron Corporation was the biggest energy company in the United States of America. It employed around 21,000 people before its bankruptcy at the end of 2001. A number of

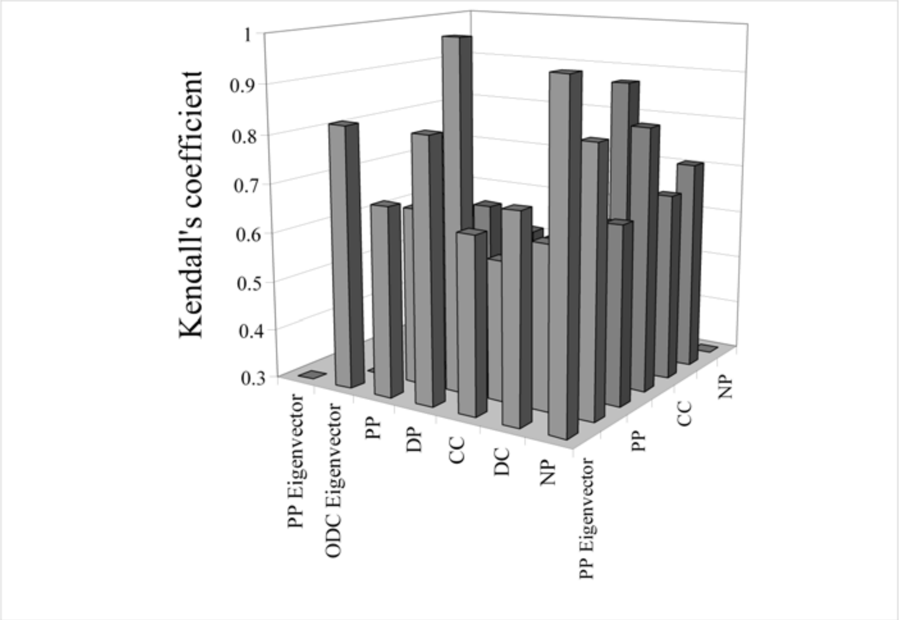


Figure 6.5: The values of Kendall's coefficient for the pairs of rankings from Table 6.4

other researches have been conducted on the Enron email dataset [88], [102], [112]. As it was mentioned before, the main aim of this part of experiments is to present the general characteristics of the method, i.e. the distribution of node position values, its minimum, maximum as well as average values. Additionally, the analysis of node position (NP) and node position with time factor ($NPwTF$) is provided.

Data about email communication comes from the period: 01.1999 — 07.2002. Before the cleansing process there were 517,431 emails whereas after there were 411,869 emails. All data were obtained from the email boxes of company employees.

First, the data has to be cleansed by the removal of bad email addresses and unification of duplicates. Every email with more than one recipient was treated as $\frac{1}{n}$ of a regular email, where n is the number of its recipients.

Equation 4.3 had to be applied to users who are not active in the email system.

After data preparation the commitment function is evaluated for each pair of members. To evaluate relationship commitment function $C(x, y)$ two formulas Equation 4.4 and Equation 4.5 were used. Equation 4.4 was utilized to calculate node position without respecting time (*NP*) whereas Equation 4.5 serves to evaluate node position with time factor (*NPwTF*). In the case of *NPwTF* calculation the whole period was split into one month periods. The initial node positions for all members were set 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. $\varepsilon = 0.01$, $\varepsilon = 0.1$, $\varepsilon = 0.3$, $\varepsilon = 0.5$, $\varepsilon = 0.7$, $\varepsilon = 0.9$.

6.2.2 Characteristics of Node Position

Distribution of Node Position Values

Interesting information about the values of node position is provided by the average node position as well as standard deviation evaluated for the entire social network. The analyses of node position values (NP) and node positions with time factor ($NPwTF$) for Enron can be found in Figure 6.6 as well as in Table 6.5 and Table 6.6.

Table 6.5: Average NP and $NPwTF$ in the Enron dataset, calculated for different values of ε

ε	Average NP	Average $NPwTF$
0.01	1.03516	0.95331
0.1	1.00450	0.98760
0.3	1.07683	0.97865
0.5	0.98975	0.99863
0.7	1.09987	0.97433
0.9	1.05439	0.98799

Table 6.6: Standard deviations of NP and $NPwTF$ in the Enron dataset, calculated for different values of ε

ε	Std. Dev. of NP	Std. Dev. of $NPwTF$
0.01	0.22333	0.19849
0.1	0.43490	0.32171
0.3	0.75407	0.74096
0.5	0.92446	1.12111
0.7	0.94420	1.44072
0.9	0.96289	1.44086

The average node position seems to be convergent to 1 in all cases.

On the other hand, the standard deviation substantially differs depending on the value of coefficient ε. The greater ε, the bigger standard deviation (Table 6.6). It shows that for greater ε the value of the distance between the members' node positions increases and it is valid for both NP and $NPwTF$ and in consequence greater ε provides the opportunity to distinguish people in a better and more precise way.

It can be noticed that the value of node position NP for over 93% (Figure 6.7) of email users in the Enron community 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 0.43% of users. This confirms that**

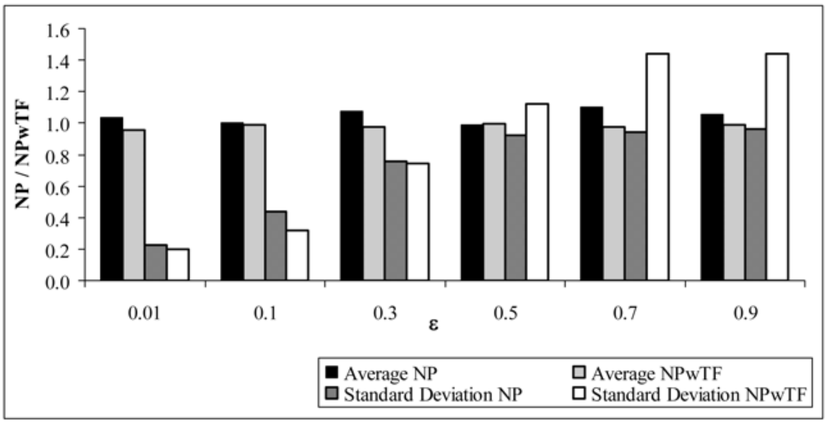


Figure 6.6: Average NP and $NPwTF$ as well as their standard deviations in the Enron dataset, calculated for different values of ε

node position can be the good measure to extract key persons in the network of Internet users. The distribution of NP depending on ε is presented in Figure 6.8.

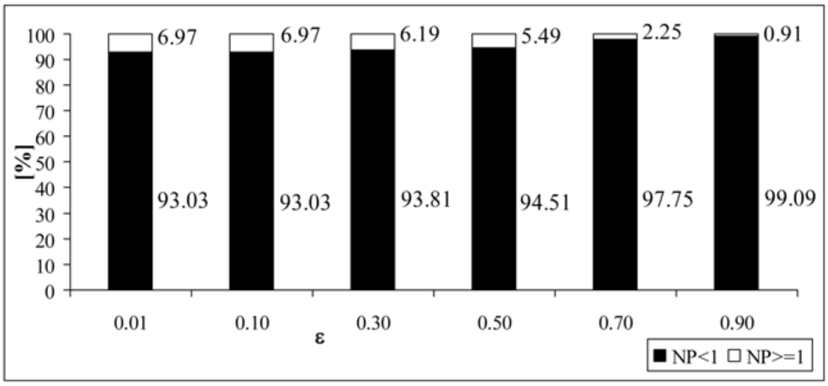


Figure 6.7: The percentage of users with $NP < 1$ and $NP \geq 1$ within the Enron network in relation to ε

Duplicates Number

Node position measure appears to be more diverse than other centrality measures. It can be visible especially while analyzing number of nodes that possess the same centrality value, Figure 6.9. Node positions are better for every value of ε, compared to indegree centrality (IDC) and outdegree centrality (ODC). Note that outdegree centrality function provides only 286 distinct values and in case of outdegree centrality there are only 383 distinct values. For that reason, the percentage of duplicates exceeds 95% for degree measures whereas it is below 60% for node positions Figure 6.9. For deeper analysis of duplicates number in comparison to other centrality measures see Sections 6.1 and 6.3.

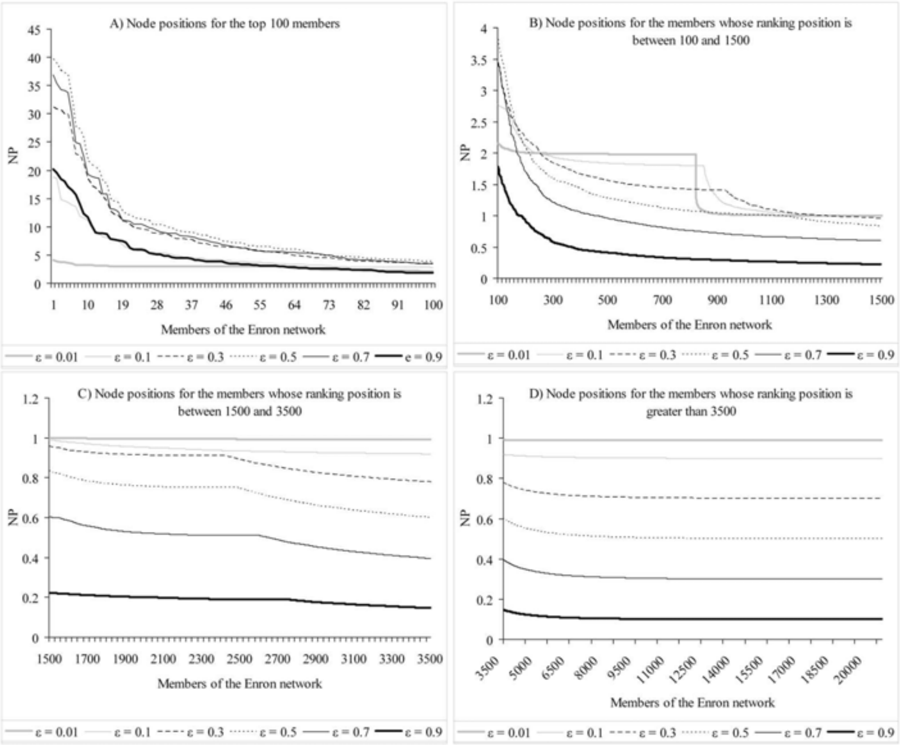


Figure 6.8: The distribution of NP depending on ε

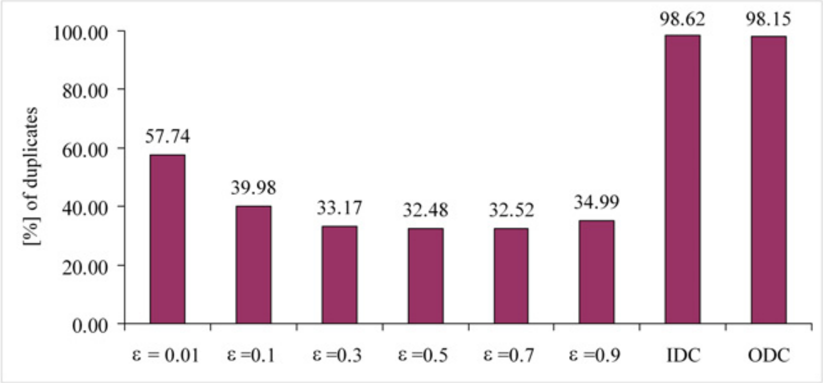


Figure 6.9: Percentage of duplicates within the set of node measures, separately for node position with different values of ε, degree prestige (IDC), and degree centrality (ODC)

Rankings 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 [74]:

$$\kappa(X,Y) = \frac{1}{n(n-1)} \cdot \sum_{i=1}^{n}\sum_{j=1}^{n} sgn(x_j - x_i) \cdot sgn(y_j - y_i) \tag{6.1}$$

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 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 outdegree centrality (ODC), indegree centrality (IDC), and node position for different ε, Table 6.7[1]. 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 ODC– and IDC–based: κ was only 0.07. The closeness between ODC– and IDC–based ranking was rather high: $\kappa = 0.35$. ODC– and IDC–based rankings are close to each other and differ from NP–based because both ODC and IDC provide big number of duplicates and do not distinguish users. It reveals that ODC and IDC deliver similar, limited knowledge about users in the network whereas node position function is the diverse, meaningful measure.

Node Position with Time Factor

The aim of the next part of the experiments is to investigate the influence of time factor on the values of node position. 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, Figure 6.10. The reason for that can be the fact that in the last periods the users were more active then before what can result from the company problems and bankruptcy. Moreover, this difference is greater for greater values of ε – up to more than 7 times in case

[1] Note that the duplicates have in the ranking the same position

Table 6.7: 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		
ODC	0.0041	0.0041	0.0084	0.0081	0.0077	0.0074	
IDC	0.0052	0.0052	0.0081	0.0079	0.0077	0.0746	0.3517

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 $NPwTF$, Figure 6.11. Overall, the 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. 6.8. 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.

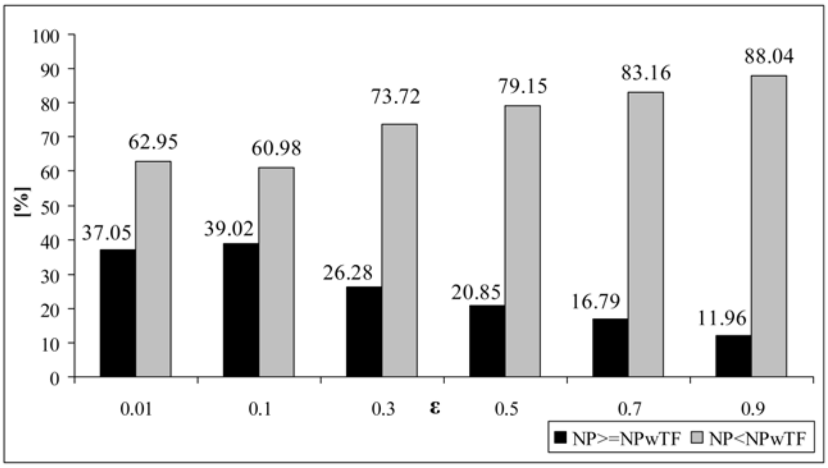


Figure 6.10: The percentage contribution of members with $NP \geq NPwTF$ and $NP < NPwTF$ within the Enron social network in relation to ε

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 Figure 6.6.

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 Equation 4.5.

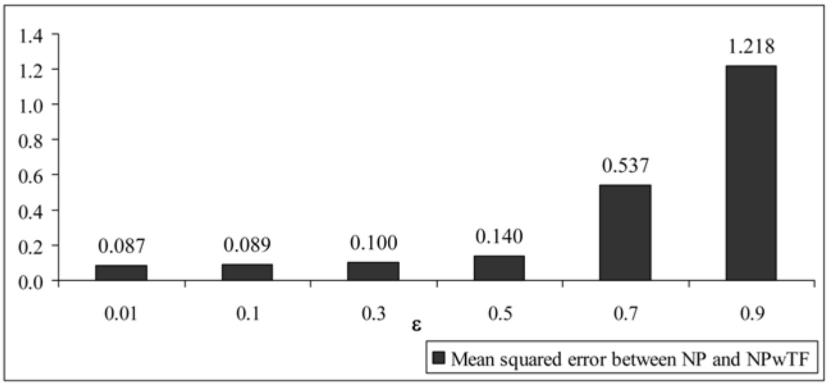


Figure 6.11: The mean squared error between NP and $NPwTF$ for the Enron dataset in relation to ε

6.3 Wroclaw University of Technology Network

The main goal of the experiments that were performed on Wroclaw University of Technology email logs was to present how the parameters of node position method, i.e. ε coefficient and stoping condition, influence the results.

6.3.1 Data Description and Preparation

The experiments were carried out on the logs from the Wroclaw University of Technology (WUT) mail server, which contain only the emails incoming to the staff members as well as organizational units registered at the university (Figure 6.12).

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. Every email with more than one recipient was treated as $1/n$ of a regular email, where n is the number of its recipients. The general statistics related to the processed datasets are presented in Table 6.8.

6.3.2 Influence of stopping condition τ and ε coefficient

This part of the experiments is devoted to the influence of different values of stoping condition τ and ε coefficient on the processing time, number of required iterations and number of distinct values of node position in the final ranking. The PIN^{edges} algorithm was utilized in all of these calculations.

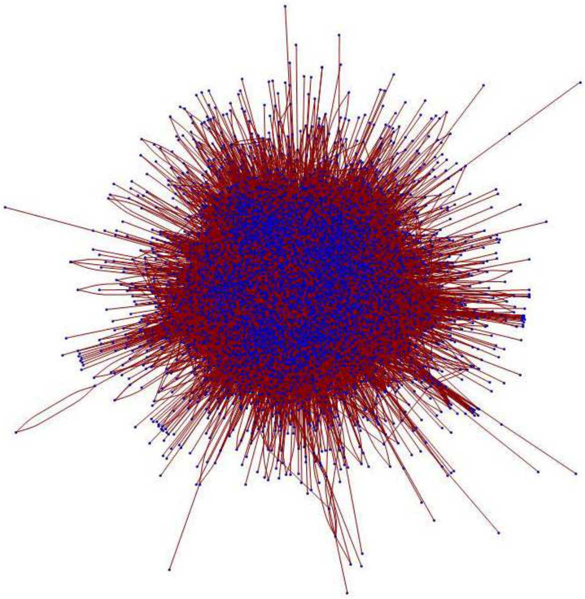


Figure 6.12: Social network discovered from the email communication between employees of WUT

Table 6.8: 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 WUT domain without isolated members	5,845
Emails per user	479
Network users with no activity	26 (0.45%)
Commitments extracted from emails	149,344
Relationships per user	30.2
Percentage of all possible relationships	0.517%

Processing Time

The processing time in relation to τ and ε values is presented in Table 6.9 and illustrated in Figures 6.13, 6.14 and 6.15. For each value of ε coefficient the processing time increases with the smaller value of stopping condition τ which is intuitive (Figure 6.15). Simultaneously, for each value of stopping condition τ the processing time increases with the greater ε value (Figure 6.14). For ε equal 0.9 the processing time is significantly longer than in the case of other ε values and for $\tau = 0.000001$ the time is six times longer than in the case of $\tau = 0.1$. On contrary for $\varepsilon = 0.1$ the processing time is 2.15 times shorter for $\tau = 0.1$ than for $\tau = 0.000001$.

Table 6.9: Processing time [min] of the PIN^{edges} depending on different values of ε coefficient and stopping condition τ

τ ε	0.1	0.01	0.001	0.0001	0.00001	0.000001
0.1	9.77	9.83	12.77	15.79	18.75	21.81
0.2	9.77	12.77	15.62	21.64	24.54	27.50
0.3	12.67	15.65	21.35	24.23	30.19	36.49
0.4	12.67	18.69	24.29	30.31	39.07	45.36
0.5	15.59	24.58	30.48	39.42	48.25	57.04
0.6	18.56	27.52	39.37	51.25	63.19	71.71
0.7	24.24	36.00	50.92	65.72	83.52	98.62
0.8	30.14	51.05	71.80	95.95	119.73	142.63
0.9	47.89	80.33	122.17	158.03	220.68	287.30

Number of Required Iterations

The second feature that is investigated is the number of iterations that is required to fulfill the set stopping condition τ in relation to coefficient ε. The outcome of the experiments is very similar as in the case of the processing time and is presented in Table 6.10 and Figure 6.16. For each value of ε coefficient the number of iterations increases with the smaller value of stopping condition τ. Simultaneously, for each value of stopping condition τ the number of required iterations increases with the greater ε value (Figure 6.14).

For $\varepsilon = 0.9$ and for $\tau = 0.000001$ the number of iterations is six times bigger than in the case of $\tau = 0.1$. On contrary for $\varepsilon = 0.1$ the number of iterations is 2.33 times smaller for $\tau = 0.1$ than for $\tau = 0.000001$. This reveals that for greater ε the number of required iterations increases faster with greater τ value than for smaller ε.

On the other hand, taking into account the value of stopping condition for $\tau = 0.1$ the number of iterations for $\varepsilon = 0.1$ is 5.(3) times smaller than in the case of $\varepsilon = 0,9$. For the smallest considered $\tau = 0.000001$ the number of iterations for $\varepsilon = 0.1$ is 13.71 times smaller than in the case of $\varepsilon = 0.9$.

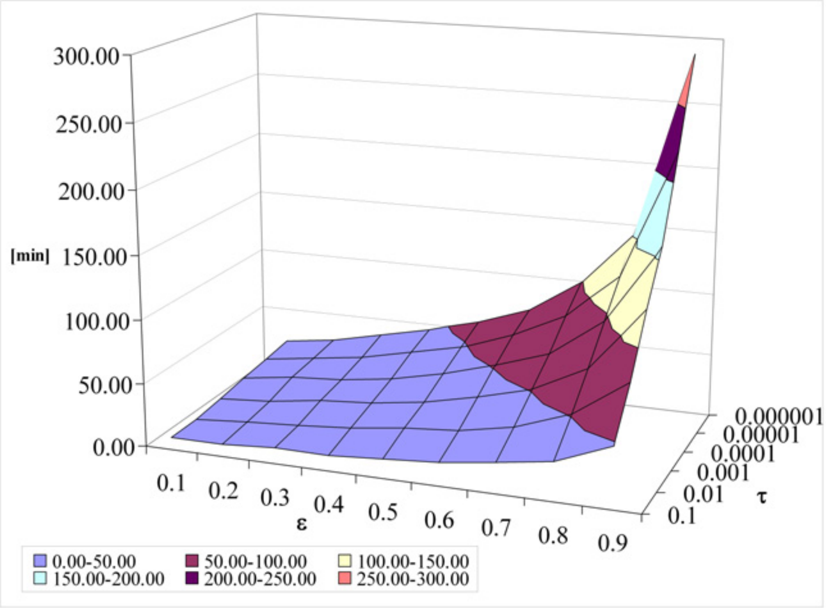


Figure 6.13: Processing time [min] of the PIN^{edges} depending on different values of ε coefficient and stop condition τ

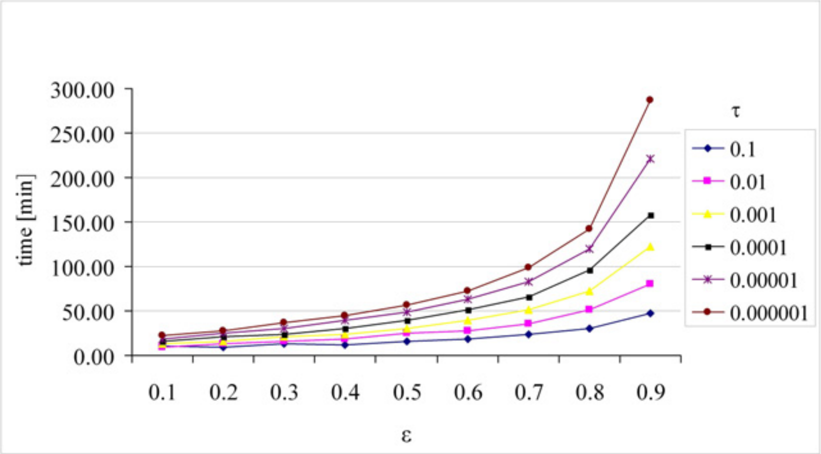


Figure 6.14: Processing time [min] of the PIN^{edges} depending on different values of coefficient ε

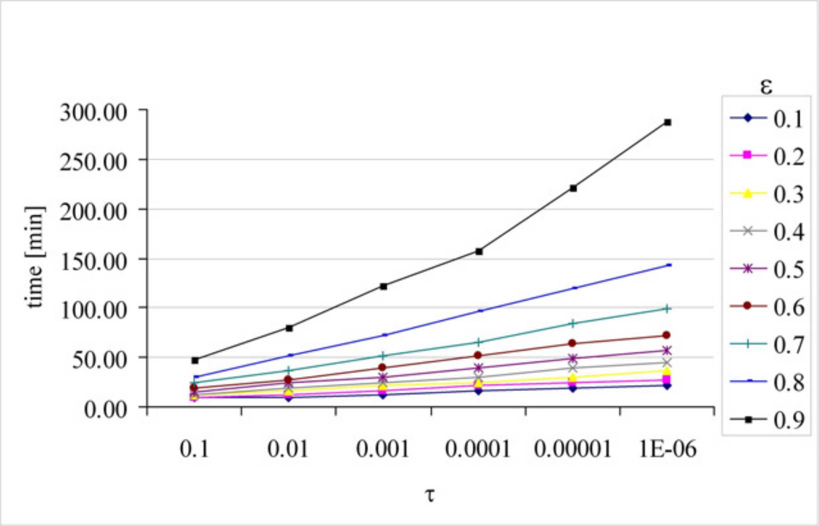


Figure 6.15: Processing time [min] of the PIN^{edges} depending on different value stop condition τ

Thus, for smaller τ the number of required iterations increases faster with greater ε value than for smaller ε.

Number of Duplicates

The last analyzed feature is the number of distinct values of node position within the WUT network. The total number of nodes in the network equals 5945. The number of duplicates for a given ε coefficient and stopping condition τ is presented in Table 6.11 and Figure 6.17. It can be noticed that for small ε and τ values the number of duplicates is significant, e.g. for $\tau = 0.1$ and $\tau = 0.01$ regardless of the ε value there are over 95% of duplicates. On the other hand for $\tau = 0.000001$ the number of duplicates is the largest for $\varepsilon = 0.1$ — 12.7% and decreases with the larger ε.

Note that for larger ε and smaller τ the number of duplicates decreases but the number of iterations needed to meet the stopping condition and in consequence processing time increases. This trade–off between the processing time and the number of distinct values enables a researcher to choose the values of parameters that best suit his/her needs. Note that for $\varepsilon = 0.1$ to $\varepsilon = 0.5$ and for the smallest $\tau = 0.000001$ the processing time does not exceed one hour and the number of duplicates is less than 755. This could be good parameters to extract key nodes. But on the other hand in the case of small ε the distribution of node position values is not very significant — the standard deviation for $\varepsilon = 0.1$ equals 0.11. Thus, all nodes will have the node position close to the average value and assessing the degree to which a given node is more important than another one is hard. With the larger ε the standard deviation increases and this distinguishes nodes in a better way but these calculations require more resources.

Table 6.10: Number of iterations depending on different values of ε coefficient and stop condition τ

τ ε	0.1	0.01	0.001	0.0001	0.00001	0.000001
0.1	3	3	4	5	6	7
0.2	3	4	5	7	8	9
0.3	4	5	7	8	10	12
0.4	4	6	8	10	13	15
0.5	5	8	10	13	16	19
0.6	6	9	13	17	21	24
0.7	8	12	17	22	28	33
0.8	10	17	24	32	40	48
0.9	16	27	41	53	74	96

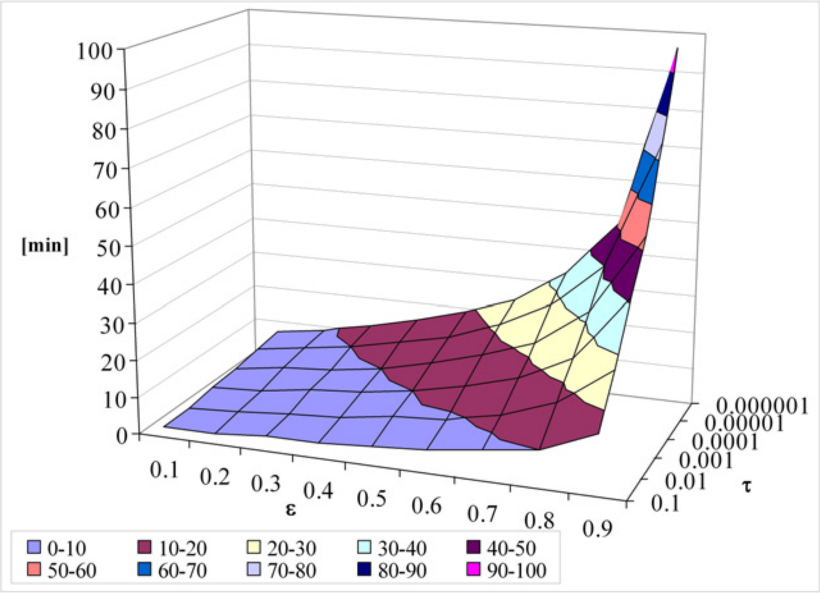


Figure 6.16: Number of required iterations depending on different values of ε coefficient and stopping condition τ

Table 6.11: Number of duplicates depending on different values of ε coefficient and stopping condition τ

τ ε	0.1	0.01	0.001	0.0001	0.00001	0.000001
0.1	5929	5857	5504	4039	1777	755
0.2	5919	5803	5234	3344	1385	631
0.3	5912	5759	5024	2917	1156	569
0.4	5904	5722	4882	2658	1039	545
0.5	5899	5689	4734	2482	976	549
0.6	5895	5658	4655	2368	926	520
0.7	5892	5644	4595	2301	883	531
0.8	5892	5635	4612	2340	913	521
0.9	5900	5680	4773	2551	985	540

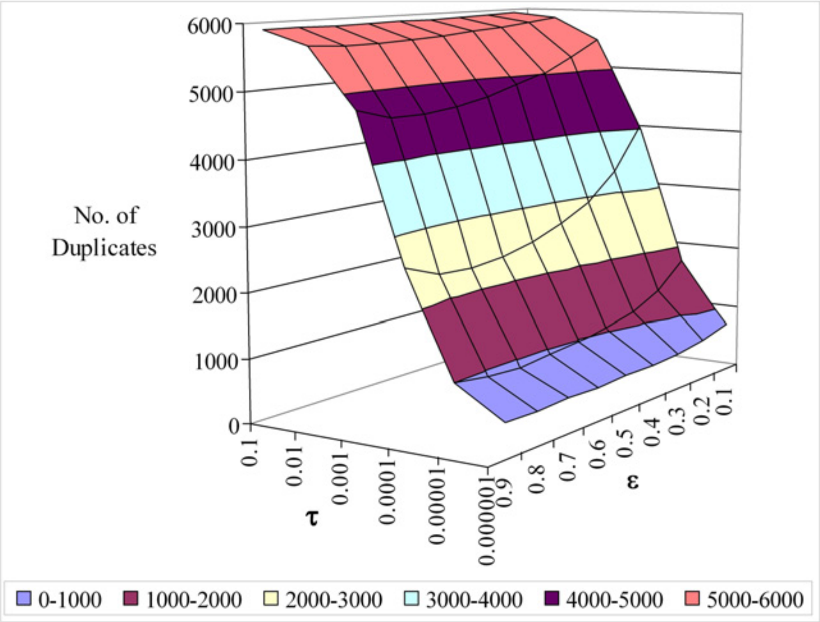


Figure 6.17: Processing time [min] of the PIN^{edges} depending on different values of ε coefficient and stop condition τ

Correlation Between Node Position Method Parameters

The conducted experiments have revealed that such features as: processing time, number of required iterations and number of distinct values of *NP* are highly influenced by the method parameters — ε and τ. Based on the experiments outcomes the correlation function between method parameters and enumerated features can be identified and analyzed.

First, the correlation between the number of iterations — n and ε was taken into consideration. A separate chart was created for each of the stopping conditions (Figure 6.18). After that the fitting process was performed. As a result we obtained the function that approximates the experimental data in the best way. This is an exponential growth function and it is described by the following formula: $n(\varepsilon) = A \cdot e^{\varepsilon/t} + n_0$. The values of A, t, n_0 and correlation rate CR diverse depending on the value of τ and they are presented in Table 6.12. The correlation rate is bigger than 0.99 for all values of τ what means that exponential growth function approximates the

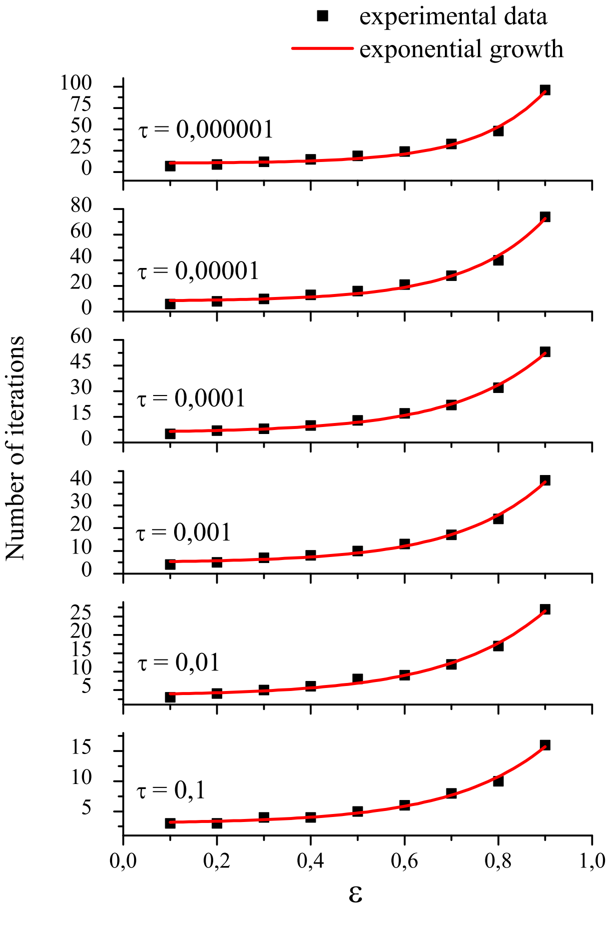


Figure 6.18: Exponential fitting function for the relation between ε and number of iterations

relation between number of iterations and ε with a very high accuracy. Note that the highest possible is 1.

Table 6.12: Parameters values for correlation function $n(\varepsilon)$

τ	A	t	n_0	CR
0.1	0.15	0.20	2.93	**0.99215**
0.01	0.32	0.21	3.40	**0.99277**
0.001	0.31	0.20	4.80	**0.99256**
0.0001	0.48	0.20	5.72	**0.99518**
0.00001	0.31	0.17	8.05	**0.99080**
0.000001	0.19	0.15	10.30	**0.99052**

As it is presented in Section 6.4.1 duration of a single iteration does not depend on the value of ε. This results in the fact that processing time is directly proportional to a number of iterations. Thus, the dependence between processing time and ε will be the same as the dependence between number of iterations and ε, i.e. it will be expressed by the exponential growth function. The longer processing time means that the pace of convergence is lower. Based on this the following assumption can be made: a pace of convergence of the proposed method is inversely proportional to processing time. In consequence the dependence between a pace of convergence and ε is best approximated by the exponential decay function.

Not only the correlation between ε and processing time can be described but also between ε and number of duplicates d. The experiments have revealed that the best fitting function that describes the dependence between ε and d is an exponential decay function that is expressed by the formula: $d(\varepsilon) = A \cdot e^{-\varepsilon/t} + n_0$ (Figure 6.19). The values of A, t, n_0 and correlation rate CR diverse depending on the value of τ and they are presented in Table 6.13. The correlation rate is the smallest for $\tau = 0.1$ and equal 0.94287 and the biggest for $\tau = 0.000001$ and equal 0.98888. It means that exponential growth function approximates the relation between number of duplicates and ε with the highest accuracy for $\tau = 0.000001$. The correlation rate is not as big as in the case of $n(\varepsilon)$ function because for $d(\varepsilon)$ there appear some fluctuations in the number of duplicates for $\varepsilon = 0.7$, $\varepsilon = 0.8$ and $\varepsilon = 0.9$ (see Table 6.11), which can result from the numerical artefact. However, it can be noticed that for bigger τ these fluctuations are smaller as the CR increases (Table 6.13).

6.3.3 Comparison with Other User Position Indices

In the last part of the experiments on WUT dataset a comparison between number of duplicates in different centrality indices is presented (Table 6.14). 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

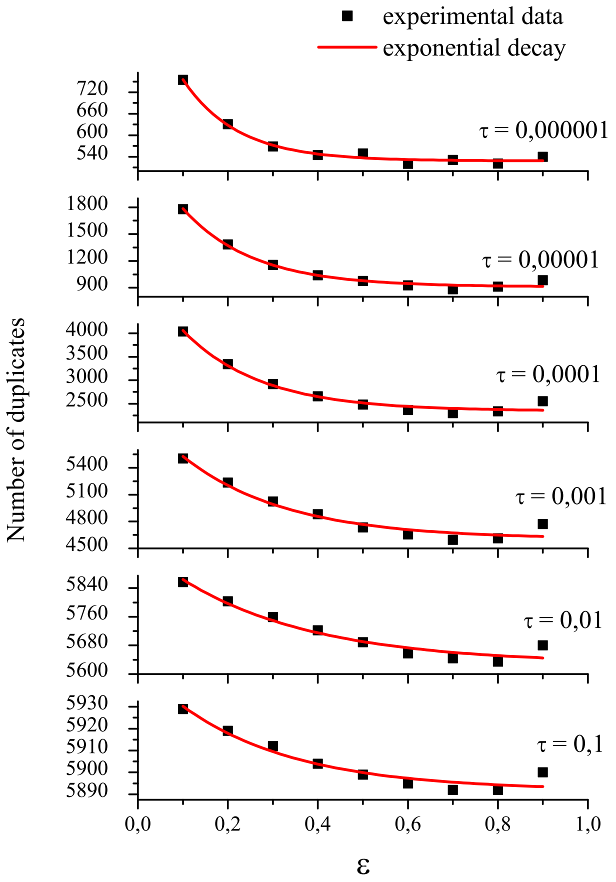


Figure 6.19: Exponential fitting function for the relation between ε and number of duplicates

Table 6.13: Parameters values for correlation function $d(\varepsilon)$

τ	A	t	n_0	CR
0.1	56.68	0.29	5891.76	**0.94287**
0.01	328.05	0.30	5629.03	**0.95228**
0.001	1425.91	0.23	4604.21	**0.95547**
0.0001	3048.13	0.17	2344. 38	**0.97913**
0.00001	1653.10	0.16	911.14	**0.98883**
0.000001	516.67	0.12	528.44	**0.98888**

processing time in comparison to larger τ values. The biggest number of 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 the case Node Position measure. Analyses of a processing time of the algorithms that were used to calculate the specific centrality measures reveal that the most efficiency 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. *NP* method is approximately 1470 times faster than measures based on the concept of shortest paths. The complexity of the algorithms for assessing the centrality based on rank prestige concept is the same as for the measures based on which the input values for rank prestige method were calculated. For example, in the case of closeness centrality eigenvector *CCE* the complexity will be the same as in the case of *CC*.

Table 6.14: Number of duplicates and processing time for different centrality indices

Measure	**Number of Duplicates**	**Percentage of Duplicates [%]**	**Processing Time [min]**
IDC	5737	96.50	2.17
ODC	5703	95.93	2.28
ECC	1943	32.68	14891.70
CC	1888	31.76	14746.86
NP, $\varepsilon = 0.1$	755	12.70	21.81
NP, $\varepsilon = 0.2$	631	10.61	27.50
NP, $\varepsilon = 0.3$	569	9.57	36.49
NP, $\varepsilon = 0.4$	545	9.17	45.36
NP, $\varepsilon = 0.5$	549	9.23	57.04
NP, $\varepsilon = 0.6$	520	8.75	71.71
NP, $\varepsilon = 0.7$	531	8.93	98.62
NP, $\varepsilon = 0.8$	521	8.76	142.63
NP, $\varepsilon = 0.9$	540	9.08	287.30

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 (e.g. 21.81 [min] for $\varepsilon = 0.1$ and $\tau = 0.000001$). On 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 the level of 30% but 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.

6.4 Efficiency Tests

The main aim of the performed efficiency tests is to investigate, which of the three developed algorithms: PIN^{nodes}, PIN^{edges} or PIN^{hybrid} is the most effective one as well as to compare the efficiency of the presented method with these presented in Chapter 4.

The efficiency tests are split into three main stages. First, the influence of ε coefficient as well as the stopping condition τ on processing time of different variants of PIN algorithms is investigated. In the second phase, the tests are performed on the networks of different size, i.e. with different number of nodes and edges. These are random networks that were generated for the needs of the experiments. The final step of these experiments is to compare the efficiency of the developed method with other measures that serve to estimate the position of the node within the social network.

6.4.1 Influence of ε on Processing Time

This part of experiments was performed on the real telecommunication dataset (phone calls). The network of users extracted from obtained data consists of over 4 million users and over 17 million connections. The tests were done for the following ε values: 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9. For each of the ε value the processing time of three developed algorithms (PIN^{nodes}, PIN^{edges} and PIN^{hybrid}) was calculated and the outcomes are presented in Table 6.15. The values in the table are the time (in seconds and in minutes) of one iteration for the given algorithm.

Table 6.15: Processing time of the PIN algorithm depending on different values of ε coefficient

ε	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
PIN^{nodes}									
time [s]	338,881	338,589	338,699	338,740	338,761	338,859	338,689	338,409	339,236
time [min]	5,648.02	5,643.16	5,644.99	5,645.67	5,646.01	5,647.65	5,644.81	5,640.14	5,653.93
PIN^{edges}									
time [s]	2,646	2,740	2,790	2,769	2,800	2,749	2,757	2,722	2,703
time [min]	44.10	45.66	46.51	46.15	46.67	45.82	45.96	45.37	45.05
PIN^{hybrid}									
time [s]	6,764	6,663	6,651	6,739	6,779	6,775	6,664	6,746	6,679
time [min]	112.73	111.04	110.86	112.32	112.99	112.92	111.07	112.43	111.32

It can be easily noticed that the processing time is the longest for PIN^{nodes} version of the algorithm and the shortest for the PIN^{edges} one. PIN^{edges} is over 120 times faster than PIN^{nodes} algorithm and about 2.5 times faster than the PIN^{hybrid}.

Table 6.16: Average processing time of one iteration and its standard deviation for different PIN algorithms

	Average time of one iteration [min]	Standard deviation [min]
PIN^{nodes}	5646.04	3.79
PIN^{edges}	45.70	0.79
PIN^{hybrid}	111.96	0.88

When the ε coefficient is taken into consideration then the average processing time of one iteration for PIN^{nodes} is over 5000 minutes, for PIN^{edges} is around 41 minutes and for PIN^{hybrid} equals 100 minutes (Table 6.16). The analysis of standard deviation of these values enables to assess how the ε coefficient influence the processing time of PIN algorithms (Table 6.16). The smallest standard deviation is in the case of PIN^{edges} algorithm and it equals 0.79 [min] whereas the longest one is for PIN^{nodes} (3.79 [min]) and this is intuitive because the average time of one iteration is also the longest. These standard deviations are small in comparison to average processing time of one iteration for different ε values so it can be assumed that the value of ε coefficient does not influence the processing time of the algorithms.

6.4.2 Influence of Network Size on Processing Time

The next stage of the efficiency test were performed on randomly generated social networks. In order to do that the algorithms used to generate the random networks were developed and 25 different directed networks were generated (Table 6.17). All further experiments in this section were performed for $\varepsilon = 0.8$ as it was shown that the value of ε value does not influence the processing time.

First, the tests were performed for PIN^{nodes} algorithm (Table 6.18 and Figure 6.20). The processing time for the largest network (100,000 nodes and the same number of edges) was approximately 1950 times longer than the processing time for the smallest network (1,000 nodes and 1,000 edges). It reveals that the network size has great influence on processing time. The larger the network, the processing time increases a lot.

The next tests were performed for PIN^{edges} algorithm (Table 6.19 and Figure 6.21). The processing time for the largest network (100,000 nodes and the same number of edges) was approximately 84 times longer than the processing time for the smallest network (1,000 nodes and 1,000 edges). It shows that the influence of the network size on processing time is smaller than in the case of the PIN^{nodes} algorithm.

The last tests were performed for PIN^{hybrid} algorithm (Table 6.20 and Figure 6.22). The processing time for the largest network (100,000 nodes and the same number of edges) was approximately 88 times longer than

Table 6.17: The random networks generated for the needs of the efficiency tests

ID	Number of Nodes	Number of Edges
1	1,000	1,000
2		5,000
3		10,000
4		50,000
5		100,000
6	5,000	1,000
7		5,000
8		10,000
9		50,000
10		100,000
11	10,000	1,000
12		5,000
13		10,000
14		50,000
15		100,000
16	50,000	1,000
17		5,000
18		10,000
19		50,000
20		100,000
21	100,000	1,000
22		5,000
23		10,000
24		50,000
25		100,000

Table 6.18: Processing time of the PIN^{nodes} depending on different size of network

Nodes Edges	1,000	5,000	10,000	50,000	100,000	Unit
1,000	15.54	27.78	39.96	138.77	255.04	[s]
	0.26	0.46	0.67	2.31	4.25	[min]
5,000	265.50	326.59	392.55	861.69	1444.97	[s]
	4.43	5.44	6.54	14.36	24.08	[min]
10,000	976.55	1,144.60	1,189.02	2,160.86	3,367.28	[s]
	16.28	19.08	19.82	36.01	56.12	[min]
50,000	10,538.83	10,937.89	11,241.66	14,701.08	17,517.50	[s]
	175.65	182.30	187.36	245.02	291.96	[min]
100,000	22,141.31	22,185.78	23,360.89	26917.69	30,304.94	[s]
	369.02	369.76	389.35	448.63	505.08	[min]

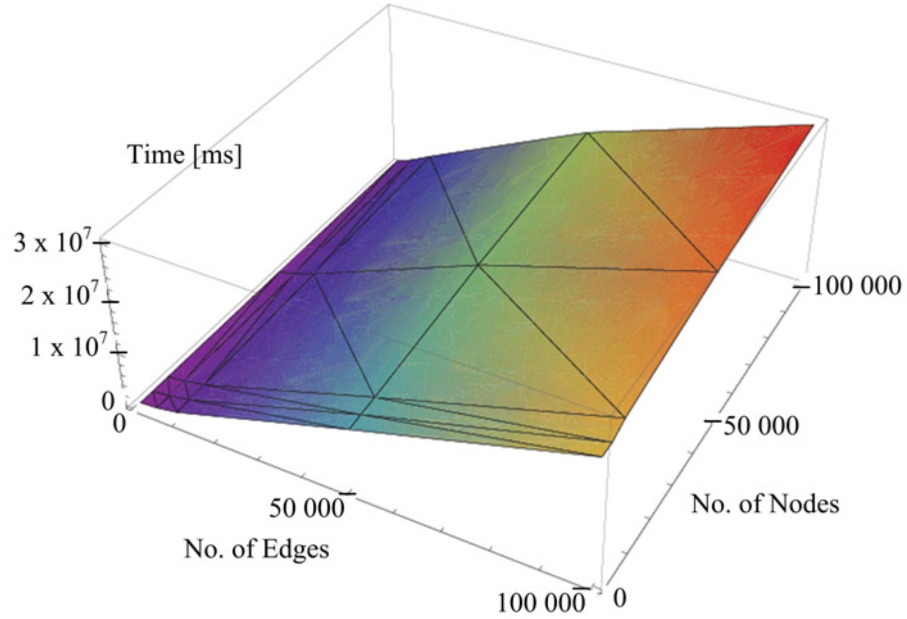


Figure 6.20: Processing time of the PIN^{nodes} depending on different size of network

Table 6.19: Processing time of the PIN^{edges} depending on different size of network

Edges \ Nodes	1,000	5,000	10,000	50,000	100,000	Unit
1,000	0.45	0.73	1.05	3.55	6.12	[s]
	0.01	0.01	0.02	0.06	0.10	[min]
5,000	1.72	2.10	2.42	5.04	8.13	[s]
	0.03	0.04	0.04	0.08	0.14	[min]
10,000	3.46	3.81	3.96	6.55	10.09	[s]
	0.06	0.06	0.07	0.11	0.17	[min]
50,000	16.57	16.20	16.39	18.95	22.79	[s]
	0.28	0.27	0.27	0.32	0.38	[min]
100,000	31.97	31.94	33.03	35.92	37.92	[s]
	0.53	0.53	0.55	0.60	0.63	[min]

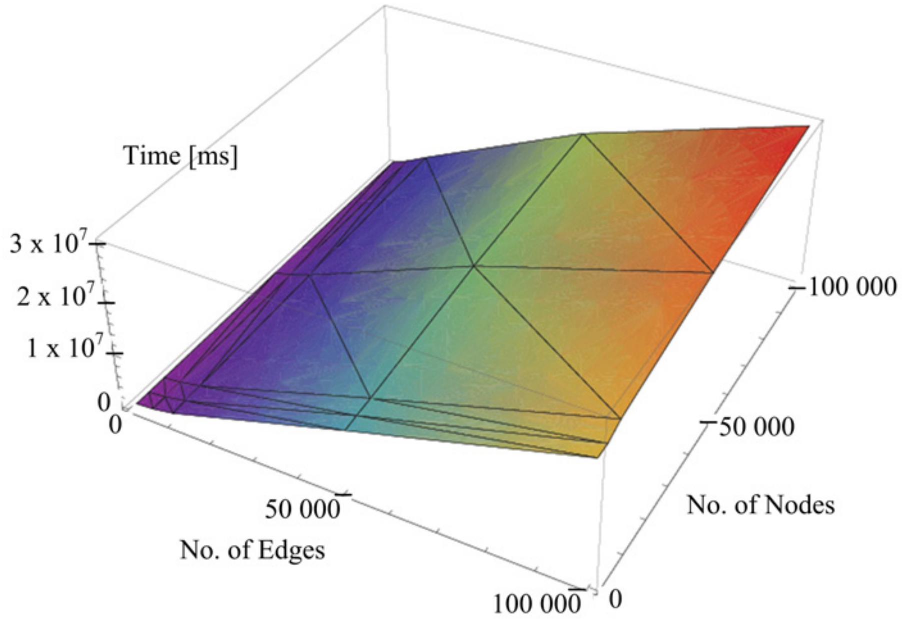


Figure 6.21: Processing time of the PIN^{edges} depending on different size of network

the processing time for the smallest network (1,000 nodes and 1,000 edges). Similarly to PIN^{edges}. It shows that the influence of the network size on processing time is smaller than in the case of the PIN^{nodes} algorithm. Moreover the influence of network size is similar when PIN^{edges} and PIN^{hybrid} algorithms are considered.

Table 6.20: Processing time of the PIN^{hybrid} depending on different size of network

Nodes / Edges	1,000	5,000	10,000	50,000	100,000	Unit
1,000	0.91	1.16	1.43	3.73	6.79	[s]
	0.02	0.02	0.02	0.06	0.11	[min]
5,000	3.76	4.03	4.38	7.31	9.84	[s]
	0.06	0.07	0.07	0.12	0.16	[min]
10,000	7.59	7.81	7.90	10.43	13.93	[s]
	0.13	0.13	0.13	0.17	0.23	[min]
50,000	35.77	35.51	35.67	38.90	43.84	[s]
	0.60	0.59	0.59	0.65	0.73	[min]
100,000	69.44	70.57	71.50	76.87	80.01	[s]
	1.16	1.18	1.19	1.28	1.33	[min]

The comparison of different variants of PIN algorithm reveals that the fastest one is always PIN^{edges}. Consider for example processing time for networks where the number of edges is constant and equals 50,000 whereas

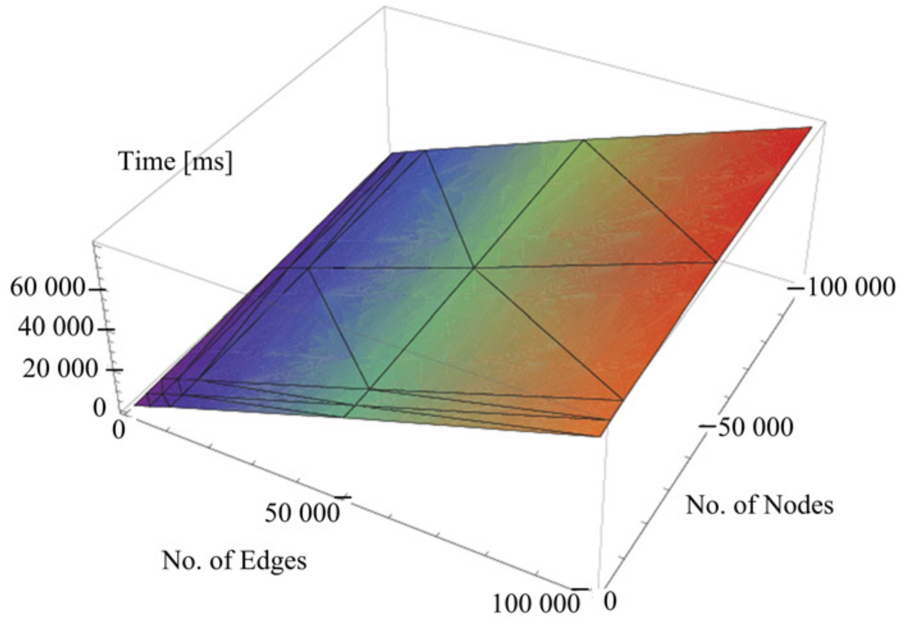


Figure 6.22: Processing time of the PIN^{hybrid} depending on different size of network

the number of nodes changes as shown in Table 6.21. Note that in case of the PIN^{edges} and PIN^{hybrid} algorithms the processing times do not differ a lot among 1,000, 5,000 and 10,000 nodes and it oscillates around 16 sec. for PIN^{edges} and 35 sec. for PIN^{hybrid}.

Table 6.21: Processing time in relation to the number of nodes in the network for fixed number of edges (50,000)

No. of Nodes	PIN^{nodes} [s]	PIN^{edges} [s]	PIN^{hybrid} [s]
1,000	10,539	16.56	35.77
5,000	10,938	16.19	35.51
10,000	11,242	16.39	35.67
50,000	14,701	18.95	38.90
100,000	17,518	22.79	43.84

The PIN^{edges} is 636.2 times faster than PIN^{nodes} for 1,000 nodes and 768.77 times faster for 100,000 nodes. Simultaneously, PIN^{edges} algorithm is approximately two times faster than PIN^{hybrid} algorithm for all types of investigated random networks where the number of edges equals 50,000 (Table 6.22).

The processing time is a monotonically increasing function of the number of nodes in the network, i.e. the larger number of nodes, the longer processing time (Figure 6.23). However, only in case of the PIN^{edges} algorithm the processing time is additionally linear function of the number of nodes in the

Table 6.22: The relation of processing times of PIN^{edges} to other PIN algorithms for fixed number of edges (50,000)

No. of Nodes	$\frac{t_{PIN^{nodes}}}{t_{PIN^{edges}}}$	$\frac{t_{PIN^{hybrid}}}{t_{PIN^{edges}}}$
1,000	636.20	2.16
5,000	675.38	2.19
10,000	685.97	2.18
50,000	775.65	2.05
100,000	768.77	1.92

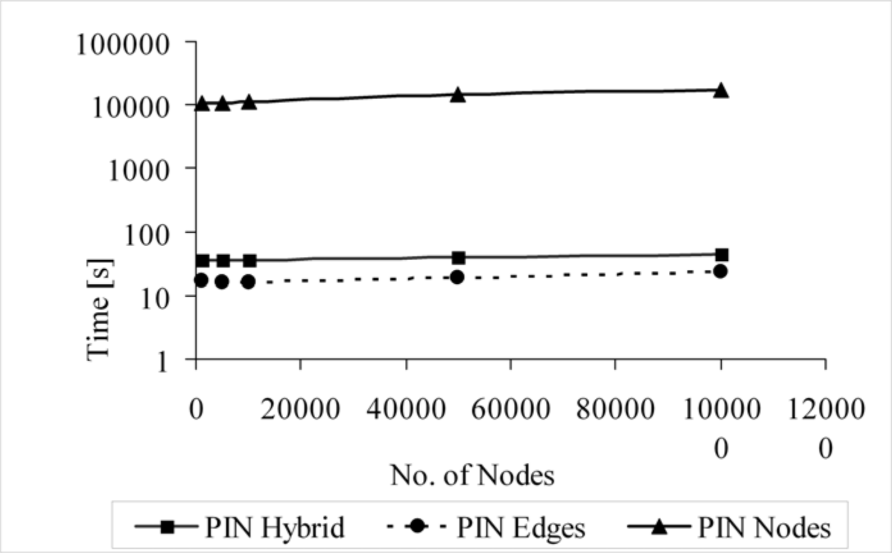


Figure 6.23: Processing time depending on the number of nodes in the network for fixed number of edges (50,000)

network. Moreover, the tangent of the slope is very close to zero and it means that the values of the function increase very slowly. In other words they are almost constant (Table 6.23).

Table 6.23: The ratio of processing time and number of nodes for different PIN algorithms for constant number of edges (50,000)

No. of Nodes	PIN^{nodes}	PIN^{edges}	PIN^{hybrid}
1,000	10.5388	0.0016	2.1591
5,000	2.1876	0.0015	2.1924
10,000	1.1242	0.0015	2.1765
50,000	0.2940	0.0013	2.0523
100,000	0.1752	0.0013	1.9238

On the other hand let us consider the processing time for networks where the number of nodes is constant and equals 50,000 whereas the number of edges changes as shown in Table 6.24. Note that, in contrast to networks when the number of edges is constant, in case of the PIN^{edges} and PIN^{hybrid} algorithms the processing times differ a lot among 1,000, 5,000, 10,000, 50,000 and 100,000 edges. It changes from 3.55 sec. for 1,000 edges to 35.92 sec. for 100,000 edges for PIN^{edges}. While it changes from 3.73 sec. for 1,000 edges to 76.87 sec. for 100,000 edges for PIN^{hybrid}.

Table 6.24: Processing time depending on the number of nodes in the network for fixed number of nodes (50,000)

No. of Edges	PIN^{nodes} [s]	PIN^{edges} [s]	PIN^{hybrid} [s]
1,000	138.77	3.55	3.73
5,000	861.69	5.04	7.31
10,000	2,160.86	6.55	10.43
50,000	14,701.08	18.95	38.90
100,000	26,917.69	35.92	76.87

The PIN^{edges} is 39.07 times faster than PIN^{nodes} for 1,000 edges and 749.28 times faster for 100,000 edges. At the same time PIN^{edges} algorithm is as fast as PIN^{hybrid} for 1,000 edges and 2 times faster for 100,000 edges. (Table 6.25).

The processing time is a monotonically increasing function of the number of edges in the network: the longer number of edges, the longer processing time (Figure 6.24). However none of them can be seen as the linear function of the number of edges in the network (Table 6.26).

The iterative nature of node position requires more or less iteration to be performed to achieve the required precision of results. However, the implementation of the general concept can be realized with different approaches.

Table 6.25: The relation of processing times of PIN^{edges} to other PIN algorithms for fixed number of nodes (50,000)

No. of Edges	$\frac{t_{PIN^{nodes}}}{t_{PIN^{edges}}}$	$\frac{t_{PIN^{hybrid}}}{t_{PIN^{edges}}}$
1,000	39.07	1.05
5,000	170.91	1.45
10,000	329.68	1.59
50,000	775.65	2.05
100,000	749.28	2.14

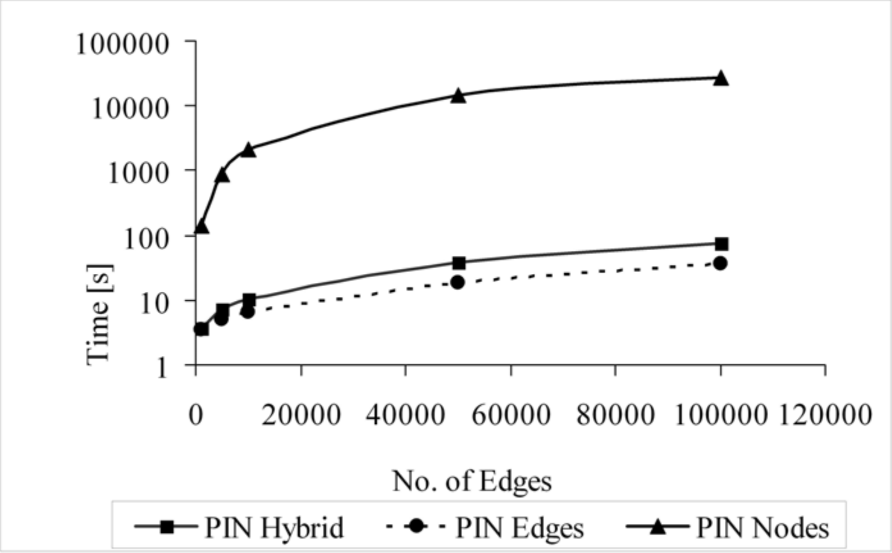


Figure 6.24: Processing time depending on the number of edges in the network for fixed number of nodes (50,000)

Table 6.26: The ratio of processing time and number of edges for different PIN algorithms for constant number of nodes (50,000)

No. of Edges	PIN^{nodes}	PIN^{edges}	PIN^{hybrid}
1,000	0.1388	0.0256	1.0514
5,000	0.1723	0.0059	1.4492
10,000	0.2161	0.0030	1.5905
50,000	0.2940	0.0013	2.0523
100,000	0.2692	0.0013	2.1397

One of the most surprising conclusions from the tests carried out is the big difference in efficiency between these three methods, even over two orders of magnitudes. The "edge approach" appears to be absolutely the best while raw, direct implementation of the concept – PIN^{nodes} remains far behind. This reveals that the implementation method for some general concepts from social network analysis may have the crucial impact on the computation efficiency.

6.4.3 Efficiency of Node Position versus Other User Position Indices

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. The measures that are taken into consideration are indegree centrality (*IDC*), outdegree centrality (*ODC*), closeness centrality (*CC*) and eccentricity centrality (*EC*) (see Chapter 4 for detailed description). 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 (Appendix A). Note that *EC* and *CC* are the least complex measures from this group (see Section 3.2). 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 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 in Table 6.27. 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 (Section 6.2 and 6.3).

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

Table 6.27: Processing time for different centrality indices

Measure	**Processing Time [min]**
IDC	2.17
ODC	2.28
ECC	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

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.

6.5 Possible Application of Node Position Method

The main goal of this case study is to illustrate that the node position method can be utilized for example to extract so called bridging nodes from the network of users. The analysis of the characteristics of the nodes that connect the whole network with the peripheral nodes or peripheral cliques that are loosely connected with the rest of the network is a very interesting research problem. The issue of identifying bridges within a given network is a complex and resource consuming task because it involves an extensive analysis of the groups and cliques existing within a given network [119], [110], [55]. Bridges can be seen as the nodes without which the network will split into two or more subgroups. This concept is similar to the idea of weak ties [52], [58], [24] that also tend to be vital bridges between the two densely knit clumps. In this case study bridges are defined as the nodes that (i) connect regular cliques with the peripheral nodes or (ii) connect regular cliques with the peripheral cliques that are loosely connected with the rest of the network. The explanations of such concepts as regular clique, peripheral clique, bridging node and peripheral node as well as the whole method for bridges properties analysis is presented. By defining the properties of the bridges we will be able to identify them without the complex calculations.

6.5.1 Method for Bridges' Properties Analysis

The main goal is to identify bridges within the social network and describe their characteristic features. The whole process is presented in the Figure 6.25. The method is split into two main groups of activities: (i) the process of bridges identification and (ii) the process of the network nodes properties calculation. Note that the network is represented as a weighted and directed graph.

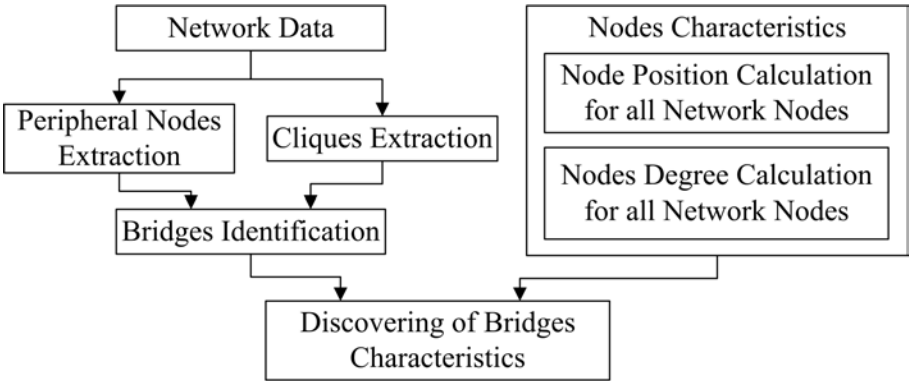

Network Data
Peripheral Nodes Extraction
Cliques Extraction
Bridges Identification
Nodes Characteristics
Node Position Calculation for all Network Nodes
Nodes Degree Calculation for all Network Nodes
Discovering of Bridges Characteristics

Figure 6.25: Crucial steps of the proposed method of bridges properties analysis

In the standard process, the first step is to extract all cliques existing in a given network. The clique is defined as "a maximal complete subgraph of three or more nodes" [119]. It means that a clique consists of a subset of nodes, all of which are connected to all nodes of the clique. Additionally, in the rest of the network where does not exist even one node that is connected to all nodes that belong to the analyzed clique [85], [56]. In the same time, the peripheral nodes and peripheral cliques are identified. The former ones are nodes that do not belong to any clique that exist within an analyzed network. The latter ones are cliques that are loosely connected with the rest of the network, i.e. they do not have any common node with the rest of the extracted cliques. The set of cliques that does not include the peripheral cliques is called a set of regular cliques. Afterwards the bridging nodes can be discovered. Bridges are nodes that belong to the regular clique and connect it with peripheral node or peripheral clique.

The second set of activities is connected with analyzing the characteristic features of the network nodes. Different characteristics from the node perspective can be analyzed, e.g. centrality, prestige, density, social capital [119], [55]. In the presented research, we take into consideration two of them: centrality index proposed in this thesis — node position as well as the number of incoming, outgoing and mutual relationships of the node [119], [26].

The last stage of the method is to find the features of the bridging nodes that are their specific and individual properties and distinguish them from the other network nodes. If it is possible to find such features, then there will be no need to perform the extensive calculations in order to locate the bridging nodes in the network structure.

6.5.2 Experiments

The experiments were conducted on the Thurman office social network (see Section 6.1) that is a non–symmetrical network of 15 people who worked in one company (Figure 6.1). Thurman spent 16 months observing the interactions among employees in the office of a large corporation [113]. The adjacency matrix for the Thurman network is presented in Table 6.1.

Process of Bridges Extraction

First step of the method is to extract the cliques existing within the network. Note, that the relationships within the Thurman network are weighted and directed. It means that group can be called a clique if and only if all the relationships between members are mutual. Four cliques were extracted from the Thurman network (the numbers indicate the ID of the specific user, see Table 6.1):

1. $C_1 = \{Ann, Amy, Lisa, Katy, Tina, Pete\} = \{1, 2, 8, 3, 6, 5\}$
2. $C_2 = \{Ann, Lisa, Pete, President\} = \{1, 8, 5, 9\}$
3. $C_3 = \{Pete, Emma, Lisa, President\} = \{5, 12, 8, 9\}$
4. $C_4 = \{Bill, Minna, Andy\} = \{4, 10, 7\}$

In order to find the peripheral nodes i.e. these that do not belong to any of the cliques, we apply the formula:

$$PN = V \setminus (C_1 \cup C_2 \cup ... \cup C_n) \tag{6.2}$$

where: PN — the set of peripheral nodes; V — the set of all nodes in a network; n — number of cliques within a given network

After utilizing the above formula the set PN is obtained:

$$PN = \{Mary, Rose, Mike, Peg\} = \{11, 13, 14, 15\}$$

Next part of the experiments is to identify the peripheral cliques. Let us remind that the peripheral clique (PC) is the clique that does not posses even one common node with all other cliques existing within the network. It means that each clique C_x is a PC if and only if the following criterion is met:

$$\sum_{i=1 \wedge i \neq x}^{n} card(C_i \cap C_x) = 0 \tag{6.3}$$

After the application of the above criterion we obtain:
$C_1 \cap C_2 = \{Ann, Lisa, Pete\} = \{1, 8, 5\}$
$C_1 \cap C_3 = \{Lisa, Pete\} = \{8, 5\}$
$C_1 \cap C_4 = \emptyset$
It does not meet the above criterion, i.e. $\sum\limits_{i=1 \wedge i \neq 1}^{4} card(C_i \cap C_1) \neq 0$ so neither C_1 nor C_2 is not the peripheral one.
$C_2 \cap C_1 = \{Ann, Lisa, Pete\} = \{1, 8, 5\}$
$C_2 \cap C_3 = \{President, Lisa, Pete\} = \{9, 8, 5\}$
$C_2 \cap C_4 = \emptyset$
This also does not meet the above criterion: $\sum\limits_{i=1 \wedge i \neq 2}^{4} card(C_i \cap C_2) \neq 0$ It means the product of sets is not empty, so C_2 is not the peripheral one.
$C_3 \cap C_1 = \{Lisa, Pete\} = \{8, 5\}$
$C_3 \cap C_2 = \{President, Lisa, Pete\} = \{9, 8, 5\}$
$C_3 \cap C_4 = \emptyset$
It does not meet the above criterion, i.e. $\sum\limits_{i=1 \wedge i \neq 3}^{4} card(C_i \cap C_3) \neq 0$ $C_4 \cap C_1 = \emptyset$
$C_4 \cap C_2 = \emptyset$
$C_4 \cap C_3 = \emptyset$
$\sum\limits_{i=1 \wedge i \neq 4}^{4} card(C_i \cap C_4) = 0$, so clique C_4 according to Formula 6.3 is the peripheral clique. From now on C_4 clique is a peripheral clique and is excluded from the list of cliques. The groups that remain in the set of cliques are named from now on as regular cliques.

After the identification of regular cliques as well as peripheral nodes and peripheral cliques (see Table 6.28) the bridges can be uncovered. In the first

Table 6.28: The values of commitment function within the Thurman network

Regular Cliques	**Peripheral Nodes**	**Peripheral Cliques**
$C_1 = \{1, 2, 8, 3, 6, 5\}$ $C_2 = \{1, 8, 5, 9\}$ $C_3 = \{5, 12, 8, 9\}$	$PN = \{11, 13, 14, 15\}$	$PC = C_4 = \{4, 10, 7\}$

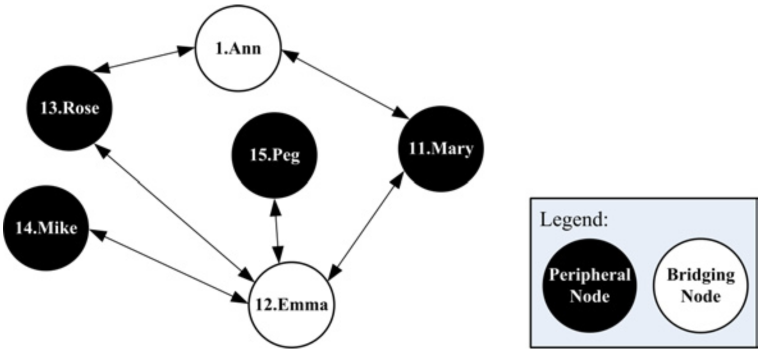


Figure 6.26: Structure of peripheral nodes in Thurman network and bridges that connect *PN* with the regular cliques

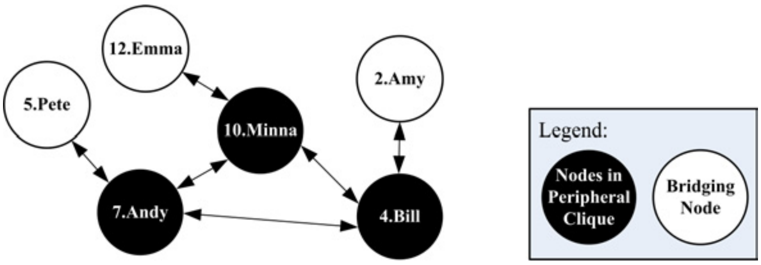


Figure 6.27: Structure of peripheral clique in Thurman network and bridges that connect *PC* with the regular cliques

step the bridges that connect the peripheral nodes with the regular cliques are uncovered. In order to perform this, all of the peripheral node's connections are analyzed (Figure 6.26). It can be easily noticed that Ann and Emma are bridges in this case, i.e. a set of bridges B_1 equals: $B_1 = \{Ann, Emma\} = \{1, 12\}$. Next, the analysis of peripheral clique and its relationships with the external network (Figure 6.27) shows that the bridging nodes in this case are $B_2 = \{Emma, Pete, Amy\} = \{12, 5, 2\}$ The final set of bridges B is the sum of the sets B_1 and B_2, so finally the set of bridges consist of:

$$B = \{Ann, Emma, Pete, Amy\} = \{1, 12, 5, 2\}$$

Properties of Bridges in Social Network

The goal of the next part of the experiments is to investigate the characteristic features of the extracted bridging nodes. For all of the nodes their node position is calculated as well as the number of mutual, incoming and outgoing edges. Before starting this part of the experiments some assumptions are made. The initial node positions $NP_0(x) = 1$ are established for every member x in the network. The value of ε is 0.9 and the stopping condition is: no difference in node position values to the precision of 5 digits after the decimal point for all the members in two following iterations, i.e. $\tau = 0.00001$. The calculated values are presented in Table 6.29. Note, that

Table 6.29: The node position values and number of edges in Thurman network

ID	Member	NP Ranking	NP	No. of mutual edges	No. of incoming edges	No. of outgoing edges
12	Emma	1	1.90025	8	0	0
1	Ann	2	1.56732	8	0	3
5	Pete	3	1.48140	8	0	5
2	Amy	4	1.38236	6	2	0
8	Lisa	5	1.36532	7	1	0
6	Tina	6	1.17424	5	2	0
3	Katy	7	1.01320	5	1	0
10	Minna	8	0.86255	3	2	2
4	Bill	9	0.79626	3	2	0
9	President	10	0.73712	4	0	9
7	Andy	11	0.63676	3	1	0
11	Mary	12	0.60897	2	2	0
13	Rose	12	0.60897	2	2	0
14	Mike	14	0.43264	1	2	0
15	Peg	14	0.43264	1	2	0

bridges identified in the first part of experiments posses high node positions and have the largest number of connections — Emma, Ann, Pete and Amy. Emma has the highest node position and she is the only community member that connects both peripheral nodes (all of them) and the peripheral clique with the regular cliques. Two of the peripheral users (Peg and Mike) communicate only with Emma so she is a crucial node when the cohesion of the whole network is concerned. Ann, who has the second highest node position and has 8 mutual connections, binds two peripheral users (Mary and Rose) with the regular cliques. However, these connections are not as crucial as relations between Emma and Peg or Mike, because not only Ann is connected with Mary and Rose. Another person that is connected with Mary and Rose is Emma. Pete, who obtained third highest node position, has also 8 mutual connections and is one of three people that connects the peripheral clique ({Minna, Bill, Andy}) with the regular cliques. Another users who connect the peripheral clique with the regular cliques are Emma and Amy. Amy was identified also as a bridging node and she obtained the fourth position in the *NP* ranking. Additionally, she possesses 6 mutual relationships — 2 relations less than the three other bridging nodes. On the other hand President, who has many connections but low node position (10th in *NP* ranking), is not a bridging node. It means that the node position is a better measure to identify the bridges than the number of the edges.

Case Study — Conclusions

The method of detecting bridging nodes in a directed and weighted social network as well as their properties were investigated. The conducted research reveals that the bridges obtain the highest node position. Two types of bridges can be distinguished: (i) these that connect peripheral nodes with regular cliques and (ii) these that connect peripheral cliques with the regular cliques. Note that the highest node position is obtained by Emma (12) who binds both all of the peripheral nodes and peripheral clique with the regular cliques. Ann (1), Pete (5) and Amy (2) also are the bridging nodes and all of them obtain high node positions: Ann — second, Pete — third and Amy — fourth. This reveals that node position is a good measure that can be used to find the bridging nodes within the network. In other words, there exists the correlation between the node position of the network and the fact if the node is a bridging one. The future work will focus on conducting the research within the complex social networks with big number of nodes and edges. We intend to develop a fast method of identification of the bridge nodes then to use it to track changes in community dynamics (group evolution). The changes in the properties of bridging nodes may be used to track processes of merging, splitting, growth and contraction of cliques in complex social networks.

Chapter 7

Conclusions and Future Work

Conclusions

This dissertation deals with Internet–based social networks, where both nodes and relations have clear technical interpretation. However, well defined in technical terms, networks of Internet users are not well analyzed due to dynamics and complexity (scale). Multidimensionality, hard–to–define before Internet, now may be investigated, but requires new algorithms and techniques. One of the algorithms proposed in this thesis that can be used in such a complex environment is the node position method that is used to discover the nodes that are important for a given Internet community. Important means that a node is perceived as the prominent by others and it is expressed by the fact that they communicate or share common activities with this node. Furthermore, the node is important if the nodes with high node position communicate with it because its position depends on the position of its neighbors.

Moreover, the definitions and detailed descriptions of such concepts as Internet Identity (*IID*), Internet Relationship R, Homogeneous Social Network *HSN*, System–based Social Network *SSN*, and Internet Multisystem Social Network *ISN* are proposed in the thesis. The author also proposes the classification of social networks existing in the Internet and presents examples for each of the created classes of social networks.

Both the formal analysis and the research on real world data was conducted in order to present the main features of the method and the influence of method parameters (coefficient ε and stopping condition τ) on the method outcome as well as to compare it with other centrality indices. Moreover, the structural neighborhood of the nodes with the highest and lowest node position were analyzed. Last part of the experiments was to investigate the complexity and efficiency of the proposed node position method.

General Characteristics of the Proposed Node Position Method

The node position provides a better opportunity to distinguish individuals within the network as opposed to other centrality measures (Section 6.1). In other words, node position measure is more diverse than the other measures. The research on Enron network confirm this assumption and it can be especially visible while analyzing number of nodes that possess the same centrality value, Figure 6.9 in Section 6.2. The number of duplicates that occurs in case of the node position method is smaller than in the case of other established centrality measures because $NP(x)$ takes into consideration not only the number of nodes who communicate to the evaluated node x but also their node positions and their contribution of activity directed to x. The fact that different aspects of node activity are taken into consideration results in **the node position reflecting better the real user position within the network of Interent users than other centrality indices**.

Moreover, the node position **provides an insight into the local structure and the role of the node within the network** (Section 6.1). The outcome of the research conducted on the Thurman network is the basis to claim that users with high *NP* tend to connect users with small number of connections i.e. these who are at the periphery of the network with the entire network. These users with high position can be treated as bridge that connects different subgroups of the network. This is confirmed by one of the possible, presented application of the measure, i.e. the extraction of the bridging nodes from the network of Internet users (see Section 6.5). The conducted research reveals that the bridges obtain the highest node position. Two types of bridges can be distinguished: (i) these that connect peripheral nodes with regular cliques and (ii) these that connect peripheral cliques with the regular cliques.

Another feature of the node position is that only few members exceed the average value that equals 1 — e.g. in the case of Enron network around 6% of users (Chapter 6.2). This confirms the assumption that the node position can be a good measure **to extract key persons in the network of Internet users**. This key users can be used in the process of target marketing and some specific services can be offered to them.

Influence of the Method Parameters on the Node Position Characteristics

In the node position method two parameters were introduced — coefficient ε and stopping condition τ.

The value of ε denotes the openness of node position measure on external influences: to what extent x's node position is 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 greater extent. Furthermore, the value of ε influences also the number of iterations that is required to meet a given stopping condition as well as other features of the method e.g. maximum value and standard deviation. The main observations regarding the influence of ε coefficient are as follows:

- ε influence the number of iterations that are required to meet the given stopping condition τ. The larger ε is, the larger number of iterations is needed and the dependence between these two variable is best fitted by the exponential growth function (Section 6.3).
- When ε is larger, then the distance between the minimum and maximum node position within community increases (Section 6.2). In consequence the larger ε the larger standard deviation. The dependence between these variables is linear. This results in rescaling the range of the node position values. For greater ε the wider range of node position values we have and this facilitates to distinguish the users. This also results in the smaller number of duplicates with the growth of ε value and the function, which approximates the dependence between ε and number of duplicates in a most accurate way, is the exponential decay function (Section 6.3).

The second parameter — stopping condition τ needed to be introduced because the calculations are iterative. All the presented experiments were performed with the following stopping condition: $\forall(x \in IID)|NP_n(x)–NP_{n\text{-}1}(x)| \leq \tau$ and $\tau \in \{0.1, 0.01, 0.001, 0.0001, 0.00001, 0.000001\}$. The iterative character of the method requires also that the initial node positions values need to be established before the beginning of the calculations. It if worth to notice that:

- The smaller τ the larger the number of iterations is required in order to meet this stopping condition and in consequence the processing time decreases (Section 6.3).
- The smaller τ the calculations are more precise, i.e. the number of duplicates decrease if τ is smaller (Section 6.3).
- Initial values of node positions — The values of initial node positions do not influence the final positions and the sum of all node positions within the network. But they influence the number of iterations. The smallest number of iterations is required when all nodes have the initial value equal 1. (Example in Section 4).

Efficiency

Note that ”the usefulness of centrality indices atands or falls with the ability to compute them quickly” [18]. Thus, one of the motivations to develop a new method of assessing the user centrality is that the existing methods tend to be very inefficient when applied to the complex social networks, such as these existing in the Internet.

The iterative nature of node position requires a certain number of iterations to be performed to achieve the required precision of results. However, the implementation of the general concept can be accomplished with different approaches. Three methods to calculate the node positions values were proposed — PIN^{nodes}, PIN^{edges}, PIN^{hybrid}. One of the most surprising conclusions from the tests carried out is the big difference in efficiency between these three methods, even over two orders of magnitudes. The "edge approach" appears to be absolutely the best while raw, direct implementation of the concept – PIN^{nodes} remains far behind. This reveals that the implementation method for some general concepts from social network analysis may have the crucial impact on the computation efficiency. The usage of edges instead of nodes to process data is also more effective for other centrality measures analyzed in the thesis, i.e. indegree and outdegree centrality. This reveals that the implementation method for some general concepts from social network analysis may have the crucial impact on the computation efficiency.

On one hand small value of τ parameter results in the large number of calculations that would slow down the process due to a large number of required iterations (Section 6.3). On the other hand too few iterations may cause the values of all node positions to be too close to each other that makes outcomes irrelevant. The second parameter — ε coefficient also influence the processing time, i.e. the greater ε the longer processing time but at the same time smaller number of duplicate values and greater distance between the minimum and maximum node position. This trade–off enables the researcher to pick the ε and τ that suit his/her needs. Such parametrization is not available in the case of centralities measures based on node degree or on the concept of shortest paths where very complex calculations must be performed in order to obtain any results.

To sum up all considerations the main achievement of this thesis is a proposition, description and analysis of the method of the node position evaluation that serves to discover the most prominent individuals within the network of Internet users (Chapter 4). Furthermore, the extensive experiments on different datasets were carried out in order to present the characteristics of the proposed method and to compare it with the existing centrality measures.

Future Work

Many possible extension of the proposed method exist. One of them is connected with the further development of commitment function that will take into consideration not only the quantitative communication between users but also its quality. It means that the text analysis of the communication can be performed in order to evaluate the relationship strength and its character. Moreover, deeper analysis of dynamics of the node position within the network of Internet users can provide the insight into the stability of the network structure and groups existing within it. Another, possible direction of the current research is to investigate the position of the whole cliques of the network not only the individuals. The node position method can be also applied to edges of the network of Internet users. A new measure called edge position can be developed based on this proposed in the thesis. This can be a new approach to identification and analysis of a nature and importance of both a single relationship and a whole tie. These various research issues connected with the node position method will be undoubtedly one of the scientific areas on which I would like to continual working.

References

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

[2] Adamic L.A., Adar E., "How to search a social network", *Social Networks, 257(3)*, pp. 187–203, 2005.

[3] Ahn Y-Y. Han S., Kwak H., Moon S., Jeong H., "Analysis of topological characteristics of huge online social networking services", In *Proceedings of the 16th International Conference on World Wide Web, ACM Press*, pp. 835–844, 2007.

[4] Alexander C.N., "A method for processing sociometric data", *Sociometry 26*, pp. 268–269, 1963.

[5] Amaral L.A.N., Scala A., Barthelemy M., Stanley H.E., "Classes of small–world networks", *Proceedings of the National Academy of Sciences USA, 97(21)*, pp. 11149–11152, 2000.

[6] Barnes J.A., "Class and Committees in a Norwegian Island Parish", *Human Relations 7*, pp. 39–58, 1954.

[7] Bavelas A., "Communication patterns in task – oriented groups", *Journal of the Acoustical Society of America 22*, pp. 271–282, 1950.

[8] Beauchamp., M.A. "An improved index of centrality", *Behavioral Science 10*, pp.161–163, 1973.

[9] Berge C., "Graphs and hypergraphs", *American Elsevier Pub. Co.*, 1973.

[10] Berkhin P., "A Survey on PageRank Computing", *Internet Mathematics 2(1)*, pp. 73–120, 2005.

[11] Bonacich P., "Factoring and weighting approaches to status scores and clique identification", *Journal of Mathematical Sociology 2*, pp. 113–120, 1972.

[12] Bonacich P., "Power and centrality: a family of measures", *American Journal of Sociology 92*, pp. 1170–1182, 1987.

[13] Bonacich P., Lloyd P., "Eigenvector-like Measures of Centrality for Asymetriv Relations", *Social Networks 23(3)*, pp. 191–201, 2001.

[14] Borgatti S.P., "Centrality and network flow", *Social Networks 27(1)*, pp. 55–71, 2005.

[15] 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.

[16] Boyd D.M., "Friendster and Publicly Articulated Social Networking", In *In Proceedings of CHI '04: CHI '04 extended abstracts on Human factors in computing systems, ACM Press*, pp. 1279–1282, 2004.

[17] Boyd D.M., Ellison N.B., "Social network sites: Definition, history, and scholarship", *Journal of Computer-Mediated Communication 13*, pp. 210–230, 2007.

[18] Brandes U., Erlebach T., "Network Analysis, Methodological Foundations", *Springer – Verlag, Berlin, Heidelberg, Germany*, 2005.

[19] Breiger R.L., "The Analysis of Social Networks", In *Handbook of Data Anaysis, Hardy, M., Bryman, A. (eds), London, SAGE Publications*, pp. 505–526, 2004.

[20] Breslin J.G., Harth A., Bojars U., Decker S., "Towards Semantically–Interlinked Online Communities, The Semantic Web: Research and Applications", In *Proceedings of the Second European Semantic Web Conference, Lecture Notes in Computer Science 3532, Springer*, 2005.

[21] Brickley D., MillerL., "FOAF Vocabulary Specification. Technical report", *Technical report, RDF Web FOAF Project*, 2003.

[22] 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.

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

[24] Burt R.S., "Structural holes", *Cambridge, MA: Harvard University Press*, 1992.

[25] Carpenter T., Karakostas G., Shallcross D., "Practical Issues and Algorithms for Analyzing Terrorist Networks", *Invited talk at WMC 2002*, 2002.

[26] Carrington P., Scott J., Wasserman S., "Models and methods in Social Network Analysis", *Cambrige University Press, Cambrige*, 2005.

[27] Cattell V., "Poor people, poor places, and poor health: the mediating role of social networks and social capital", *Social Science and Medicine 52(10)*, pp. 1501–1516, 2001.

[28] Caverlee J., Webb S., "A Large-Scale Study of MySpace: Observations and Implications for Online Social Networks," In *Proceedings of the 3rd International AAAI Conference on Weblogs and Social Media, AAAI Press*, pp. 36–44, 2008.

[29] Chiu P.Y., Cheung C.M.K., Lee M.K.O., "Online Social Networks: Why Do "We" Use Facebook?", In *Proceedings of the First World Summit on the Knowledge Society, Springer, Communications in Computer and Information Science, 19*, pp. 67–74, 2008.

[30] Creese J., Cribb J., Spicer J., "Social Networking: never mind the students, what about us? Use of Social Networking Softwares for professional networking and development for library staff", In *Proceedings of Beyond The Hype: Web 2.0 Symposium*, 2008.

[31] Culotta A., Bekkerman R., McCallum A., "Extracting social networks and contact information from email and the Web", In *Proceedings of the First Conference on Email and Anti-Spam*, 2004.

[32] Davis J. A., "Clustering and structural balance in graphs", *Human Relations 20*, pp. 181–187, 1967.

[33] Degenne A., Forse M., "Introducing social networks", *London: SAGE Publications Ltd*, 1999.

[34] Desautels B., "Social networking: LinkedIn.com", *Journal of Leadership Studies, 2(2)*, pp. 103–104, 2008.

[35] Dijkstra E.W., "A note on two problems in connection with graphs", *Numerische Mathematik 1*, pp. 269–271, 1959.

[36] DiMicco J., Millen D.R., "Identity management: multiple presentations of self in facebook", In *Proceedings of the International ACM SIGGROUP Conference on Supporting Group Work, ACM Press*, pp. 383–386, 2007.

[37] DiMicco J., Millen D.R., Geyer W., Dugan C., Brownholtz B., Muller M., "Motivations for Social Networking at Work", In *Proceedings of the Computer Supported Cooperative Work 2008 Conference, ACM Press*, pp. 711–720, 2008.

[38] Donath J.S., "Identity and deception in the virtual community", *Chapter 2 in Smith M.A., Kollock P. (eds.), Communities in cyberspace, Routledge, London & New York*, pp. 29–59, 1999.

[39] Dunbar R., ”Coevolution of neocortical size, group size and language in humans”, *Behavioral and Brain Sciences 16*, pp. 681–735, 1993.

[40] Dynes S., Gloor P., Laubacher R., Zhao Y., ”Temporal Visualization and Analysis of Social Networks”, *North American Association for Computational Social and Organizational Science Conference, Pittsburgh PA*, 2004.

[41] Flake, G., Lawrence, S., Lee Giles, C., ”Efficient identification of web communities”, In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.150–160, 2000.

[42] Floyd R.W., ”Algorithm 97: Shortest Path”, *Communications of the ACM 5(6)*, pp.345, 1962.

[43] Freeman L.C., ”A set of measures of centrality based on betweenness”, *Sociometry 40*, pp.35–41, 1977.

[44] Freeman L.C., Roeder D., Mulholland R.R., ”Centrality in Social Networks: II. Experimental Results”, *Social Networks 2(2)*, pp.119–141, 1980.

[45] Friedkin N.E., ”Theoretical foundations for centrality measures”, *American Journal of Sociology 96*, pp.1478–1504, 1991.

[46] Garton L., Haythorntwaite C., Wellman B., ”Studying Online Social Networks”, *Journal of Computer-Mediated Communication 3(1)*, 1997.

[47] Gibson D., Kleinberg J., Raghavan P., ”Inferring Web communities from link topology”, In *Proceedings of the Ninth ACM Conference on Hypertext and Hypermedia*, 1998.

[48] Golbeck J., ”Computing and Applying Trust in Web-Based Social Networks”, *Dissertation Submitted to the Faculty of the Graduate School of th Universtity of Maryland, College Park in partial fulfillment of the requirements for the degree of Doctor of Philosophy*, 2005.

[49] Golbeck J., Hendler J., ”FilmTrust: movie recommendations using trust in web-based social networks,” In *Proceedings of Consumer Communications and Networking Conference, IEEE Conference Proceedings 1*, pp. 282–286, 2006.

[50] Golder S., Huberman B.A., ”Usage Patterns of Collaborative Tagging Systems”, *Journal of Information Science 32(2)*, pp. 198–208, 2006.

[51] Golbeck J., Rothstein M., ”Linking Social Networks on the Web with FOAF: A Semantic Web Case Study”, In *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence, AAAI Press*, pp. 1138–1143, 2008.

[52] Granovetter M.S., ”The strength of weak ties: A Network Theory Revisited”, *Sociological Theory 1*, pp. 201–233, 1983.

[53] Gross R., Acquisti A., ”Information revelation and privacy in online social networks”, In *Proceedings of ACM workshop on Privacy in the electronic society Alexandria, ACM Press*, pp. 71–80, 2005.

[54] Hakimi S.L., ”Optimum location of switching centers and the absolute centers and medians of a graph”, *Operations Research 12*, pp.450–459, 1964.

[55] Hanneman R., Riddle M., ”Introduction to social network methods,” *online textbook, available from Internet: http://faculty.ucr.edu/ hanneman/nettext/, (01.04.2006)*, 2005.

[56] Harary F., Norman R.Z., Cartwright D., ”Structural Models: An Introduction to the Theory of Directed Graphs”, *New York, John Wiley and Sons*, 1965.

[57] Harth A., ”SECO: Mediation Services for Semantic Web Data”, *IEEE Intelligent Systems, Special Issue on Semantic Web Challenge 19(3)*, pp. 66–71, 2004.

[58] Hatala J.P., ”Social Network Analysis in Human Resources Development: A New Methodology”, *Human Resource Development Review 5(1)*, pp. 45–71, 2006.

[59] Heymann P., Koutrika G., Garcia-Molina H., ”Fighting Spam on Social Web Sites: A Survey of Approaches and Future Challenges”, *IEEE Internet Computing 11(6)*, pp. 36–45, 2007.

[60] Hill R., Dunbar R., ”Social Network Size in Humans”, *Human Nature 14*, pp. 53–72, 2002.

[61] Howard B., ”Analyzing online social networks”, *Communication of the ACM 51(11)*, pp. 14–16, 2008.

[62] Hubbell C.H., ”An input–output approach to clique detection”, *Sociometry 28*, pp. 277–299, 1965.

[63] Jeffreys H., Jeffreys B.S., ”Increasing and Decreasing Functions”, *Methods of Mathematical Physics, 3rd ed., Cambridge University Press, Cambridge, England, 22*, 1988.

[64] Jordan K., Hauser J., Foster S., ”The Augmented Social Network: Building identity and trust into the next-generation Internet”, *First Monday 8(8),http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/-article/viewArticle/1068*, 2003.

[65] Juszczyszyn K., Musiał A., Musiał K., Bródka P., "Molecular Dynamics Modelling of the Temporal Changes in Complex Networks", In *Proceedings of IEEE Congress on Evolutionary Computing, IEEE Proceedings*, accepted, 2009.

[66] Katz L., "A new status derived from sociometrics analysis", *Psychometrica 18*, pp.39–43, 1953.

[67] Kazienko P., Musial K., "Recommendation Framework for Online Social Networks", In *Proceedings of the 4th Atlantic Web Intelligence Conference, Studies in Computational Intelligence, Springer Verlag*, pp. 111–120, 2006.

[68] Kazienko P., Musiał K., "Assessment of Personal Importance Based on Social Networks", In *Proceedings of the 6th Mexican International Conference on Artificial Intelligence, Lecture Notes in Artificial Intelligence LNAI 4827, Springer Verlag*, pp. 529–539, 2007.

[69] Kazienko P., Musiał K., "On Utilising Social Networks to Discover Representatives of Human Communities", *International Journal of Intelligent Information and Database Systems, Special Issue on Knowledge Dynamics in Semantic Web and Social Networks 1(3/4)*, pp.293–310, 2007.

[70] Kazienko P., Musial K., "Social position of Individuals in Virtual Social Networks", *Journal of Mathematical Sociology, submitted*, 2009.

[71] Kazienko P., Musiał K., Kajdanowicz T., "Profile of the Social Network in Photo Sharing Systems", In *Proceedings of the 14th Americas Conference on Information Systems, Mini-track: Social Network Analysis in IS Research, Association for Information Systems (AIS), ISBN: 978-0-615-23693-3*, 2008.

[72] Kazienko P., Musiał K., Kajdanowicz T., "Multidimensional Social Network and Its Application to the Social Recommender System", *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, accepted, 2009.

[73] Kazienko P., Musiał K., Zgrzywa A., "Evaluation of Node Position Based on Email Communication", *Control and Cybernetics 38(1)*, in press, 2009.

[74] Kendall, M.G., "Rank correlation methods", *Charles Griffin & Company, Ltd., London*, 1948.

[75] Kennedy H., "Beyond anonymity, or future directions for internet identity research", *New Media & Society 8(6)*, pp. 859–876, 2006.

[76] Kinsella S., Breslin J.G., Passant A., Decker S., "Applications of Semantic Web Methodologies and Techniques to Social Networks and Social Websites, Reasoning Web", In*Proceedings of the 4th International Summer School 2008, Lecture Notes in Computer Science LNCS 5224, Springer*, pp.171–199, 2008.

[77] 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.

[78] Kostakos V., O'Neil E., Jones S., "Social networking 2.0", In *Proceedings of the Conference on Human Factors in Computing Systems, ACM Press*, pp. 3381–3386, 2008.

[79] Krebs V., "The Social Life of Routers", Internet Protocol Journal 3, pp. 14–25, 2000.

[80] Kumar R., Raghavan P., Rajagopalan S., Tomkins A., "The Web and Social Networks", *IEEE Computer 35(11)*, pp. 32–36, 2002.

[81] Lazega E., "The Collegial Phenomenon. The Social Mechanism of Co–operation Among Peers in a Corporate Law Partnership", *Oxford University Press, Oxford*, 2001.

[82] Leskovec J., Backstrom L., Kumar R., Tomkins A., "Microscopic evolution of social networks", In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM Press*, pp. 462–470, 2008.

[83] Lévy P., "Collective Intelligence: Mankind's Emerging World in Cyberspace", *Perseus Books Cambridge, MA, USA*, 1997.

[84] Liben-Nowell D., Kleinberg J., "The Link Prediction Problem for Social Networks", In *Proceedings of the 12th International Conference on Information and Knowledge Management, ACM Press*, pp. 556–559, 2003.

[85] Luce R.D., Perry A.D, "A method of matrix analysis of group structure", *Psychometrica, 14*, pp. 95–116, 1949.

[86] Marsden P., Campbell K.E., "Measuring tie strength", *Social Forces 63*, pp. 482–501, 1984.

[87] Marx K., "Selected Writings in Sociology and Social Philosophy", tr. T.B. Bottomore, *New York: McGraw–Hill*, 1956.

[88] McCallum A., Corrada-Emmanuel A., Wang X., "The Author-Recipient-Topic Model for Topic and Role Discovery in Social Networks, with Application to Enron and Academic Email", In *Proceedings of the SIAM International Conference on Data Mining, Workshop on Link Analysis, Counterterrorism and Security*, pp. 33–44, 2005.

[89] Mesnage C., Jazayeri M., "Specifying the Collaborative Tagging System", In *Proceedings of the 1st Semantic Annotation and Authoring Workshop co-located with The 5th International Semantic Web Conference, available at: http://myunderstanding.files.wordpress.com/2006/09/- mesnage-jazayeri06bfinal.pdf (accessed March 2009)*, 2006.

[90] Mika P., "Social networks and the semantic web: the next challenge", *IEEE Intelligent Systems 20(1)*, pp. 82–85, 2005.

[91] Milgram S., "The Small–World Problem", *Psychology Today 2*, pp. 60–67, 1967.

[92] Millen D., Feinberg J., Kerr B., "Social bookmarking in the enterprise", *Queue 3(9), ACM Press*, pp. 28–35, 2005.

[93] Montgomery J., "Social Networks and Labor-Market Outcomes: Toward an Economic Analysis", *American Economic Review 81(5)*, pp. 1407–1418, 1991.

[94] 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.

[95] Morris M., "Sexual network and HIV", *AIDS 11*, pp. 206–219, 1997.

[96] Musial K., Bródka P., Kazienko P., "A Performance of Centrality Calculation in Social Networks", In *Proceedings of the International Conference on Social Computing*, submitted, 2009.

[97] Musial K., Kazienko P., Kajdanowicz T., "Multirelational Social Networks in Multimedia Sharing Systems", *Chapter in: Knowledge Processing and Reasoning for Information Society, Eds. N.T. Nguyen, G. Kołaczek, B. Gabrys, Academic Publishing House EXIT*, pp. 275–292, 2008.

[98] Newman, M.E.J., "The structure of scientific collaboration networks", *National Academy of Sciences USA 98*, pp. 404–409, 2001.

[99] O'Murchu I., Breslin J.G., Decker S., "Online Social and Business Networking Communities", In *Proceedings of the Workshop on Application of Semantic Web Technologies to Web Communities, CEUR Workshop*, pp. 107, 2004.

[100] O'Reilly T., "What is Web 2.0: Design Patterns and Business Models for the Next Generation of Software", *Communications & Strategies 1*, pp. 17, 2007.

[101] Pagel M., Erdly W., Becker J., "Social networks: we get by with (and in spite of) a little help from our friends", *Journal of Personality and Social Psychology 53(4)*, pp. 793–804, 1987.

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

[103] 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.

[104] Rak J., "The Digital Queer: Weblogs and Internet Identity", *Biography 28(1)*, pp. 166–182, 2005.

[105] Recordon D., Reed D., "OpenID 2.0: a platform for user-centric identity management", In *Proceedings of the 2006 Workshop on Digital Identity Management, ACM Press*, pp. 11–16, 2006.

[106] Robins G.L., Alexander M., "Small worlds among interlocking directors: network structure and distance in bipartite graphs", *Computational & Mathematical Organization Theory 10*, pp. 69–94, 2004.

[107] Rogers E., "Progress, problems and prospects for network research: Investigating relationships in the age of electronic communication technologies", In *Proceedings of the Clearwater Beach, FL: Sunbelt Social Network Conference*, 1987.

[108] Sabidussi G., "The centrality index of a graph", *Psychmetrica 31(4)*, 1966.

[109] Schuler D., "Social Computing", *Communications of the ACM 37(1)*, pp. 28–29, 1994.

[110] Scott J., "Social network analysis: A handbook (2nd ed.)", *London: Sage*, 2000.

[111] Shaw M.E., "Group structure and the behavior of individuals in small groups", *Journal of Psychology 38*, pp. 139–149, 1954.

[112] Shetty J., Adibi J., "Discovering Important Nodes through Graph Entropy The Case of Enron Email Database", In *Proceedings of the 3rd International Workshop on Link Discovery, ACM Press*, pp. 74–81, 2005.

[113] Thurman B., "In the office: Networks and coalitions", *Social Networks 2*, pp. 47–63, 1979.

[114] Travers J., Milgram S., "An experimental study of the small world problem," *Sociometry 32(4)*, pp. 425–443, 1969.

[115] Tutzaue F., "Entropy as a measure of centrality in networks characterized by path-transfer flow", *Social Networks 29(2)*, pp. 249–265, 2007.

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

[117] Walker K., "It's Difficult to Hide It, The Presentation of Self on Internet Home Pages", *Qualitative Sociology, 23(1)*, pp. 99–120, 2000.

[118] Wang F.Y., Carley K.M. Zeng D., Wenji M., "Social Computing: From Social Informatics to Social Intelligence", *Intelligent Systems, IEEE*, pp. 79–83, 2007.

[119] Wasserman S., Faust K., "Social network analysis: Methods and applications", *New York: Cambridge University Press*, 1994.

[120] Wellman B., Salaff J., Dimitrova D., Garton L., Gulia M., Haythornthwaite C., "Computer Networks as Social Networks: Collaborative Work, Telework, and Virtual Community", *Annual Review Sociology 22*, pp. 213–238, 1996.

[121] Watts D. J., "Small Worlds: The Dynamics of Networks Between Order and Randomness", *Princeton University Press, USA*, 1999.

[122] Weaver A.C., Morrison B.B., "Social Networking", *Computer 41(2)*, pp. 97–100, 2008.

[123] Wellman B., Wortley S., "Different strokes from different folks: Community ties and social support", *American Journal of Sociology 96*, pp. 558–588, 1990.

[124] Wood A.F., Smith M.J., "Online Communication: Linking Technology", *Identity & Culture. Second Edition. Routledge, New York*, 2004.

[125] Xing W., Ghorbani A., "Weighted PageRank Algorithm", In *Proceedings of the 2nd Annual Conference on Communication Networks and Services Research*, pp. 305–314, 2004.

[126] Yang W.S., Dia J.B., Cheng H.Ch., Lin H.T., "Mining Social Networks for Targeted Advertising", In *Proceedings of the 39th Hawaii International International Conference on Systems Science, IEEE Computer Society*, pp. 425–443, 2006.

[127] Zhu W., Chen C., and Allen R.B., "Visualizing an enterprise social network from email", In *Proceedings of the ACM/IEEE-CS joint Conference on Digital Libraries, ACM Press*, pp. 383, 2006.

Appendix A — Methods to calculate the shortest path in the network

Dijkstra Single Source Shortest Path Algorithm

Dijkstra [35] provided the first polynomial–time algorithm for the Single Source Shortest Path (SSSP) for graphs with non–negative edge weights. The output of the algorithm are the shortest path distances $d(x, v)$ from the node x to all $v \in M$.

```
**Input:** Graph $G = (V, E)$, edge weights $E \to \mathcal{R}$, source vertex $x \in V$
**Output:** Shortest path distances $d(x, v)$ to all $v \in V$
$P = \phi$ (empty set), $T = V$
$d(x, v) = \infty$ for all $v \in M$, $d(x, x) = 0$, $pred(x) = 0$
while $P \neq V$ do
  begin
    $v = argmind(x, v) | v \in T$
    $P := P \cup v, T := T \setminus v$
    for $w \in N(v)$ do
      if $d(x, w) > d(x, v) + \omega(v, w)$ then
      begin
$d(x, w) := d(x, v) + \omega(v, w)$
$pred(w) = v$
      end
  end
```

The algorithm starts by marking the source vertex x as permanent and inserting it into P, scanning all its neighbors $N(x)$, and setting the labels for the neighbors $v \in N(x)$ to the edge lengths: $d(x, v) = \omega(x, v)$. Next, the algorithm repeatedly removes a non–permanent vertex v with minimum label $d(x, v)$ from T, marks v as permanent, and scans all its neighbors $w \in N(v)$. If this scan discovers a new shortest path to w using the edge (v, w), then the label $d(x, w)$ is updated accordingly. The algorithm relies upon a priority queue for finding the next node to be marked as permanent. Implementing this priority queue as a Fibonacci heap, Dijkstra's algorithm runs in time $O(m + nlogn)$. For unit edge weights, the priority queue can be replaced by a regular queue. Then, the algorithm boils down to Breadth-First Search (BFS), taking $O(m + n)$ time.

Floyd–Warshall All–Pairs Shortest Paths Algorithm

Floyd–Warshall's All–Pairs Shortest Paths Algorithm (APSP) [42] first initializes all distance labels to infinity, and then sets the distance labels $d(u, v)$, for $u, v \in E$, to the edge weight $\omega(u, v)$. After this initialization, the algorithm basically checks whether there exists a vertex triple u, v, w for which the distance labels violate the condition:

$$d(u, w) \leq d(u, v) + d(v, w) \forall (u, v, w \in V)$$

If so, it decreases the involved distance label $d(u, w)$. This check is performed in a triple for–loop over the vertices. Since we are looking for all–pairs shortest paths, the algorithm maintains a set of predecessor indices $pred(u, v)$ that contain the predecessor vertex of v on some shortest path from u to v.

```
**Input:** Graph $G = (V, E)$, edge weights $E \rightarrow \mathcal{R}$, source vertex $x \in V$
**Output:** Shortest path distances $d(x, v)$ to all $v \in V$
$d(u, v) = \infty$, $pred(u, v) = 0$ for all $u, v \in V$
$d(v, v) = 0$ for all $v \in V$
$d(u, v) = \omega(u, v)$, $pred(u, v) = u$ for all $u, v \in E$
for $v \in V$ do
  for $u, w \in VxV$ do
    if $d(u, w) > d(u, v) + d(v, w)$ then
      begin
        $d(u, w) := d(u, v) + d(v, w)$
        $pred(u, w) := pred(v, w)$
      end
```

Appendix B — Node Position and Commitment Function in Multirelational Network of Internet Users

Relation Layers in the Flickr Photo Sharing System

The concept of *SSN* (see Section 2.2.1) and ties (see Section 2.4.4) that aggregate different types of relationships was applied to the Flickr photo sharing system [72], in which multimedia objects (*MO*s) are photos. Users can publish their pictures in Flickr, mark them with tags, create groups and attach their photos to them, build their own lists of favorite photos published by others, maintain contact lists linking to their acquaintances as well as comment photos authored by others. All these activities reflect common interests or acquaintances between users and enable to create the multirelational social network.

Eleven types of relations were identified in the system: relations based on contact lists — R^c, R^{rc}, R^{coc}, shared tags used by more than one user — R^t, user groups — R^g, photos added by users to their favorites — R^{ff}, R^{fa}, R^{af}, and opinions about pictures created by users — R^{oo}, R^{oa}, R^{ao}. Relations based on contact lists (R^c, R^{rc}, R^{coc}) represent direct intentional relations. Tag–based (R^t), group–based (R^g), favorite–favorite (R^{ff}), and opinion–opinion relations (R^{oo}) are typical object–based relations with equal roles, whereas favorite–author (R^{fa}), author–favorite (R^{af}), opinion–author (R^{oa}), and author–opinion (R^{ao}) are object–based relations with different roles. All these relations correspond to eleven separate layers in one multirelational social network, Figure 7.1.

Each relationship is extracted from the data about user behavior and for all relationships the values of commitment function can be assigned. These values express the strength of the relationships and are specific for each relation layer. Overall, the greater user iid_i's activity towards user iid_j among all activities of iid_i, the higher value of commitment function for relationship from iid_i to iid_j.

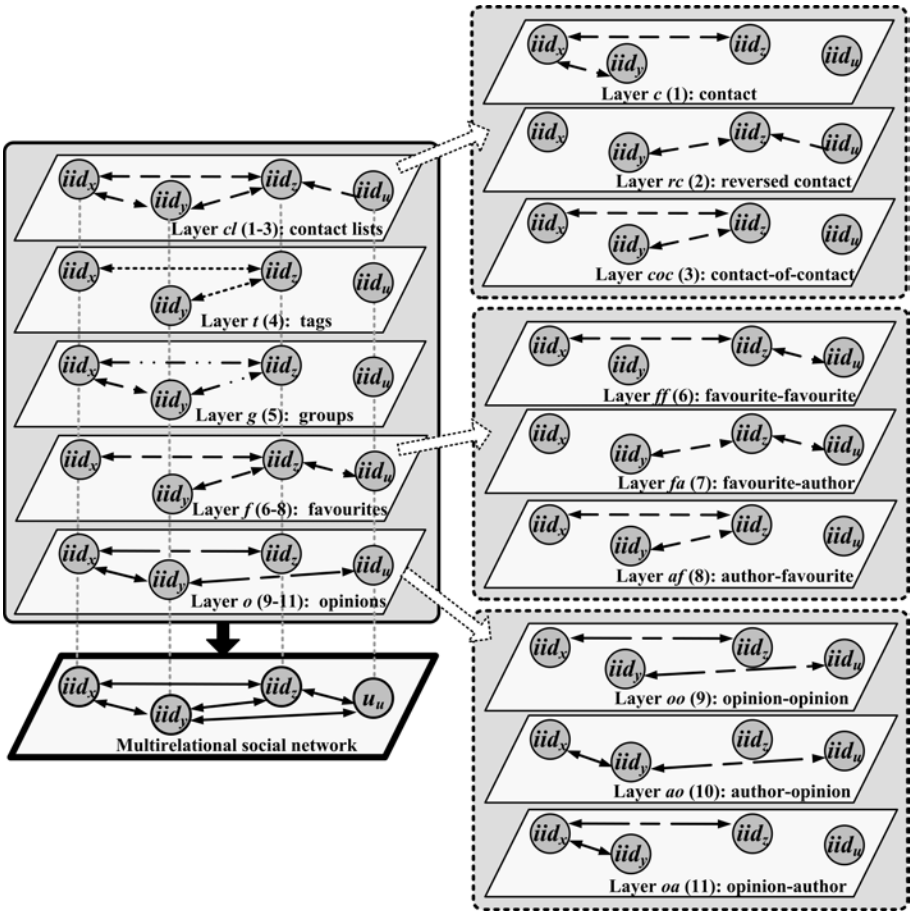

iid_x
iid_y
iid_z
iid_u
Layer *cl* (1-3): contact lists
Layer *t* (4): tags
Layer *g* (5): groups
Layer *f* (6-8): favourites
Layer *o* (9-11): opinions
u_u
Multirelational social network
Layer *c* (1): contact
Layer *rc* (2): reversed contact
Layer *coc* (3): contact-of-contact
Layer *ff* (6): favourite-favourite
Layer *fa* (7): favourite-author
Layer *af* (8): author-favourite
Layer *oo* (9): opinion-opinion
Layer *ao* (10): author-opinion
Layer *oa* (11): opinion-author

Figure 7.1: The relation layers in Flickr

Relationships Based on Contact List

The information about user iid_i's relationships based on contacts is derived directly from iid_i's contact list (CL_i), Figure 7.2. The relation r^c_{ij} from user iid_i to iid_j denotes that iid_j belongs to iid_i's contact list, Figure 7.2a. The strength value $s^c(iid_i, iid_j)$[1] of the relation $r^c(iid_i, iid_j)$[2] is calculated as follows:

$$s^c_{ij} = 1/n^c_i \text{ if } iid_j \text{ is in the } iid_i\text{'s contact list.} \tag{7.1}$$

where:
$n^c_i = card(CL_i)$ is a number of all iid_i's relations derived from iid_i's contact list, i.e. the length of iid_i's contact list CL_i.

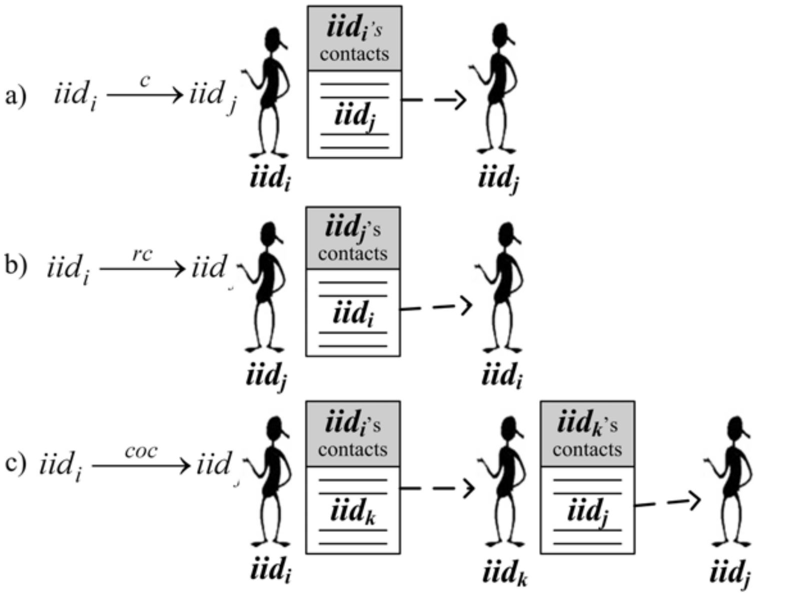


Figure 7.2: Relation layers extracted from contact lists: R^c (a), R^{rc} (b), and R^{coc} (c)

Based on the proposed relationship strength function s^c_{ij} the commitment function C^c_{ij} that denotes the contribution of activity of user iid_i towards user iid_j is calculated:

$$C^c_{ij} = s^c_{ij} \tag{7.2}$$

The relation r^{rc}_{ij} from user iid_i to iid_j denotes that iid_i belongs to iid_j's contact list and is called reversed–contact relation, Figure 7.2b. The strength value s^{rc}_{ij} of the relation r^{rc}_{ij} is calculated as follows:

$$s^{rc}_{ij} = 1/n^c_j \text{ if } iid_i \text{ is in the } iid_j\text{'s contact list.} \tag{7.3}$$

[1] In order to simplify the notation: $s^c(iid_i, iid_j) \equiv s^c_{ij}$

[2] In order to simplify the notation: $r^c(iid_i, iid_j) \equiv r^c_{ij}$

where:
$n_j^c = card(CL_j)$ is a number of all iid_j's relations derived from iid_j's contact list.

Another way of s_{ij}^{rc} calculation could be considered: $s_{ij}^{rc} = 1/n_i^c$. In contrary to Equation 7.3, it would underline the importance of user iid_i for user iid_j denoting whether iid_i is either one of many or one of only few among iid_j acquaintances.

Based on s_{ij}^{rc}, the commitment function C_{ij}^{rc} from user iid_i to iid_j from a reversed contact layer can be proposed:

$$C_{ij}^{rc} = \frac{s_{ij}^{rc}}{\sum\limits_{m=1}^{m_i^c} s_{im}^{rc}} \tag{7.4}$$

where:
$m_i^c = n_i^c = card(CL_i)$ is a number of all iid_i's relations derived from iid_i's contact list.

The indirect relation r_{ij}^{coc} from user iid_i to iid_j denotes that there exists another user iid_k that belongs to iid_i's contact list and iid_j is on the contact list of iid_k, Figure 7.2c. Therefore, it represents 'contact of contact' relation, which refers to 'friend of the friend' type of relationship. The strength value s_{ij}^{coc} of the relation r_{ij}^{coc} is calculated as follows:

$$s_{ij}^{coc} = \frac{n_i^{coc}}{n_i^c} \tag{7.5}$$

where:
${n_i}^{coc}$ is a number of different users iid_k on iid_i's contact list who simultaneously have user iid_j on their contact lists.

Based on s_{ij}^{coc}, the commitment function C_{ij}^{co} for relationships from a 'contact of contact' layer can be established:

$$C_{ij}^{coc} = \frac{s_{ij}^{coc}}{\sum\limits_{m=1}^{m_i^c} s_{im}^{coc}} \tag{7.6}$$

where:
$m_i^c = n_i^c = card(CL_i)$ is a number of all iid_i's relations derived from iid_i's contact list.

Relationships Based on Tags

The tag–based relationship r_{ij}^t between user iid_i and iid_j can be derived from information about tags they share. The general idea of extraction of tag–based relations from raw data is illustrated in Figure 7.3.

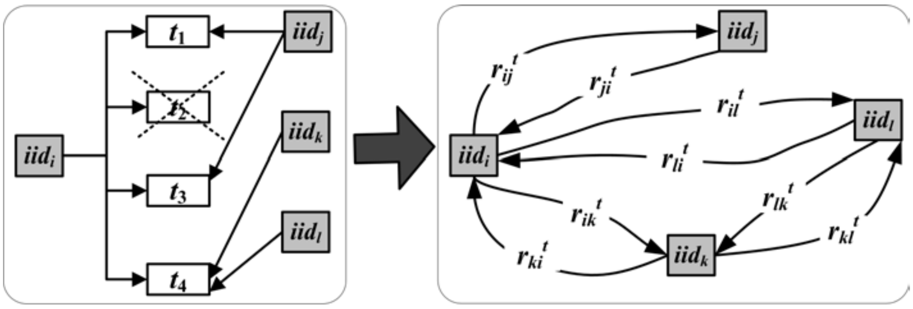


Figure 7.3: Extraction of tag–based relations

All tags that have already been used by at least two users form a set of shared tags. The relation r_{ij}^t between two users iid_i and iid_j exists if both of them have used at least one common tag to describe their photos. The strength value s_{ij}^t of such relation is expressed in the following way:

$$s_{ij}^t = \frac{n_{ij}^t}{n_i^t} \tag{7.7}$$

where:
n_{ij}^t is a number of tags common for users iid_i and iid_j;
n_i^t is a number of tags used by iid_i.

Note that it is not important how many times tag t_k was used by two users (to how many photos) but crucial is the fact that t_k was shared at least once.

Based on s_{ij}^t, the commitment function C_{ij}^t for relationships from a tag layer can be established:

$$C_{ij}^t = \frac{s_{ij}^t}{\sum\limits_{m=1}^{m_i^t} s_{im}^t} \tag{7.8}$$

where:
m_i^t is a number of all iid_i's relationships derived from tag layer.

Tag–based relation is an object–based relation with equal roles since all users have the same role towards the picture they tag.

Relationships Based on Groups

The data about groups to which user iid_i belongs enable to create the relations based on groups. A group contains some *MO*s published by a set of authors and for that reason it aggregates authors (group members) of photos placed in it. Let G be the set of all groups that consists of more than one member. The group–based relation r_{ij}^g from user iid_i to iid_j means that users iid_i and iid_j belong to at least one common group $g_k \in G$ or to be precise there are some groups that contain photos authored by iid_i and simultaneously some photos published by iid_j. The strength value s^g_{ij} of r^g_{ij} is:

$$s^g_{ij} = \frac{n^g_{ij}}{n^g_i} \tag{7.9}$$

where:
n^g_{ij} is a number of groups to which belong *MO*s published by both users iid_i and iid_j;
n^g_i is a number of groups containing user iid_i's photos.

Based on s^g_{ij}, the commitment function C^g_{ij} from user iid_i to iid_j for relationships derived from a group layer can be established:

$$C^g_{ij} = \frac{s^g_{ij}}{\sum_{m=1}^{m^g_i} s^g_{im}} \tag{7.10}$$

where:
m^g_i is a number of all iid_i's relationships derived from group layer.

Relationships Based on List of Favorites

The next three types of relations are obtained from the data about photos that have been added by some users to their favorites (Figure 7.4). The relation favorite–favorite r^{ff}_{ij} from user iid_i to iid_j exists if both users marked at least one common photo as their favorite. The relation author–favorite r^{af}_{ij} from author iid_i to user iid_j means that user iid_j has marked at least one iid_i's photo as iid_j's favorite. The relation r^{af}_{ij} simultaneously results in another relation: favorite-author r^{fa}_{ji} from user iid_j to author iid_i. Similarly, r^{ff}_{ij} results in r^{ff}_{ji} . For example, when the photo MO_m authored by the new user iid_i was marked as favorite by the first user iid_j, then this fact creates two new relations r^{af}_{ij} and r^{fa}_{ji}. When another user iid_k marks the same photo MO_m then four new relations are generated: r^{af}_{ik}, r^{fa}_{ki}, r^{ff}_{jk} and r^{ff}_{kj}.

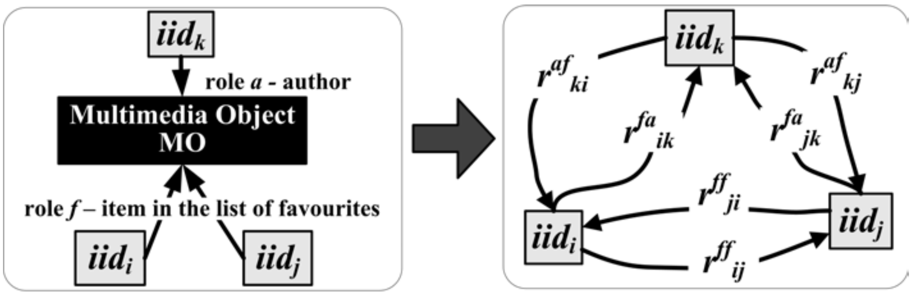


Figure 7.4: Extraction of relations based on favorites

The strength value s_{ij}^{ff} of relation r_{ij}^{ff} is calculated as follows:

$$s_{ij}^{ff} = \frac{n_{ij}^{ff}}{n_i^f} \tag{7.11}$$

where:
n_{ij}^{ff}, n_i^f is a number of photos marked as favorite simultaneously by user iid_i and iid_j or only by user iid_i, respectively.

To evaluate strength value s_{ij}^{af} of relation r_{ij}^{af} the following formula is used:

$$s_{ij}^{af} = \frac{n_{ji}^{fa}}{n_i^a} \tag{7.12}$$

where:
n_{ji}^{fa} is a number of photos marked as favorite by user iid_j and authored by iid_i;
n_i^a is a number of all photos added to a system by iid_i and marked by others as their favorite.

Finally, the formula for strength s_{ij}^{fa} of relation r_{ij}^{fa} is:

$$s_{ij}^{fa} = \frac{n_{ij}^{fa}}{n_i^f} \tag{7.13}$$

where:
n_{ij}^{fa} is a number of photos marked as favorite by user iid_i and authored by iid_j;
n_i^f is a total number of photos marked as favorite by user iid_i.

Relations based on favorites are kind of object–based relation with either equal (R^{ff}) or different roles (R^{af}, R^{fa}).

Based on s_{ij}^{ff}, the commitment function C_{ij}^{ff} from user iid_i to iid_j for relationships derived from a favorite–favorite layer can be established:

$$C_{ij}^{ff} = \frac{s_{ij}^{ff}}{\sum\limits_{m=1}^{m_i^{ff}} s_{im}^{ff}} \tag{7.14}$$

where:
m_i^{ff} is a number of all iid_i's relationships derived from favorite–favorite layer.

The commitment values for r_{ij}^{af} and r_{ij}^{fa} are calculated in analogous way, i.e.:

$$C_{ij}^{af} = \frac{s_{ij}^{af}}{\sum\limits_{m=1}^{m_i^{af}} s_{im}^{af}} \tag{7.15}$$

where:
m_i^{af} is a number of all iid_i's relationships derived from author–favorite layer.

$$C_{ij}^{fa} = \frac{s_{ij}^{fa}}{\sum\limits_{m=1}^{m_i^{fa}} s_{im}^{fa}} \tag{7.16}$$

where:
m_i^{fa} is a number of all iid_i's relationships derived from favorite–author layer.

Relationships Based on Opinions

The last three types of relations can be extracted from information about commented pictures. A relationship opinion–opinion r_{ij}^{oo} from user iid_i to iid_j exists if both users commented at least one common photo. A relation author–opinion r_{ij}^{ao} from author iid_i to commentator iid_j exists if user iid_j commented at least one iid_i's photo. A relation opinion–author r_{ij}^{oa} from commentator iid_i to author iid_j exists if user iid_i created opinions to at least one iid_j's photo.

Note that the existence of relation r_{ij}^{oa} results in the reverse relation r_{ji}^{ao}. Note that a single new opinion provided to the given MO_m may create as many new relations as many distinct users commented this MO_m.

The strength values of opinion–based relations are evaluated as follows:

$$s_{ij}^{oo} = \frac{n_{ij}^{oo}}{n_i^o} \tag{7.17}$$

$$s_{ij}^{ao} = \frac{n_{ji}^{oa}}{n_i^a} \tag{7.18}$$

$$s_{ij}^{oa} = \frac{n_{ij}^{oa}}{n_i^o} \tag{7.19}$$

where:
n_{ij}^{oo} is a number of photos commented simultaneously by user iid_i and iid_j;
n_i^o is a total number of photos commented by iid_i;
n_{ji}^{oa}, n_{ij}^{oa} is a number of photos commented by user iid_j and authored by iid_i and vice versa commented by iid_i and authored by iid_j, respectively;
n_i^a is a total number of pictures authored by iid_i and commented by others.

Similarly to favorites, relations based on opinions are kind of object–based relation with either equal (R^{oo}) or different roles (R^{ao}, R^{oa}).

In the case of opinion–based layers the commitment functions are calculated as follows:

$$C^{oo}_{ij} = \frac{s^{oo}_{ij}}{\sum\limits_{m=1}^{m^{oo}_i} s^{oo}_{im}} \tag{7.20}$$

where:
m^{oo}_i is a number of all iid_i's relationships derived from opinion–opinion layer.

$$C^{ao}_{ij} = \frac{s^{ao}_{ij}}{\sum\limits_{m=1}^{m^{ao}_i} s^{ao}_{im}} \tag{7.21}$$

where:
m^{ao}_i is a number of all iid_i's relationships derived from author–opinion layer.

$$C^{oa}_{ij} = \frac{s^{oa}_{ij}}{\sum\limits_{m=1}^{m^{oa}_i} s^{oa}_{im}} \tag{7.22}$$

where:
m^{oa}_i is a number of all iid_i's relationships derived from opinion–author layer.

Aggregation of Layers

According to Definition 2.2.2 multirelational social network SSN contains a set T of ties derived from data about direct intentional links between users or their shared activities. Ties (linkages) can be created as aggregation of all previously discovered relation layers. As a result, we obtain combined multirelational social network (Figure 7.1). Thus, a tie l_{ij} from user iid_i to user iid_j exists in the multirelational social network, if there exists at least one relation from iid_i to iid_j of any kind. As a result, set T is the sum of all relation sets identified within the system:

$$T = R^c \cup R^{rc} \cup R^{coc} \cup R^t \cup R^g \cup R^{ff} \cup R^{fa} \cup R^{af} \cup R^{oo} \cup R^{ao} \cup R^{oa} \tag{7.23}$$

However, tie $l_{ij} = (iid_i, iid_j) \in T$ reflects only the fact of connection from iid_i to iid_j. Hence, similarly to relations, we can assign real values called strength of tie s^l_{ij} to each existing tie $l_{ij} \in T$ based on strengths of all component relations:

$$s^l_{ij} = \frac{\sum_k \alpha_k \cdot s^k_{ij}}{\sum_k \alpha_k} \tag{7.24}$$

where:
k is the index of relation layer (Figure 7.1); for the Flickr system, we have k=1 for R^c, 2 — R^{rc}, 3 — R^{coc}, 4 — R^t, 5 — R^g, 6 — R^{ff}, 7 — R^{fa}, 8 —

R^{af}, 9 — R^{oo}, 10 — R^{ao}, 11 — R^{oa};
α_k is a static coefficient of the kth layer importance;
s_{ij}^{k} is a strength of the kth relation from iid_i to iid_j, e.g. $s_{ij}^{1} = s_{ij}^{c}$, $s_{ij}^{2} = s_{ij}^{rc}$, $\cdots$, $s_{ij}^{11} = s_{ij}^{oa}$.

Strength of linkage aggregates all strengths from all relation levels discovered in the system. Note that values of all strengths both for relations and for ties are from the range [0;1].

The commitment function that aggregates the commitment function values for all relationships existing between iid_i and iid_j is calculated as follows:

$$C_{ij}^{l} = \frac{s_{ij}^{l}}{\sum\limits_{m=1}^{m_i^l} s_{im}^{l}} \tag{7.25}$$

where:
m_i^l is a number of all iid_i's ties derived from all extracted layers.

Note that one can use many different formulas for the commitment function. For example, we could incorporate the time factor into simple quantities of individual activities. In this case, each historical activity would not be counted as 1 but as $\frac{1}{\lambda^{tp}}$, where λ is the constant and tp denotes the number of fixed periods, which have passed since the time of the activity.

Based on such definition of evaluating the commitment functions the whole process of calculating node position values can be performed.

Appendix C — Node Position Based on Outgoing Relations

Node Position Centrality — General Concept

The node position method can be applied not only to calculate the prestige of a node but also the its centrality. In such a case a basic concept of evaluating node position remains unchanged but the commitment function needs to be reformulated in the way presented below. This results in taking into consideration not incoming relations but outgoing ones.

The centrality of the node in the weighted and directed *NIU*, expressed by the node position centrality function, tightly depends on the strength of the relationships that a given user maintains with other members of the network as well as on the node positions centralities of these members — called acquaintances. In other words, the member's node position is inherited from others but the level of inheritance depends on the activity of this user directed to members. The activity contribution of one user absorbed by another is called commitment centrality.

Node position centrality function $NP_c(x)$ of a member x in the social network of Internet users, respects both the value of node position centralities of all other network members as well as the level of activity of x to other members:

$$NP_c(x) = (1 - \varepsilon) + \varepsilon \cdot \sum_{i=1}^{m_x} (NP_c(y_i) \cdot C(x, y_i)) \tag{7.26}$$

where:
ε – the constant coefficient from the range $(0; 1]$;
y_i — x's acquaintances, i.e. the members with whom x is in direct relationship: $C(x, y_i) > 0$;
$C_c(x, y_1)$,...,$C_c(x, y_m)$ – the commitment function that denotes the contribution in activity of x directed to y_1,$\cdots$,y_m.
m_x — the number of x's acquaintances.

In general, the greater NP_c one possesses the more valuable this member is for the entire community because he/she can effectively communicates with other users. Such people are likely to be connected via strong relations with many other members. The node position centrality of the user x is inherited from the others but the level of inheritance depends on the activity of user

x directed to her/his acquaintances, i.e. intensity of communication. Thus, NP_c depends both on the number and quality of relationships.

There are five important constraints regarding commitment centrality function derived from the relationships $C_c(x,y)$ in $NIU(IID,R)$:

1. Commitment centrality function $C_c(x,y)$ reflects the strength of the relationship from x to y in $NIU(IID,R)$, $x,y \in IID$, $x \neq y$. If there exists the relationship $(x,y) \in R$ then $C_c(x,y) > 0$. If there is no relationship from x to y, i.e. $(x,y) \notin R$ then $C_c(x,y) = 0$, except in the case of condition 5.

2. The value of commitment is within range $[0;1] : \forall(x,y \in IID)C_c(x,y) \in [0;1]$.

3. Commitment centrality function to itself equals 0: $\forall(x \in IID)C_c(x,x) = 0$.[3]

4. The sum of all commitments directed to a given network member has to equal 1:
$$\forall(x \in IID) \sum_{y \in IID} C(y,x) = 1 \tag{7.27}$$

5. If a member y is active to other users but none of them is active to y and since no isolated members are allowed in $NIU(IID,R)$, in this case, to satisfy condition 4 (Equation 7.27), the sum 1 is distributed equally among all y's acquaintances – x, i.e. all values of $C_c(x,y)$:
$$\sum_{z \in IID} C_c(z,y) = 0 \Rightarrow$$
$$\Rightarrow \forall(x \in IID : C_c(y,x) > 0) \tag{7.28}$$
$$C_c(x,y) = \frac{1}{card(\{x \in IID : C_c(y,x) > 0\})}$$

Note that the network of internet users $NIU(IID,R)$ must not contain any isolated members. This restriction is derived from the lack of possibility to satisfy all enumerated above conditions for such members, especially condition 4 (Equation 7.27).

The consequence of the 4th constraint is that if member x is active to only one other member y, then $C_(y,x) = 1$.

Node Position Centrality — Commitment Evaluation

To assess the strength of the relationship between two individuals y and

[3] In the case when the user e.g. sends emails to himself/herself then this communication is not taken into consideration and is excluded from the further analysis

x within the network of Internet users the commitment centrality function $C_c(y,x)$ is used. It denotes the amount of the member y's activity that person y passes to member x.

The commitment centrality $C_c(y,x)$ of member y directed to x is directly evaluated from source data as the normalized sum of all contacts, cooperation, and communications from y to x in relation to all activities incoming to y from external world:

$$C_c(y,x) = \begin{cases} \dfrac{A_c(y,x)}{\sum\limits_{y \in IID} A_c(y,x)}, & \text{when } \sum\limits_{y \in IID} A_c(y,x) > 0 \\ \text{apply Condition 5,} & \text{when } \sum\limits_{y \in IID} A_c(y,x) = 0 \end{cases} \tag{7.29}$$

where:
$A_c(y,x)$ — the function that denotes the activity of person y directed to member x, e.g. number of emails sent by y to x;
m — the number of people within the virtual social network.

Note that according to requirement 3 for the commitment function we need to ensure that $A(y,y) = 0$, e.g. emails sent to themselves are excluded.

As it can be easily proved Equation 7.29 fulfills also all other requirements for relationship commitment function. Note that there may exist some members y to whom nobody is acyive, for which $\sum_{x \in IID} A_c(x,y) = 0$ and in consequence $\sum_{x \in IID} C_c(x,y) = 0$. In all such cases the process described in condition 5 (Equation 7.28) needs to be performed, in order to fulfill the fourth condition (Equation 7.27).

For example in the case of email communication the commitment centrality function $C_c(y,x)$ will be calculated as the number of emails sent by user y to x divided by the number of all emails received by user x.

Analogous to commitment function presented in Chapter 4 the time factor can also be considered in the Equation 7.29. In this situation 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 in each period are considered separately for each individual:

$$C_c(y,x) = \begin{cases} \dfrac{\sum\limits_{i=0}^{k-1} (\lambda)^i \cdot A_c^i(y,x)}{\sum\limits_{y \in IID} \sum\limits_{i=0}^{k-1} (\lambda)^i \cdot A_c^i(y,x)} & \text{when } \sum\limits_{y \in IID} \sum\limits_{i=0}^{k-1} (\lambda)^i \cdot A_c^i(y,x) \geq 0 \\ \text{apply Condition 5,} & \text{when } \sum\limits_{x \in IID} \sum\limits_{i=0}^{k-1} (\lambda)^i \cdot A_c^i(y,x) = 0 \end{cases} \tag{7.30}$$

where:
i — the index of the period: for the most recent period $i = 0$, for the previous

one: $i = 1, \cdots$, for the earliest one $i = k$–1;
$A_c^i(y, x)$ — the function that denotes the activity level of person y directed to member x in the ith time period, e.g. number of emails sent by y to 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 person y 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. For example, if one year's data set is processed and a period is a month then $k = 12$. For $\lambda = 0.9$, the data from January is considered with the factor $0.9^{11} = 0.31$, for February we have $0.9^{10} = 0.35, \cdots$, for October $0.9^2 = 0.81$, for November — 0.9 and finally for December $0.9^0 = 1$. This in a sense is similar to an idea which was used in the personalized systems to weaken older activities of recent users.

One of the concepts that can be also utilized in the time analysis is as it was presented in Chapter 4 the sliding time window [40], [65].

One of the activity types is the communication via chat. In this case, $A_c^i(y, x)$ is the number of chats that are common for x and y in the particular period i; and $\sum_{y \in IID} A_c^i(y, x)$ is the number of all chats in which x took part in the ith period. If person y had many common chats with x in comparison to the number of all x's chats, then y has greater commitment within activities of x, i.e. $C_c(y, x)$ will have greater value and in consequence the node position of member y will grow.