VU UNIVERSITY AMSTERDAM

BMI PAPER

Social Networking Analytics

Abstract

In recent years, the online community has moved a step further in connecting people. Social Networking was born to enable people to share more, on social and professional level. Due to its potential, significant scientific and technological efforts are made to better understand, control and extend this phenomenon. The public accessibility of web-based social networks stimulated extensive research in this domain. Understanding how networks grow and change, and being able to predict their behavior, contributes to the evolution of other domains such as business, education, social, biology, fraud detection, criminal investigation etc. This paper surveys fundamental concepts of social networking analytics as well as a set of established models for the problem of link prediction. Two case studies are supporting the paper: the first study treats the problem of influential behavior by measurements of centrality and power; the second study compares the accuracy of three classification algorithms for a case of co-authorship link prediction.

Author: Elena Pupazan

Supervisor: Dr. Sandjai Bhulai

APRIL 2011

Preface

This paper is written as a compulsory part of the Business Mathematics & Informatics master program at the VU University Amsterdam. The purpose of the paper is to engage the student to research on a subject of his choice as extension to the knowledge acquired during the study. The addressed problem should be business related and a computer science and/or mathematical method should be used to find answers. During the project the student is supervised by a staff member who is specialized in the chose subject.

The topic of this paper is Social Networking Analytics, with focus on underlying concepts of the discipline, behavior aspects in social networks and link prediction modeling. The choice of the topic is reasoned by the personal interest in the phenomenon of Social Networking as well as in Predictive Analytics. The core message of the paper is the significant role of social networking analytics in various activity domains.

The paper is structure in seven chapters: *Chapters 1* introduces the studied topic; *Chapter 2* presents briefly the evolution of Social Networking and Social Networking Analytics; in *Chapter 3* there are introduced fundamental concepts and metrics in Social Networking Analytics; *Chapter 4* focuses on behavioral aspects of Social Networking, introducing a set of established link prediction models. The presented theoretical aspect are supported and extended by two case studies: *Chapter 5* presents a study of influence within the Bernard & Killworth fraternity, over a determined period of time. The analysis is based on measurements of centrality and power, using the UCINET 6 technology. *Chapter 6* proposes a comparison of accuracy of three learning algorithms: *Support Vector Machine, K-Nearest Neighbor* and *Naïve Bayes* for the link prediction problem in the DBLP co-authorship community. The specific of this study is the particular set of features applied in learning. *Chapter 7* presents the conclusions of the conducted research in Social Networking Analytics.

I would like to thank my supervisor, Dr. Sandjai Bhulai, for his support and guidance in scoping and writing this paper.

Elena Pupazan
Amsterdam, 2011

Table of contents

1. Introduction

Before Twitter, there was Facebook, before Facebook there was Flickr, before Flickr there was MySpace and so on. All these virtual communities brought people together from all sides of the world, encouraging social and professional interaction.

As the interest of individual in virtual social networking grows, more scientific attention is given to them. Systems are being developed for understanding how and who acts in such social networks. These are tracking every possible social networking activity: usage, topics, who interact with who, for how long, user specific interests etc. Social Networking Analytics (SNA) is the discipline incorporating such scientific interests, arose from a long standing practice called Social Network Analysis. Social scientists trained in the latter study how people and groups are connected to each other (similar to the "Six Degrees of Separation" game). After the introduction of virtual social networks, it was a natural progression to apply the learned concepts and practices in the internet world.

Due to the high popularity and flexibility of social networking sites, companies had to develop unique strategies of reaching customers through those channels. Many SNA services and applications are today cloud-based and offer organizations various ways to track and interpret customer activity on such sites.

Understanding how people interact and what they are interested in will not only help the sales sector, but also the areas of marketing, HRM, CRM and so on. Some of the benefits of SNA are the abilities to better segment customers and estimate customer life cycles. A company is able to better see the key influencers who are leading the conversations. From there, an untapped pool of potential customers can be found and reached. Most of the services available now are offered at reasonably low costs. When leveraged by the right people and in the right way, businesses have the potential to grow and expand in a way that wasn't thought possible just a few years ago.

From a CRM perspective, being able to interact with customers on a personal level, in ways that are comfortable for them, will strengthen those customer relationships and make them last. Being proactive does not go unnoticed by customers, when their problems are recognized and fixed in a timely manner, these customers are more likely to continue using the company products or services. SNA should not be used to replace a customer service or the CRM program, it should be seen and used as an extension to the overall system.

As networks continue to increase in numbers and technology becomes more advanced, even more tools for social networking analytics will come on the market, each delving deeper into the system and offering more and more insight. If used correctly, social networking analytics may be a key tool in helping an organization to find and connect to the right markets and audiences, on a personal level.

2. Evolution of Social Network Analytics

Due to the recent globalization of the commercial environment and the impact of the new technologies, the analysis of social networks represents a major interest**.** This rather new area of research grew out of social and exact sciences, computers supporting today modeling and complex mathematical calculations, previously impossible. The analysis of social networks is driven by business and social interests, combining various academic fields.

The term *social networks* was used for the first time in 1950 in *sociometrics,* the science that seeks to obtain data on social behavior and to analyze it. The latter incorporation of mathematical tools and computing triggered the evolution of Social Network Analysis and Analytics.

The mathematical basis of SNA arose out of the fields of graph theory, statistical and probability theory, game theory as well as algebraic models. In fact, it was from these theories, especially graphs, that the Internet and various virtual networking concepts were derived.

Networks are generally studied based on the participants and their actions in the network, with little or no emphasis on the relationships. Particularly, in Social Networking and SNA the *type* and the *forms* of relationships between the network members are fundamental.

Social networking data comes today in many forms: blogs (Blogger, LiveJournal), micro-blogs (Twitter, FMyLife), social networking (Facebook, LinkedIn), wiki sites (Wikipedia, Wetpaint), social bookmarking (Delicious, CiteULike), social news (Digg, Mixx), reviews (ePinions, Yelp), and multimedia sharing (Flickr, Youtube).

Online social networking represents a fundamental shift of how information is being produced, transferred and consumed. User generated content, in any data form, establishes a connection between producers and consumers of information. For consumers, the abundance of share data and opinions is a support in making more informed decisions.

SNA is applicable in various domains and fields: organizational behavior, terrorist networking, political and economic systems, inter-relationships between banks and companies, social influence, educational systems and many others. Some of the current interests and challenges in the discipline of SNA are:

- Collecting massive amounts of data and preventing information overload for the users
- Extracting and modeling temporal patterns of information growth and fade over time
- Correcting effects and biases generated by incomplete or missing data
- Handling unreliable or conflicting information
- Classification and tracking of topics
- Identification of topic relevance
- Predicting and identifying emerging or popular topics
- Detecting, quantifying and maximizing the individuals influence
- Determining implicit links between users
- Understanding of sentiment flow through networks and polarization

This paper treats fundamental concepts in social networking and addresses in particular two topics of interest in SNA: influence and link prediction.

3. Basic concepts in Social Networking

A *social network* can be defined as a finite set of *actors* and their *relationships*. This is a simple and direct concept, allowing everyone to understand the social network according to the complete data and the connectivity of a considered network. This definitions does not say much though over the types of relationships of certain groups (i. e. the number of times they take part in the same programs or activities).

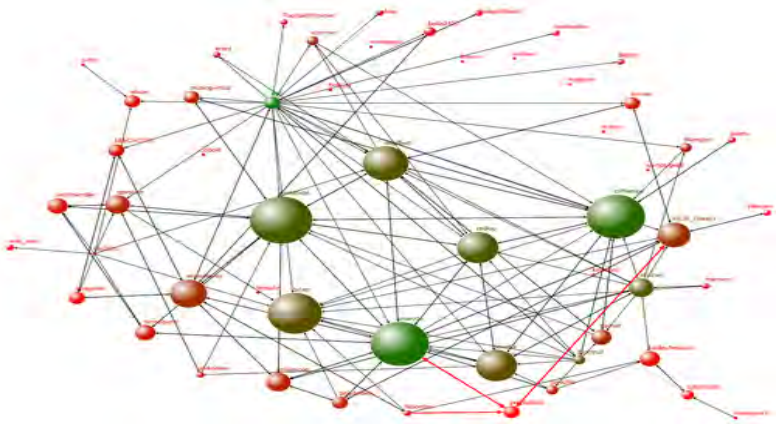

Fig.1 Social Network

An *actor* is the social entity who participates in a certain network and who is able to act and form connections with other actors. It could be an individual, a corporation or a social body. Examples of actors could be the students in a classroom, the departments in a company, the states of a federation, the web sites of a given business sector, the member nations of the UN etc. When all the actors of a network are of the same type, the network is called *monomodal*. But there are cases in which there are different actors in a network. In a multi-agent system, the actor is called an *agent*.

A link between two actors in a social network is called a *connection*. It is defined by some type of *relationship* between these actors, depending on the type of *society*. Between companies, the connection could be a business contract of supply, between people in a company, it could be the hierarchic relationship, if considering the organizational structure, or it could be the sending of e-mails in a network of relationships between friends. Other examples include the relationships of friendship or respect between students in a classroom, the biological relationships (in a family), the associations of members to clubs, the diplomatic relationships between countries etc. In the graph theory section presented later in the paper it will be shown that connections may have a value as well as a direction.

To study networks of various relationships in an objective way, models need to be created to represent them. There are three *notations* currently in use in the social network analysis:

- *Graph Theory* – the most common model for visual representation, it is graph based
- *Sociometrics* – proposes matrices representation, also called *sociomatrices*
- *Algebraic* – proposes algebraic notations for specific cases, especially for multiple relationships (Wasserman [1994])

Each notation scheme has different applications and will enable different developments and analyses. Further, this chapter presents concepts and notations used for representations with *graphs* and *sociomatrices*. The combination of these two techniques has helped significantly the evolution of social network analysis.

3.1 Graph theory

The Graph theory has been widely used in analyses of social networks due to its representational capacity and simplicity. Basically, the graph consists of **nodes** (*n*) and of **connections** (*l*) which connect the nodes. In social networks the representation by graphs is also called **sociogram**, where the nodes are the *actors* or *events* and the lines of connection establish the set of *relationships* in a two dimensional drawing.

Dyad is the simplest network, composed of only two nodes, that may be connected or not. If connected, this represents a property of the **pair**.

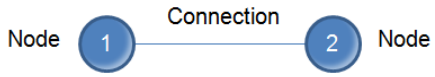

Fig. 2 Example of Dyad

Triad is a network formed by three nodes and the possible connections between them. The triad brings some important concepts into question, such as the equilibrium and the transitivity which are presented later on. There are maximum three dyads in a triad. In business relationships, this can be an important factor because if Node 1 has a relationship with Node 2, and they in turn with Node 3, there is a possible path through Node 2 and on to Node 1 to make transactions with Node 3.

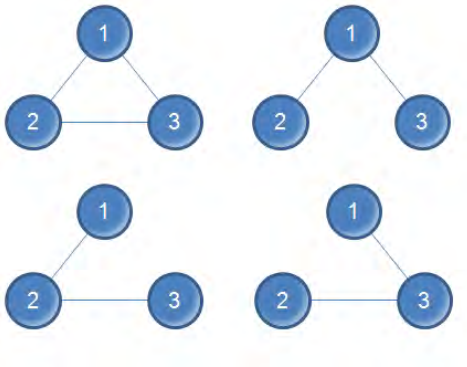

Fig. 3 Example of Triad

Group. A group can be defined as the set of all the nodes and their connections, considering a limit defined for the group. For example, the set of nations belonging to the UN and the business transactions could define a group, with the links between the countries being the connections between them. The definition of the limit is important to be able to study the group. Of course the students in one classroom have relationships of friendship with others outside of this limit, just as nations may have business relationships with countries outside the UN. But for the purposes of analysis of the social networks, the definition of the limits defines the group.

Subgroup. Within a group, there are many dyads and triads, but the concept of small sets of nodes can be extended within a group to be a *subgroup*. This can be very important in the study of complex and large social networks with the analysis of specific subgroups defined within the group.

Relationship. The set of connections of a given type defines the *relationship* found in the social network under analysis. Whereas a connection is only between two actors or nodes, the relationship is defined for the whole set of connections. Thus, we can talk about social relationships, business relationships, educational relationships etc. In the social network, there may be a connection between two actors (a situation where often the variable is set to “1” in a table or matrix), or there is none (represented with a “0”).

There are also relationships which imply values, when there is a connection and this connection can be attributed with a value (i.e. the financial worth of the business relationships between companies). The social networks where values are also involved, have a greater degree of complexity. This also due to the possibility of direction within a graph (i.e. a given company buys from another, but sells nothing to it).

Adopting next some of the nomenclatures, as in Wasserman and Faust [1994], the actors of a network will be noted *n*, and the set of actors as *N.* The connections of a network will have notation *l*, and the set of connections will be *L.* Thus, a network of *"f"* actors and of *"h"* connections will have the sets of actors and of connections defined respectively by: $N = \{n_1, n_2, \dots, n_f\}$ and $L = \{l_1, l_2, \dots, l_h\}$.

As the connection is always between two actors, then the connection defines a pair of actors (or dyad). If saying that a connection l_1 refers to the connection between actors n_2 *and* n_5, then we can write: $l_1 = < n_2, n_5 >$.

Up to this point it has been defined a connection between two actors without being concerned about the type of relationship. Many of these connections are **non-directional**, meaning that a connection between two actors is established and that the relationship is not in any specific direction. For example, marriage establishes a relationship which is non-directional as it is not possible for a member to be married to another and that the inverse is not also true. If considering that the type of connection between companies to be the existence or otherwise of a contract, such a connection is non-directional.

A **directional** connection is that which represents a connection which goes from an actor (origin) and ends at another (destination). For example, if making an analysis which considers purchases and sales between companies of a network, there will be a direction in the connections. The image below (Figure 4) exemplifies the concept. In the first case, the direction of the arrow shows that actor 1 sells to actor 2; in the second, actor 2 sells to 1, and in the last case, the graph represents that actor 1 sells to actor 2 and also that actor 2 sells to actor 1.

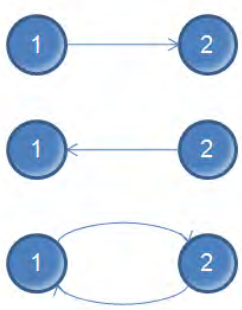

Fig. 4 Directional connection in graphs

So if a connection l_1 refers to the directional connection of actor n_2 to actor n_5: $l_1 = < n_2 \rightarrow n_5 >$.

For a network with the number of actors equal to *"f"*, the maximum number l_{max} of connections in a non-directional graph can be written using the expression:

$$l_{max} = \frac{f(f-1)}{2}$$

In other words, for two actors the maximum is one connection, for three the maximum is three, for four, it's six, and so on, as shown in Figure 5 below:

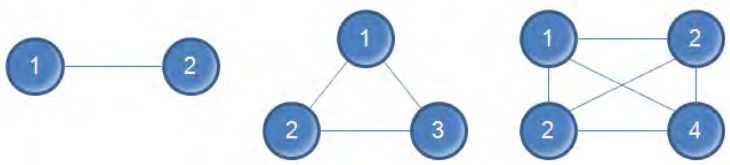

Fig. 5 Maximum number of connections in non-directional graphs

In **directional** graphs, the maximum number of connections (arrows) between two actors is two arrows (one in each direction), for three actors the maximum is six, and so on. The expression which defines the maximum number of directional connections is: $l_{maxdir} = f(f-1)$.

One example of directional graph which has the maximum number of connections is the Brazilian soccer championship. There are twenty teams playing for the championship, each team plays against all the other teams, once at home and once away (outward game and return match, two directions). The total of the connections (games) in this network (championship) will be 380.

Graphs enable many interesting analyses to be made and have visual appeal which help us to understand the structure and behavior of social networks. However, for networks with many actors and connections, this becomes impossible. Similarly, some important information, such as the frequency of occurrence and specific values, are difficult to apply in a graph.

3.2 Sociomatrices

For making possible the analysis of networks with many actors and connection, the matrices developed by sociometrics, *sociomatrices,* are being used. Thus, sociometrics and its sociomatrices complement the Graph theory, establishing a mathematical basis for analyses of social networks.

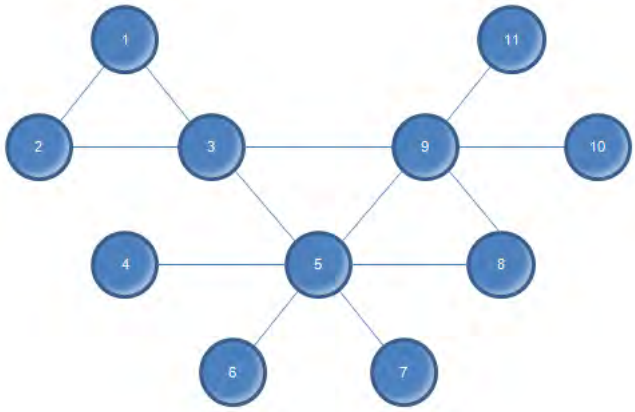

Fig. 6 Network of non-directional business relations

Figure 7 presents a **matrix** which shows the existence of the connections between the various actors of the network proposed in Figure 6, represented by a non-directional graph. In being **non-directional**, a matrix is symmetrical.

	1	2	3	4	5	6	7	8	9	10	11
1	0	1	1	0	0	0	0	0	0	0	0
2	1	0	1	0	0	0	0	0	0	0	0
3	1	1	0	0	1	0	0	0	1	0	0
4	0	0	0	0	1	0	0	0	0	0	0
5	0	0	1	1	0	1	1	1	1	0	0
6	0	0	0	0	1	0	0	0	0	0	0
7	0	0	0	0	1	0	0	0	0	0	0
8	0	0	0	0	1	0	0	0	1	0	0
9	0	0	1	0	1	0	0	1	0	1	1
10	0	0	0	0	0	0	0	0	1	0	0
11	0	0	0	0	0	0	0	0	1	0	0

Fig. 7 Symmetrical matrix for the non-directional graph in Fig. 6

Each element of the matrix shows a connection, or the lack of it, between two actors and is notated "$x_{line,\ column}$", with the sub-indices indicating the actor of a given line and the actor of a given column. If considering the values of "*i*" and "*j*" as these indices, each element will be identified by x_{ij} or algebraically:

- x_{ij} = 1 - when there is a connection between n_i and n_j
- x_{ij} = 0 - when there is no connection
- $x_{ii} = x_{jj}$ = 0 - when the connection does not exist

and in the symmetrical matrix: $x_{ij} = x_{ji}$.

Therefore, if the connections are **directional**, the graph is directional, and in this case the notation will be:

- x_{ij} = 1 - when there is a connection from n_i to n_j
- x_{ji} = 1 - when there is a connection from n_j to n_i
- x_{ij} = 0 - when there is no connection

and here the matrix is rarely symmetrical. In Figure 8 is presented a directional graph where the companies have selling relationships between each other. The arrows point in the direction of the sale.

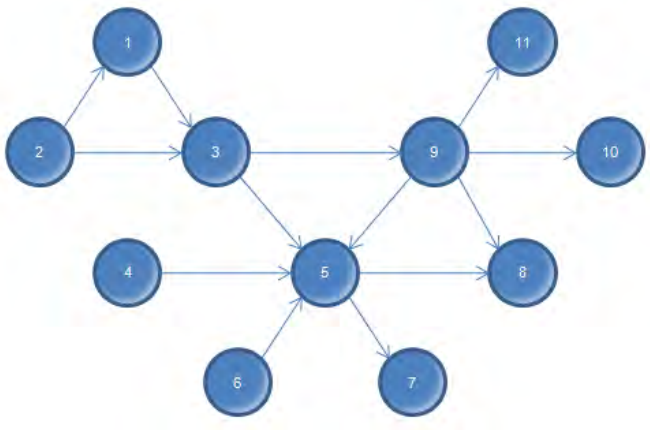

Fig. 8 Directional graph with sales connections between companies

Figure 9 presents the corresponding sociomatrix, where can be seen the asymmetry and that the main diagonal is empty.

	1	2	3	4	5	6	7	8	9	10	11
1	-	0	1	0	0	0	0	0	0	0	0
2	1	-	1	0	0	0	0	0	0	0	0
3	0	0	-	0	1	0	0	0	1	0	0
4	0	0	0	-	1	0	0	0	0	0	0
5	0	0	0	0	-	0	1	1	0	0	0
6	0	0	0	0	1	-	0	0	0	0	0
7	0	0	0	0	0	0	-	0	0	0	0
8	0	0	0	0	1	0	0	-	0	0	0
9	0	0	0	0	1	0	0	1	-	1	1
10	0	0	0	0	0	0	0	0	0	-	0
11	0	0	0	0	0	0	0	0	0	0	-

Fig. 9 Sociomatrix corresponding to the directional graph in Fig. 8

In the next section, using the basic knowledge of graphs and sociomatrices, various characteristics of the networks of business relationships, such as prestige, social role of the actors and other definitions which are useful in the practical analyses in business and social environments are being defined.

3.2 Measures in Social Networking

The use of *graphs* and *sociomatrices* is necessary in order to create **models**, or simplified representation systems of networks of relationship. However, with graphs and sociomatrices it is not possible to represent the whole of the characteristics and attributes of a network, nor all of its limits and variations. In order to make analyses therefore, the model is simplified and the analysis is based on various measures. The main measures used for social network analysis are presented in this section.

Nodal degree

In a non-directional network, it is measure the number of connections at a node and this number is called the nodal degree. The degree of a node can vary from zero, when there is no connection at this node to any other node of the network, through to the value *f – 1*, when there is a connection at this node with all the other nodes on the network. The measure of the degree of a node can define its importance, for example, in a network where there are various connections, this is something of interest to the members of the network.

To obtain a graph of the degree of a given node, $g(n_i)$, count the number of lines which are connected to this node. Considering the example shown in Figure 6 and then checking the degree of each node, in decreasing order, as follows:

- $g(n_5) = 6$
- $g(n_9) = 5$
- $g(n_3) = 4$
- $g(n_1) = g(n_2) = g(n_8) = 2$
- $g(n_4) = g(n_6) = g(n_7) = g(n_{10}) = g(n_{11}) = 1$

An important piece of data in business networks is the average number of relationships between the members of the network. This can be measured by obtaining the average degree of the network. The **average degree** is defined by the sum of all the degrees divided by the number of actors in the network or algebraically:

$$\bar{g} = \frac{\sum_{i=1}^{f} g(n_i)}{f} = \frac{2L}{f}$$

where *L* is the number of connections of the network and *f* is the total number of actors (nodes). For the network from the previous example, the value of $\bar{g}$ = 2.36.

Nodal degree (directional graph)

In directional graphs, the measure of the degree is slightly different, as it is interesting to know how many connections the origin node has and how many connections it has as destination.

The number of connections this node has as destination is called **nodal-in degree**. For the nodal-in degree of node n_i, obtained by counting the number of arrows pointing towards it. The used notation is $gi(n_i)$.

The number of connections this node has as origin is called **nodal-out degree**. For the nodal-out degree of node n_i, obtained by counting the number of arrows pointing from it. The used notation is $go(n_i)$.

These measures are very important in a network, as the **nodal-out degree** can indicate the capacity of expansion of a given actor, whilst the nodal-in degree can represent their popularity. The measure of the nodal-in degree, for example, is one of the factors which determines the status of a given web site when making a search using Google. The position in the ranking of a page shown in the search results is determined by the number of sites which link to that page on the network, in other words, the nodal-in degree of the page.

For the business network considered in Figure 8, showing the directed connections for sales from one actor to another, the next nodal-out degree and the nodal-in degree are calculated for each node:

Nodal-out degree	Nodal-in degree
$go(n_1) = 1$	$gi(n_1) = 1$
$go(n_2) = 2$	$gi(n_2) = 0$
$go(n_3) = 2$	$gi(n_3) = 2$
$go(n_4) = 1$	$gi(n_4) = 0$
$go(n_5) = 2$	$gi(n_5) = 4$
$go(n_6) = 1$	$gi(n_6) = 0$
$go(n_7) = 0$	$gi(n_7) = 1$
$go(n_8) = 1$	$gi(n_8) = 2$
$go(n_9) = 4$	$gi(n_9) = 1$
$go(n_{10}) = 0$	$gi(n_{10}) = 1$
$go(n_{11}) = 0$	$gi(n_{11}) = 1$

Table 1 Nodal-out and Nodal-in degrees corresponding to the directional graph in Fig. 8

In the table above can be seen that for the same node the nodal-out degree and the nodal-in degree may be either equal or not. Based on the differences of in and out degrees, the theoreticians of directional graphs have created different names for the roles of the nodes (Wassermann [1994]). This is of special interest in business networks, as they define the behavior of the actor in the network of relationships.

Furthermore, depending on the number and type of connection, different *types of node* are defined:

- **Isolated** if $gi(n_i) = go(n_i) = 0$ - *neither the origin nor destination of connections*
- **Transmitter** if $gi(n_i) = 0$ and $go(n_i) \geq 1$ - *not the destination of connection, but the origin*
- **Receptor** if $go(n_i) = 0$ and $gi(n_i) \geq 1$- *not the origin of connection, but the destination*
- **Carrier** if $gi(n_i) \geq 1$ and $go(n_i) \geq 1$ - *the origin and destination of connection*

For the considered example, the company node 5 is a *carrier* and acts as intermediary as a seller in this network, but also concentrates most of the buying (its nodal-in degree is by far the highest).

As for the non-directional graph, it is important to find the average nodal-in degree and the average nodal-out degree of the members of such a network. The average nodal-in degree, denoted by $\overline{ge}$, is defined as the sum of all the nodal-in degrees divided by the number of actors of the network, that is:

$$\overline{ge} = \frac{\sum_{i=1}^{f} ge(n_i)}{f}$$

where *f* is the total number of actors (nodes). Similarly, the average nodal-out degree, denoted by $\overline{gs}$, is defined as the sum of all the nodal-out degrees divided by the number of actors of the network, that is:

$$\overline{gs} = \frac{\sum_{i=1}^{f} gs(n_i)}{f}$$

The total number of “ins" have necessarily to be equal to the total of the “outs” (the sum of all the origins should be equal to the sum of all the destinations). The next formulation is possible:

$$\overline{gs} = \overline{ge} = \frac{L}{f}$$

where *L* is the number of connections of the network. For the network in the above example, the value of $\overline{ge} = \overline{gs} = 1{,}27$, which represents a directional network with low connectivity.

Density of the network. Whilst the degree of the node is important to define the number of relationships of a given actor, another important piece of data of a network is its **density**, in other words, the measurement of the number of existent connections. Dense networks are those in which there are many connections and sparse networks are those where there are few connections. Environments where there are intense business relationships, such as between the countries of the European Union form dense networks.

The measurement of the density of a non-directional network is denoted by Δ and it is defined by the number of connections *L* of this network divided by the maximum number l_{max} of connections.

The expression for the density for the non-directional graph is:

$$\Delta = \frac{L}{\frac{f(f-1)}{2}} = \frac{2L}{f(f-1)}$$

If the graph has no connections, it is said to be empty and the density is equal to 0. If it has the maximum number of connections, then it is said to be full and the density is equal to 1. Figure 10 exemplifies the empty, the full and the intermediate graph, for a network with four nodes.

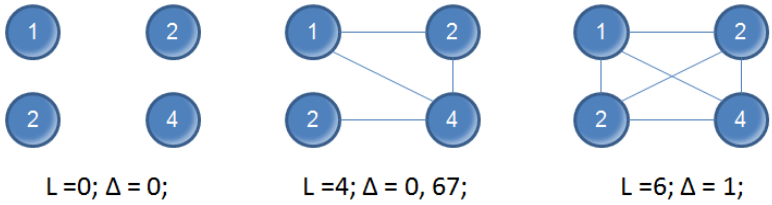

Fig. 10 Density of different non-directional graphs

For a directional network, the measurement of the density is denoted by and is defined by the number of *L* connections (arrows) of this network divided by the maximum number $l_{max.dir}$. The expression for the density for the directional graph is:

$$\Delta = \frac{L}{f(f-1)}$$

Walk, trail and path

In a network, there may be some type of relationship between two nodes, even if there is no direct connection between them, but through a third node, for example, with which both nodes have a connection. An example can be: if Mary is a friend of Jon and Jon is a friend of Joyce, it is possible that Mary and Joyce get to know each other and also become friends.

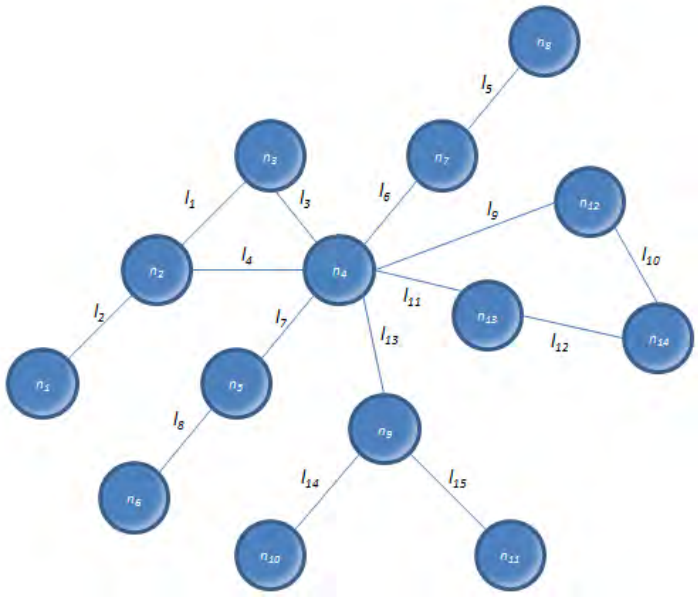

Fig. 11 Walks, Trails and Paths in a network

The various connections function as a kind of network of channels, and as the network becomes more complex, the complexity of *paths* through these various channels becomes greater. In a graph representing a network, from one actor to any other one it is possible to trace paths passing through various connections. For these paths, there are used the following definitions:

Walk – sequence of nodes and connections, starting out from one node and ending at another node, passing through the connections which join the various nodes of the route made. Nodes and connections can be repeated or not, with the length of the walk being defined by the number of connection lines travelled. In the example in Figure 11, the sequence {n_6, l_8, n_5, l_7, n_4, l_6, n_7, l_6, n_4} is one walk in which the nodes n_4 and l_6 are repeated, and the total length of the walk equal to 4.

Trail – a trail is a special type of walk in which all the connection lines are distinct, but the nodes can be repeated. In Figure 11, an example of a trail is the sequence {n_5, l_7, n_4, l_3, n_3, l_1, n_2,l_4, n_4}, in which the node n_4 is repeated. In this trail the total length is equal to 4.

Path – the path is another special case of walk in which all of the nodes and connection lines are distinct, and there can be no repetitions. One example of path in Figure 11 is the sequence {n_6, l_8, n_5, l_7, n_4, l_6, n_7}, whose length is equal to 3.

Note*:* In a network of relationships these concepts are fundamental for calculating the distances between actors, and then to set up, between companies, for example, possible negotiations based on mutual relationships.

If the graph is **directional**, as in the example in Figure 12, these paths can be interpreted slightly differently. The idea of direction has to be attributed and if designating the connections as “arrows”, the next measures can be considered:

Directed walk – sequence of nodes and arrows, leaving from a node and ending at a node, *passing along the arrows always in the same direction*, which link the various nodes of the path travelled. Nodes and arrows may be repeated or not, and the length of the directed walk is defined by the number of arrows. In the example in Figure 12, the sequence {n_7, l_6, n_4, l_4, n_2, l_1, n_3, l_3, n_4, l_4, n_2, l_2, n_1} is the directed walk in which the nodes n_4 and n_2 and the arrow l_4 are repeated, and the total length of the walk is equal to 6.

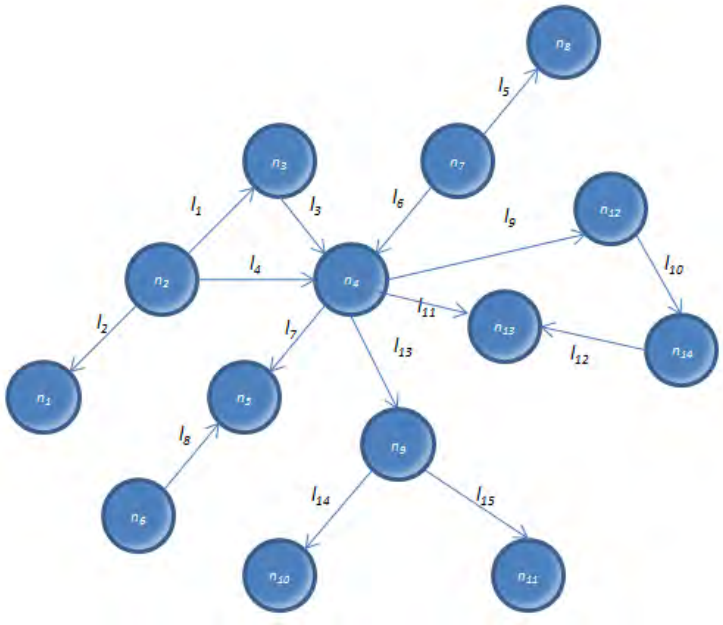

Fig. 12 Directed walks, Trails and Paths in a directional network

Directed Trail – similarly, the directed trail is a special type of walk in which all the arrows of the connection are distinct and *always in the same direction*, but the nodes may be repeated. For the proposed example, the sequence {n_7, l_6, n_4, l_4, n_2, l_1, n_3, l_3, n_4, l_{13}, n_9} is a directed trail in which the node n_4 is repeated, and the total length is equal to 5.

Directed Path – in this case, all the nodes and connection arrows are distinct, and the *arrows are always in the same direction*, without repetitions. An example of directed path in the previous figure, Fig. 12, is the sequence {n_2, l_1, n_3, l_3, n_4, l_{13}, n_9, l_{14}, n_{10} whose length is equal to 4.

Note*:* If for the previous three cases some of the arrows on the path travelled had the opposite direction, then the denominations would be ***semi**-walk, **semi**-trail and **semi**-path* respectively.

Closed walk

A sequence is called a **closed walk** when the walk begins and ends on the same node. There is no problem if some lines and nodes are repeated. An example of closed walk in the graph in Figure 11 is the sequence {n_5, l_7, n_4, l_3, n_3, l_1, n_2, l_4, n_4, l_7, n_5} in which nodes n_4 and n_5 are repeated, and the walk begins and ends at node n_5.

Cycle

A sequence is called a **cycle** when there are at least three nodes and the start-node is the same as the end-one and the connection lines are not repeated. An example of a cycle in the graph in Figure 11 is the sequence {n_4, l_3, n_3, l_1, n_2, l_4, n_4}. The concept of cycle is the same for **directional** graphs, provided that all the arrows point in the same direction on the path travelled. In Figure 12 a cycle is defined by the sequence {n_2, l_1, n_3, l_3, n_4, l_4, n_2}.

Semi-cycle

In a **directional** graph a **semi-cycle** sequence is a cycle in which at least one of the arrows points in the opposite direction to the others. An example of semi-cycle in the graph in Figure 12 is the sequence {n_4, l_9, n_{12}, l_{10}, n_{14}, l_{12}, n_{13}, l_{11}, n_4}.

Searchability and directional connectivity

In a network, if there is a path between two nodes, this means that these two nodes can establish some type of relationship along this path formed by the path, that is, a node can *find* the other node along the path. This possibility of relationship is called **searchability**.

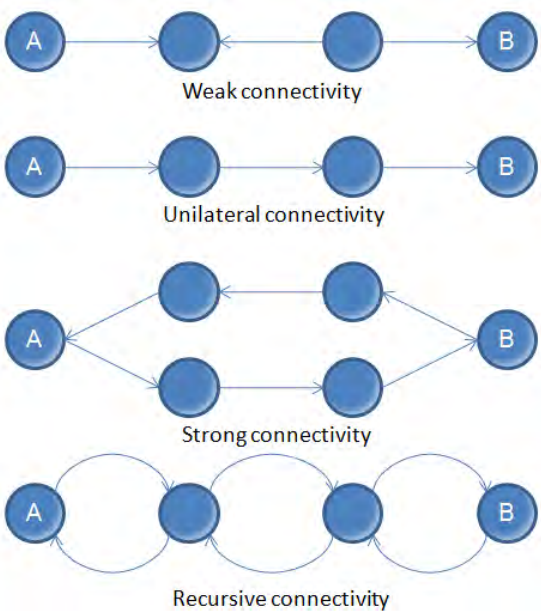

Fig. 13 Types of connectivity in directional graphs

In a **directional** graph, *searchability* can be established at different levels, depending on the direction of the arrows along the path. For a node to be able to *find* the other node in a directional network, there are four types of connectivity, as shown in the example of types of paths between nodes *A* and *B* in Figure 13. These are the four types of connectivity:

- The nodes *A* and *B* have **weak connectivity** between then when there is a semi-path between them (at least one arrow in the opposite direction)
- The nodes *A* and *B* have **unilateral connectivity** between then when there is a directional path from *A* to *B* or from B to A between them (all arrows point in the same direction)
- The nodes *A* and *B* have **strong connectivity** between then when there is a directional path from A to B and another directional path from B to A (passing through different nodes and connections)
- The nodes *A* and *B* have **recursive connectivity** between then when there is a directional path from A to B and from B to A passing through the same nodes and connections.

Every directional graph comes within one of these types of connectivity. Their interpretation is:

- The directional graph has **weak connectivity** if **all** the pairs of nodes have weak connectivity
- The directional graph has **unilateral connectivity** if **all** the pairs of nodes are connected unilaterally
- The directional graph has **strong connectivity** if **all** the pairs of nodes have strong connectivity
- The directional graph has **recursive connectivity** if **all** the pairs of nodes have recursive connectivity

Note: These ideas are important for the analysis of **cohesion** between the members of a given network. If there is weak connectivity between *A* and *B* in a business network of sales, the possibility of *A* selling to *B* is less than if the connectivity were strong.

Connected and disconnected network

A network is considered **connected** if there is a path between any pair of nodes of this network, that is, if any actor in the network can establish a relationship with any other, even if it means going through various intermediate connections and actors. If this is not possible, the network is **disconnected**.

This concept is very important because it allows one to see if a business relationship can be established using a given network, or because it enables one to see which connections could be taken out to “disconnect” the network and, for example, the connections of a terrorist network could be destroyed, if that were so desired.

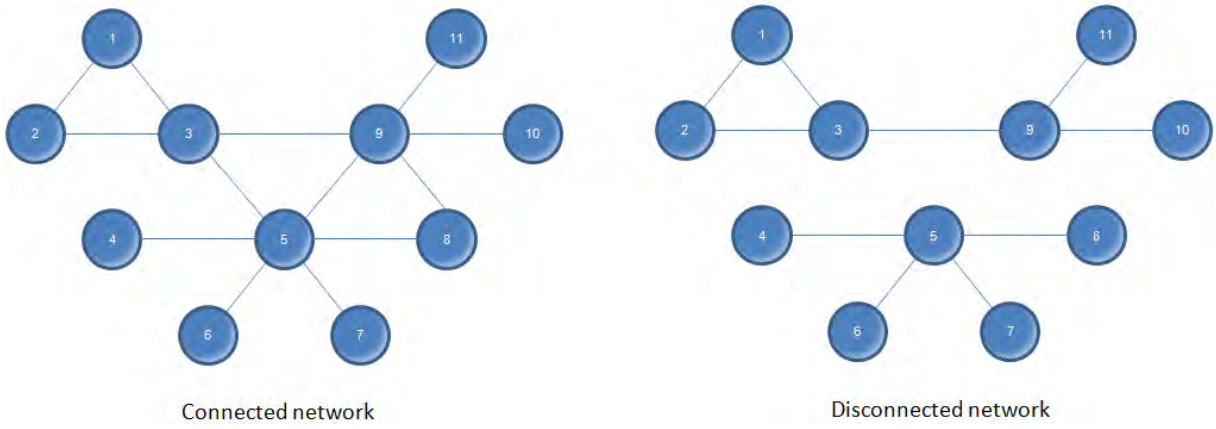

Fig. 14 Examples of connected and disconnected network

Geodesic

The shortest path between two nodes is called **geodesic**, and the length of this path, in number of intermediate connections, is called **geodesic distance**. This minimum distance is very interesting because it allows the analyst to see how many connections and how many nodes are intermediaries in a relationship between two actors of a network. The geodesic distance between any two nodes n_i and n_j, *is noted* $d(n_i, n_j)$.

If there is no geodesic for any two nodes, that is, if there is no possibility of any path between them, their distance is considered infinite and the network will be disconnected.

For a directional network, the geodesic is considered as the shortest directed path between two nodes. Considering that in a directed path all the arrows have to be in the same direction, the geodesic from n_i to n_j will not always be the same geodesic from n_j to n_i. See an example of this type in figure 2.15. The sequence which defines the geodesic from n_1 to n_3 is $\{n_1, l_2, n_2, l_3, n_4, l_4, n_3\}$, with the geodesic distance $d(n_1, n_3)=3$. Whereas for the geodesic from n_3 to n_1, the sequence is $\{n_3, l_5, n_1\}$, with the geodesic distance $d(n_3, n_1)=1$.

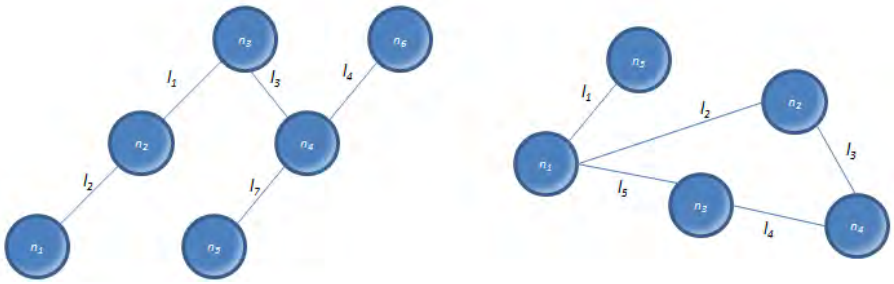

Fig. 15 Examples of geodesic in an undirected and a directed network

Diameter

Having established the geodesic distances of a connected network, the greatest distance will determine the **diameter** of this network. In the example in Figure 15, the diameter of the undirected network is equal to "4", as the greatest distance is established by the geodesics:

- $d(n_1, n_5) = 4$ and $d(n_1, n_6) = 4$

The diameter for a directional network follows the same principle, considering the greatest directional geodesic distance of any pair of nodes of the network. For the directional network in Figure 15, the diameter is equal to 4, defined by the geodesic distance from n_5 to n_3 :

- $d(n_5, n_3) = 4$

"Cut node" - Cut-point

A **cut-point** is a node which if it was taken out, would make the network disconnected, dividing it into different "components". There are very important cut-points, because they can divide the network into different and non-communicating parts, which weakens the network considerably. The taking-out of a node implies the disappearance of all its connections. As example, if taking-out node n_4 from the network in Figure 12.10, the result would be as shown below, in Figure 16.

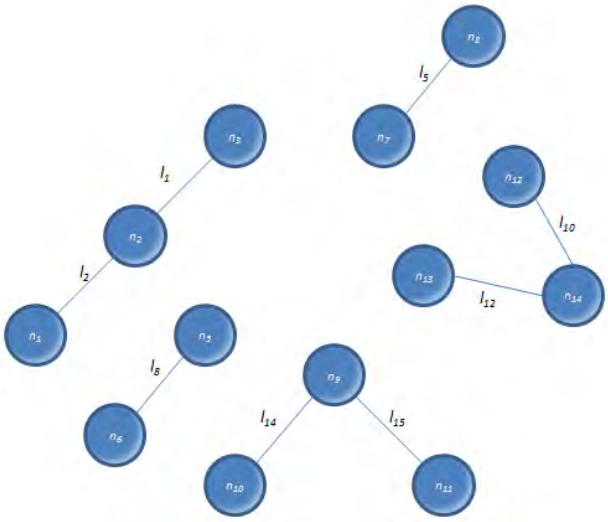

Fig. 16 Disconnected network by taking out the cut-point n_4

The network became fragmented (disconnected) and five sub-graphs or components resulted from the original graph. No other cut-point from this network can cause so much damage. Most of the nodes in this network are not cut-points (as they do not separate the network into different components). In the considered example, other cut-points are n_2, n_5, n_7 and n_9. Obviously taking- out another node affects the network quite differently than the previous one. This type of study in terrorist or organized crime networks has been done to find out which are the most important cut-points that could weaken the organization.

Bridge

The idea of **bridge** is similar to that of cut-point, but it refers to the connection which if it was taken out from the network, would make the network disconnected, dividing it into different "components". All the nodes remain in the network, and just the connection which represents the bridge is taken out, resulting in a disconnected network. In the example in Figure 17, line l_3 is the bridge. If it is taken out, the network becomes two components and nodes n_1, n_2 and n_3 are not paths to nodes n_4, n_5 and n_6. In a business environment, the connection which acts as a bridge could be a contract or an agreement. The termination of such an agreement could cause isolation, for example, of two groups in the business network who would no longer relate to each other.

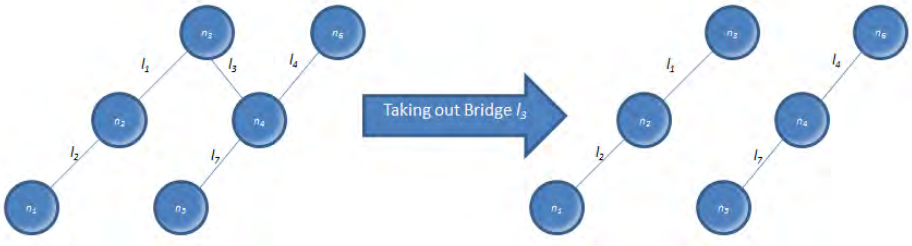

Fig. 17 Disconnected network by taking out bridge l_3

Cyclic graph and tree

Every graph which contains cycles can be called a **cyclic graph**. However, if a network represented by this graph has no cycle, it will be called a **tree**. The tree is a special network because it is weakly connected and each connection is a bridge. Any connection, if it is a taken-out, will cause a disconnection of the network. For this reason, networks in the form of a tree are not good for the business environment and any problem with an actor or a connection will affect the development capacity of the network.

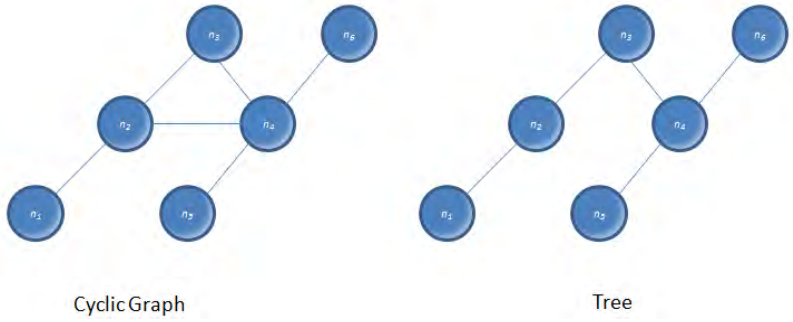

Fig. 18 Example of cyclic graph and tree type network

Bipartite graph

A graph can be considered **bipartite** if the relationships get established between two sets of actors, but with no connections between the actors within the same set. This is a special case of networks, and a practical example is in the formation of the network of distance learning relationships.

Suppose that one set of actors consisting of teacher-tutors and the students are using the tutoring tool. The teachers will be in one of the sets and the students in the other, and the connections are the various questions and answers. Not all the students establish communication with all the teachers, and not all the teachers answer all the students. If there are connections from all the actors in one set to all the actors in the other, it will be a fully bipartite graph. The concept is presented also graphically below, in Figure 19.

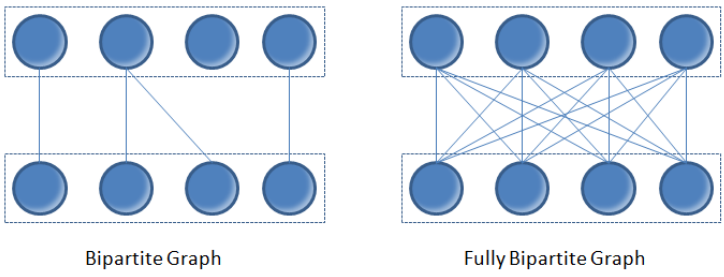

Fig. 19 Examples of bipartite graph and fully bipartite graph

Graphs with sign and with value

For each relationship established by a connection in a graph, two further pieces of additional information can be included: a **sign** and a **value**. The inclusion of a *positive or negative sign* for a connection can show us that a relationship is good or bad. An example of this type of network is a graph showing the relationships of affinity between students in a classroom. Usually (+) indicates that there is friendship and (–) indicates enmity.

The inclusion of *value* can add a number to a connection. An is indicating on the graph the business relationship between companies, the value of the connection representing the amount in millions of dollars in a sale transaction.

Centrality and prestige

Two important concepts in a network are the ideas of **centrality** and **prestige** of an actor. There are various definitions and forms of calculating centrality.

For a given actor n_i, the **centrality** is denoted as $C(n_i)$ and the measure will be given by the degree of the node, that is, by the number of connections of this node in the network. *Centrality* can be also considered the measure that gives the indication of power and influence of the individual nodes of the network based on how well they are *connected.* The fundamental measures of centrality are: *Betweenness*, *Closeness*, and *Degree*.

Betweenness measures the number of subjects whom an individual is connecting indirectly, through their direct links.

Closeness indicates how near is a subject to all other individuals in a network, directly or indirectly. Closeness centrality is the inverse measure of the sum of the shortest distances between each individual and everyone else in the network.

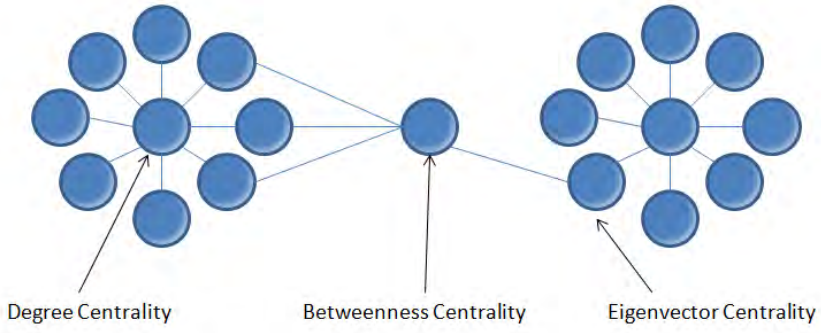

Fig. 20 Example of centrality

Centralization is the difference between the numbers of links of each node in the network divided by maximum possible sum of differences. A centralized network will have many of the links dispersed around a certain node(s) while a decentralized will have nodes with comparable number of links.

The concept of **prestige** of an individual n_i is related to the concept of directional networks. The centrality of an individual n_i , considering the arrows directed towards them (i.e. their nodal-in degree) defines his prestige, $P(n_i)$, in the network.

Other metrics in Social Networking

- **Clustering coefficient** is the measure representing the probability of a future link between two unconnected neighbors of a considered node.
- **Cohesion** represents the degree in which nodes are connected directly among each other by cohesive bonds.
- **Radiality** represents the degree with which the network of a certain individual reaches out into the global network providing content and inducing influence.
- **Reach** represents the degree in which any node of a network can reach the other nodes.
- **Structural cohesion** measures the minimum number of nodes that would disconnect the network or the group if removed.
- **Structural equivalence** represents the degree in which nodes share a common set of links connecting them to other nodes in the network

Analysis of network data can be done on different levels: *node level* (i.e. centrality, prestige, node roles such as bridges, isolates etc.); *dyadic level* (referring structural distance and reachability, notions of equivalence, reciprocity etc.); *triadic level* (referring aspects of balance, transitivity etc.); s*ubset level* (i.e. cliques, (cohesive) subgroups, network components etc.); *global network level* (referring aspects of connectedness, diameter, centralization, density etc.).

4. Behaviour and Dynamics in Social Networks

Connectedness in social networks implies two aspects: the structural connectivity (which network entity is linked to which network entity) and the behavioral connectivity (individual actions affect all other entities in the network). This is why, aside the understanding of the network structure, it is important to understand the network interaction and behavioral dynamics.

Both the structural and behavioral levels present high complexity. Considering the behavioral aspect, if the entities in the network are actively involved in the considered community, these will appreciate their influence and will consider it with their new actions. A fundamental consideration must be the fact that community behavior is continuously changing. In any social network, the behavioral shifting and evolution is caused by both internal and external factors. The network behavioral models are based on the network entities strategic reasoning and behavior, considered in social context and not in isolation.

Behavioral impact in a network should be considered both from the network and the individual point of view: as prior mentioned, an individual action has impact on the behavior or / and structure of the entire network and vice versa, the network behavior and evolution has impact on the individual behavior.

Considering the first type of dependency, the network behavior and evolution are influenced either by individual characteristics, structural positioning, connectivity activity etc. This type of dependencies is described as the *selection processes.* An exemplification of such dependency is the *homophily process*: the creation of network relationships based on entity similarity.

On the other hand, networks can affect the individual characteristics and their behavioral development. Such dependency is described as *influence processes*. An exemplification is the *assimilation process*: similarity of individuals that are highly socially connected.

Influence and *selection processes* are strongly interdependent. Separating these is difficult as the network data is inherently interdependent.

A future connection of two individuals might depend on their relationship with third-party individuals. Not many statistical models can detect such network dependencies. An established approach for the analysis of longitudinal network data is the *Stochastic actor-based modeling*, Snijders [1996]. Extensions of this approach can help the simultaneous analysis of *selection* and *influence processes,* based on the interdependence between these processes - Snijders, Van de Bunt and Steglich [2010]; Steglich, Snijders and Pearson [2010].

Large real-world networks are highly dynamic and exhibit a range of interesting properties and patterns. One of the recurring themes in the line of behavioral and dynamics research is to design models that predict and reproduce the emergence of such network structures. Research then seeks to develop models that will accurately predict the global structure of the network. The following sections introduce the concept of structural balance in Social Networks and present specific models for the Link Prediction problem.

Chapter 5 presents a behavioral case study of *structural influence* based on the centrality and power graph theory approach. The technology presented is UCINET, a dedicated tool for social networking analysis, the investigation being conducted on the UCINET dataset, *Bernard & Killworth Fraternity (BFRAT).*

Chapter 6 presents a second study, with focus on the accuracy performance of link prediction models. A set of three models are used: *Support Vector Machine (SVM), K-Nearest Neighbor* and *Naïve Bayes* and the link prediction investigation is done on the DBLP co-authorship dataset using the following set of features: *sum of papers, weighted sum of neighbors, weighted sum of secondary neighbors* and w*eighted shortest distance.*

4.1 Structural Balance of Social Networks

Social relationships have a profound impact on human development, in all life stages. Such relationships are of positive nature (i.e. friendship, collaboration, trust, support etc.) or of negative nature (i.e. oppression, dislike, harassment, intimidation etc.). A social network captures all such types of relations defined between a finite set of members. Individual characteristics and shared relationships change in time and continuously impact the entire community (social network).

Clearly, the tension executed between every two network entities, be it positive or negative is a fundamental aspect in social networking. The framework of this type of analysis is the *structural balance*, which aims to extract and store the relationship information in a clear and structured way. The structural balance concept is based on social psychology theories being helped by graphical and mathematical representations. The *structural balance theory* is based in fact on pure mathematical analysis.

The structural balance theory is based on identifying the nature of relationship between two individuals by initially isolating them. If these individuals share some level of friendship, support or collaboration, their link is marked positive: "+", else the link is marked negative: "-". The theory looks at subgroups of three individuals sharing a particular configuration of positive and negative values. In fact, there are possible four distinctive configuration cases between three individuals A, B, C. These are presented in Figure 21 below.

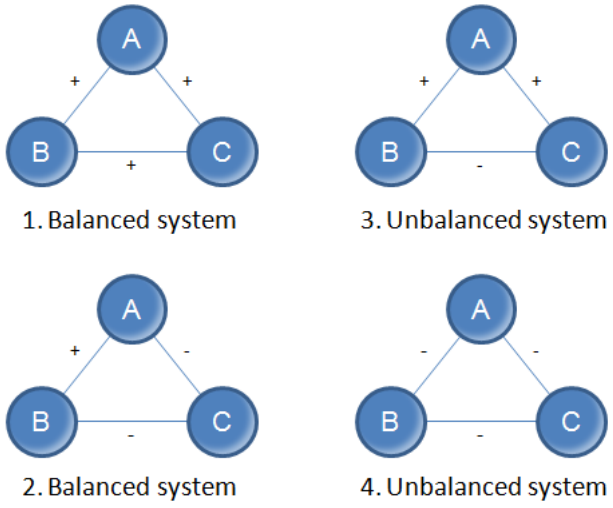

Fig. 21 Structural balance for sets of three nodes

In such reduced systems, clear conclusion of structural balance can be drawn:

Case 1: *A, B, C are mutual friends***.** This is a natural situation of three persons that are mutually friends. There are no instability sources in such system, therefore the system is *balanced.*

Case 2: *A, B are friends and C is a mutual enemy.* This is also a natural situation between three individuals, two of the three are in a relationship of friendship and both dislike the third individual. As the system has clear friendship and enemy bindings and therefore no instability sources, this system is *balanced.*

Case 3: *A is friend with B and C, but B and C are enemies.* In such system there is present, in some degree, a psychological *stress* or *instability* into the formed relationships: one individual is in a friendship relation with two other individuals that dislike each other. The instability source comes from the fact that individual A might try changing the negative relation between B and C in positive one or might take side and become enemies with one of the individuals B or C. Based on this instability reasoning, this system in *unbalanced.*

Case 4: *A, B, C are mutual enemies.* In this type of system there are also present instability aspects. The reasoning is based on the fact that two individuals might start collaborating against the third individual in the system. In this case a negative link might transform in a positive one. This is why this system too is considered *unbalanced.*

In conclusion, the *structural balance* of a sub-system of three individuals connected by three links is achieved if*: either all three links are positive or else, only exactly one of the links is positive.* This consideration is known as the *structural balanced property* and is at the basis of the global structural balance of the network.

The global structural balance of the network is expressed as the problem of eliminating the unbalanced triangles. This expression is not convenient due to the involved computation, but it represents the basic start point in the concept of *structural balance* of social networks. A more mature formulation of structural balance in a social network is the *Balance Theorem,* given by Frank Harary [1953]:

"*If a labeled complete graph is balanced, then either all pairs of nodes are friends, or else the nodes can be divided into two groups, X and Y, such that every pair of nodes in X like each other, every pair of nodes in Y like each other, and everyone in X is the enemy of everyone in Y."*

Today structural balance is highly relevant in the *on-line social media* where individual opinion is intensively expressed, often in a context of influence. Another example is the *international relations,* representing the relationship between various countries.

Understanding the mechanism of positive and negative relationships helps the studies of behavior, structure and influence in the social field. These are important aspects in managing social or business contexts. Research is only starting exploring these fundamental questions, aiming to understand how, out of large scale datasets, balance and related theories can bring out knowledge.

4.2 Link Prediction Models

Social networks present high dynamics and a continuous transformation by adding new nodes and edges. This behavior causes changes in the nature of the social interaction and the structure of the network. For various domains it would be a great benefit to be able to understand and therefore control the mechanism of evolution of social networks.

Apart from influence, another fundamental topic in the evolution of social networks is the link prediction problem. This subject has captured the attention of various scientists, especially in the artificial intelligence sector and data mining.

Many studies refer business and professional collaborations generated by informal social interactions in such networks. Other studies focus on the impact of the social hierarchy in the professional network or inferring missing links. It is interesting to notice that most of these studies conclude that effective and concrete link prediction methods can be used to analyze social networks so to predict future interactions that might help organizations, businesses or investigations.

The social network analysis proved a significant role in domains as security, terrorism, biology, sales and many others. In some of the domains, such as security and terrorism, the type of prediction is of a link between groups of individuals that collaborate, but not by an obvious connection. In domains similar to sales, a typical type of link predictions regards the potential collaboration based on observations of business and informal interests and actions.

Today, due to the large amount of available social networking data, studies and simulation of different nature are possible. These contribute significantly to understanding the properties and the behavior of social networks.

4.2.1 Mathematical framework

Consider a social network $G = (V, E)$, where V is the set of network nodes and E is the set of edges between the network nodes, the problem of link prediction is the task to predict how likely a new link $e_{ij} \notin E$ will exist between a pair of existing nodes in the network (v_i, v_j).

Often the time dimension is added to the link prediction problem so to measure the growth of the network. In this case the discussed problem should be seen as the task of accurate prediction of the edges that will be added to the network between two deterministic points in time.

The link prediction problem addresses four main aspects: *link existence, link type, link weight* and *link cardinality.* Many link prediction studies concentrate on the problem of *link existence* - whether a new link between two nodes in a given social network will exist in the future or not. The link existence problem is extended by the other two problems of link prediction: *link weight* – the links between different network nodes are given different weights and *link cardinality* – two nodes of a given social network are connected to more than one link. The fourth problem, the *link type* is a more particular problem - it refers to possible different roles of the one relationship between the same two nodes of the given social network.

The link prediction problem can be treated with techniques of various natures: statistics, probability, graph theory, machine learning etc. Depending though on the approach of analysis, the techniques can be classified in three groups:

- **Models based on node similarity** – regards the similarity measurement between two nodes
- **Models based on topological patterns** - local or global patterns that could define the network
- **Methods based on probabilistic models** – a defined model that could abstract the network

4.2.2 Models based on node similarity

The models based on node similarity propose measurements of similarity for pairs of network nodes. In this context, the task of link prediction is the consideration of new edges between network nodes presenting a considerable similarity, usually measured against a threshold. In general, the measurement of similarity is either (*pre)defined* or *learned* (using machine learning techniques), depending on the studied domain or the type of network.

According to Lin [1998] the similarity between two network nodes $(v_1; v_2)$ can be defined by the percentage of the common information in the total set of properties characterizing the two nodes. The measurement is applicable in case of a probabilistic model for the studied case:

$$sim\ (v_1; v_2) = \frac{log\ P\ (\ common(\vartheta_1; \vartheta_2)\ )}{log\ P(\ description(\vartheta_1; \vartheta_2)\ )}$$

where $\vartheta_1,\ \vartheta_2$ are the sets of properties characterizing the two nodes $v_1; v_2$.

Another similarity distance measurement was given by Bennett and Li [2004] and refers to the *Kolmogorov* complexity measurement between the set of properties of the two nodes $(v_1; v_2)$. The *Kolmogorov* complexity measurement of a binary string v is defined as the length of the shortest program for an Universal Turing Machine (UTM) to correctly reproduce the considered string, v. Consider $\mathrm{v}_i, \mathrm{v}_j$ the binary strings corresponding to the set of properties of the two nodes, for a given UTM, the *Kolmogorov* complexity measurement $K(\mathrm{v}_i|\mathrm{v}_j)$ is the length of the shortest program for the UTM to output v_i when given v_i as input. In this context, the similarity measurement is formulated as:

$$dis\ (\mathrm{v}_1; \mathrm{v}_2) = \frac{max\ \{\ K(\mathrm{v}_2|\mathrm{v}_1),\ K(\mathrm{v}_1|\mathrm{v}_2)\ \}}{max\ \{\ K(\mathrm{v}_1),\ K(\mathrm{v}_2)\ \}}$$

The disadvantage of such predefined similarity measurements is that they do not consider the network context. For this reason, the adaptive similarity functions are frequently learned using supervised learning techniques. Some of the most representative techniques are: *Binary classifiers, Kernel methods* and *Statistical Relational Learning (SRL).*

Binary classifiers are proposing training a binary classifier to determine the similarity between two network nodes, based on their content information. A mapping feature function is used to extract the content features of the two network nodes in a single vector $\hat{a}(v_1; v_2)$. Considering a simple linear regression, the objective of the function is learning a set of parameters w that can indicate best similarity. For a candidate node pair, the link prediction problem is reduced to:

$$link\ (v_1; v_2) = \begin{cases} Does\ Exist, & if\ w'\hat{a}(v_1; v_2) > 0 \\ Does\ Not\ Exist, & if\ w'\hat{a}(v_1; v_2) < 0 \end{cases}$$

Within the set of pairs not selected as candidates (negative examples), it is possible and should be considered that new links might exist. Another conclusion is that in networks with few or sparse links, the number of candidates and non-candidates pairs is considerably unbalanced.

The binary classifiers are best applicable when nodes of a certain class have many features in common, else finding pairs is very difficult and the consequence is a high recall.

Kernel matrices methods are proposing an alternative to the binary classifiers that suit also the case when the set of common features between nodes of the same class is reduced. One approach is capturing the content information of the network nodes in Cartesian products for pairs of features $< v^{\alpha}; v^{\beta} >$:

$$\hat{a}_{Cart}^{<\alpha,\beta>} = (v_1^{\alpha}v_1^{\beta}, v_1^{\alpha}v_2^{\beta}, \dots, v_1^{\alpha}v_n^{\beta}, v_2^{\alpha}v_1^{\beta}, v_2^{\alpha}v_2^{\beta}, \dots, v_n^{\alpha}v_1^{\beta}, v_n^{\alpha}v_2^{\beta}, \dots, v_n^{\alpha}v_n^{\beta})$$

The problem with this approach is that the dimension of the feature set is n^2. Clearly, the involved computation is not practical in the case of networks with a large set of node-features. Also, conducting learning in a high dimensional feature-space is challenging and may lead to over-fitting.

A better solution is the approach of *Support Vector Machine (SVM)* learning algorithms, suggesting pairing nodes as inner products $< v_1; v_2 >$ and not considering the nodes individually. In this way, by using kernel functions $K(v_1; v_2)$ for the defined inner products, the challenge of classification in higher dimensional feature-space can be solved. Oyama and Manning [2004] suggested the next kernel, for any instance of node-pairs in the original feature-space:

$$K(\widehat{v_1}, \widehat{v_2}) = K\left(({v_1}^{\alpha}, {v_1}^{\beta}), ({v_2}^{\alpha}, {v_2}^{\beta})\right) = < {v_1}^{\alpha}, {v_2}^{\alpha} >< {v_1}^{\beta}, {v_2}^{\beta} >$$

when $(\widehat{v_1}) = ({v_1}^{\alpha}, {v_1}^{\beta})$ and $(\widehat{v_2}) = ({v_2}^{\alpha}, {v_2}^{\beta})$ are instances of feature-pairs of the considered nodes. The proposed kernel is actually a tensor product between two linear kernels representing the inner products.

The link prediction problem considers the space of node-pairs as input space of nodes and the similarity between such pairs is defined by the explicit form of the proposed kernel. A high value of the kernel indicates high node-similarity. This approach has a wide applicability, especially in prediction of rating or collaborations. One specific domain of collaboration is the scientific co-authorship, and link prediction in such a community represents the subject of the second study presented in the paper.

Statistical Relational Learning (SRL) incorporates a variety of approaches and techniques. The nature of these methods can be statistical, probabilistic, logic-based algorithms etc. An established approach suggested for link prediction was established by Popescul [2003] and suggests using aggregation of relational features for measuring similarity. Various classification algorithms have been proposed and studied, many known from other disciplines such as data mining and machine learning. A particular approach is translating the link prediction problem in an optimization problem by mapping the network nodes to Euclidean spaces.

4.2.3 Models based on topological patterns

This approach is focused on identifying global or local topological patterns in the entire network or partial network. For fundamental concept in this approach is scoring the weight of the link between the nodes of a pair $(v_1; v_2)$, in rapport to the determined topological pattern(s).

Depending on the leading element in determining the topological patterns, there can be distinguished three types of topological patterns approaches: *Node based, Path based or Graph based.*

Node based approaches take into consideration the neighborhood information of a node, for example the set of first neighbors that a node has. One consideration in this area is that two network nodes would more probably establish a link if they have a large number of common neighbors.

In the proposed link prediction study in the co-authorship world, due to the nature of the domain, such information is relevant and important. Scientists and researchers tend to set new collaboration with colleagues in the same area, based on the recommendations received from their collaboration partners, the first neighbors. In other words there is a high probability that a scientist will collaborate with his second neighbors. This is one topological feature considered in the algorithm comparison.

A number of measurements of this nature have been already formulated and standardized. These intend to define a scoring function for a potential link between two nodes $(v_i; v_j)$, most often based on structural considerations such as the number of direct neighbors a node has, noted $\Gamma(v_i)$ and respectively $\Gamma(v_j)$. The most common node-based scoring functions are:

- **Common neighbors method** – proposes a scoring function of the link between two nodes $(v_1; v_2)$ based on the number of common neighbors these nodes share:

$$score\ (v_1; v_2) = |\Gamma(v_1) \cap \Gamma(v_2)|$$

- **Jaccard coefficient** – proposes a scoring function of the link between two nodes $(v_1; v_2)$ based on the ratio between their common neighbors and the total number of their neighbors:

$$score\ (v_1; v_2) = \frac{|\Gamma(v_1) \cap \Gamma(v_2)|}{|\Gamma(v_1) \cup \Gamma(v_2)|}$$

- **Adamic/Adar coefficient** – proposes a scoring function of the link between two nodes $(v_1; v_2)$ based on the number of their common neighbors, weighting more those neighbors $x \in \Gamma(v_1) \cap \Gamma(v_2)$ that the two nodes share least with other nodes in the network:

$$score\ (v_1; v_2) = \sum_{x \in \Gamma(v_1) \cap \Gamma(v_2)} \frac{1}{log\ |\Gamma(x)|}$$

- **Preferential attachment method** – proposes a scoring function of the link between two nodes based on the premise that node v_1 will receive a connection from node two v_2 with a probability proportional to the number of neighbors of v_2, $|log\ \Gamma(v_2)|$. And vice versa:

$$score\ (v_1, v_2) = |\Gamma(v_1)|\,|\Gamma(v_2)|$$

Path based approaches take in consideration the path connectivity information between two network nodes. The main idea of this type of approaches is that the more indirect paths are connecting two nodes the higher the possibility that a link will connect them directly. Many studies contributed to the theory of shortest-path distance based on analysis of the entire set of indirect links connecting two network nodes.

As in the case of the node similarity approach, a number of measures based on the path similarity have been already established. The main ones are:

- **Katz measure** - proposes a scoring function of the link between two nodes based on the sum of the total number of paths weighted according their length. If the $paths_{v_1,v_2}^{(l)}$ denotes all paths of length l between two network nodes $(v_1; v_2)$ then the formulation of the Katz measure is:

$$score\ (v_1; v_2)\ = \sum_{l=1}^{\infty} \delta^l |\ \text{paths}_{v_1;v_2}^{(l)}\ |$$

  where $\delta^l > 0$ is a parameter of the predictor.

- **Hitting time measure** – proposes a scoring function of the link between two nodes based on the required steps to reach one of the nodes, v_2 when starting from a certain node v_1 and when using a random walk to move through the neighborhoods or the considered start node. The required number of steps is also called hitting time and is often notated with $H_{v_1;v_2}$. It is important to realize that this measure is not always symmetric. This is also why often an extension of the hitting time measure is used, the commute time: $C_{v_1;v_2} = H_{v_1;v_2} + H_{v_2;v_1}$. The scoring function $score\ (v_1; v_2)$ is obtained by negating one of the two measures, hitting time or commute time.

- **PageRank measure** – proposes a scoring function of the link between two nodes $(v_1; v_2)$ that measures the probability with which node v_2 is present in a random walk that is returning to v_1. The measurement uses a parameter $\delta \in [0,1]$ considering that, at every step, the stationary probability of v_2 in the walk is δ and the probability of a move to another random neighbor is $1 - \delta$.

- **SimRank measure** – proposes a scoring function of the link between two nodes $(v_1; v_2)$ indicating weather the similarity of the considered two nodes is shared by also with other neighbors of theirs. The measure is in fact a fixed point of the previous recursive formulation defined by the condition that for a parameter $\delta \in [0,1]$ the scoring function $score\ (v_1; v_1) = 1$. In this context, the SimRank measure is formulated as:

$$score\ (v_1; v_2)\ = \delta\ \frac{\sum_{x\in\Gamma(v_1)}\sum_{y\in\Gamma(v_2)} score(x,y)}{|\Gamma(v_1)||\Gamma(v_2)|}$$

Graph based approaches consider that global structure patterns of a network can be captured by low-rank matrices and propose an approximation of the adjacency matrix *M of* the graph data *G* by a M_k product of such matrices. The link prediction problem is in this case a score function $score\ (v_1; v_2)$, where $(v_1; v_2)$ is an entry in the product matrix M_k. Using for prediction M_k and not *M* can be considered a *noise-reduction* technique that generates most of the structure in the matrix, with a more simple representation.

A known approach of approximating the adjacency matrix *M* with M_k is minimizing the *sum-squared distance*. As *M* represents the data of graph *G,* any link prediction problem could be formulated as finding a low-rank approximation $M_k$for *M.* A representative example is *Collaborative Filtering (CF),* which is the problem of user interest prediction, considering patterns in prior preference observation. The adjacency matrix $M \in \mathbb{R}^{n*m}$ captures such observations of user preference for items as ratings. Methods of matrix factorization assume that the user-item ratings in *M* are determined by a reduced number of factors corresponding to the user and the item. Consider two low-approximations matrices $O \in \mathbb{R}^{n*k}$ and $P \in \mathbb{R}^{m*k}$ to approximate *M* then $M_k = OP'$ where *k* is the rank of the resulted approximation. M_k minimizes efficiently the *sum-square distance* to the target rating matrix *M* as:

$$\int(O, P) = \sum_{i,j} (M_{i,j} - (OP')_{i,j})^2$$

The definition of the low rank matrices *O, P* is based on the leader component(s) in the adjacency rating matrix *M*. The rank of the defined approximation matrices, *k,* is in general smaller than the rank of the rating matrix.

An extended formulation of the low rank approximation was given by Srebro [2004], the *Maximum Margin Matrix Factorization* (MMMF)*,* considering the case of collaborative prediction where only some entries in *M* are based on observation and M_k, minimizing the sum-squared distance to these *M* entries, can no longer be defined in terms of a singular value decomposition. margining the norms of the low rank matrices:

$$\int(O, P) = \frac{1}{2}\left(\| O \|_{F_{ro}} + \| P \|_{F_{ro}}\right) + \lambda \sum_{i,j} (M_{i,j} - (OP')_{i,j})^2$$

where $\| . \|_{F_{ro}}$ denotes the *Frobenius* norm of the low-rank approximation matrices *O* and *P*. By considering the two low-rank matrices identical, the MMMF formulation can be generalized for a normal form of the graph.

Another method of determining global topological patterns is *Graph Factorization Clustering (GFC),* formulated by Yu et al.[2005]. The fundamental concept of this method is detecting hidden clusters based on a random walk approach that could bridge arbitrary node pairs. For a bipartite graph *G = (A, B, M)* - where *A*= $\{a_i\}_{i=1}^n$, *B*= $\{b_p\}_{p=1}^m$ are the two disjoint set of nodes corresponding to users, respectively items, *M* = $\{m_{ip}\}$ – represents the preference adjacency matrix, with $m_{ip} \geq 0$ the weight of a user-item link (a_i, b_p). The GFC method formulates the similarity between a pair of nodes (a_i, a_j) as:

$$w_{ij} = \sum_{p=1}^{m} \frac{m_{ip} m_{jp}}{\lambda_p} = (M \Lambda^{-1} M')_{ij}$$

where $\lambda_p = \sum_{i=1}^{n} m_{ip}$ is the node degree of $b_p \in B$ and $\Lambda = diag(\lambda_1, .., \lambda_m)$. Considering the node similarity formulation w_{ij} in the context of Markov random walks on graphs, w_{ij} represents then a quantity proportional to the stationary probability of direct links between the pair of nodes (a_i, a_j)

denoted $p\,(a_i, a_j)$. In the case of a bipartite graph, the nodes $\{a_i\}_{i=1}^{n}$ are not linked directly (the user set), needing to pass through nodes of *B* (the item set).

Considering $d_i = p(a_i)$ the degree of a node and denoting the probability of conditional transition with $p(a_j|a_i)$, the stationary probability can be written as:

$$w_{ij} = p(a_i, a_j) = p(a_i)\,p(a_j|a_i) = d_i \sum_p p(a_j|b_p)\,p(b_p|a_i) = \sum_p \frac{p(a_i, b_p)\,p(b_p, a_j)}{\lambda_p}$$

If the number of user nodes is greater than the number of items (*n>m)* then $p(b_p|a_i)$ represents the likelihood that the data point *i* belongs to item node *p.*

Yu et al.[2005] proposed also an approximation of a general data graph *G = (V,E)* by a bipartite graph *G= (A, B, M)* so to define a soft clustering structure where the node set *V = A* and the adjacency matrix *E* are observed, *B* represents the set of hidden clusters and *M* is the node-cluster association to be solved. Denoting $H = M\Lambda^{-1}$ where $\Lambda \in \mathbb{D}_{+}^{m*m}$, $\mathbb{D}_{+}^{m*m}$ representing the set of *m x m* diagonal matrices with positive diagonal values, the graph approximation can be found minimizing the divergence distance l between the matrices: $\min_{H,\Lambda}\ l(E, H\Lambda H')$ s.t. $\sum_{i=1}^{n} h_{ip} = 1$.

The diverge distance between two matrices *X* and *Y* was defined as:

$$l\,(X, Y) = \sum_{ij} (x_{ij} \log \frac{x_{ij}}{y_{ij}} - x_{ij} + y_{ij})$$

4.2.4 Models based on a probabilistic model

The fundamental concept of this approach is to learn a model based on a given network based on certain strategies of optimization such as *Maximum Likelihood (ML) and Maximum a Posteriori (MAP)*. Consider a network graph *G = (V,E)* and ∂ the set of parameters of the learned model, the candidate future links $l_{i,j}$ are defined as variables in probabilistic models and can be defined as: $P(l_{i,j}|\partial)$.

In this approach, three significant sub-categories of models are: *Probabilistic Relational Models (PRM), Bayesian Relational Models (BRM)* and *Stochastic Relational Models (SRM).* The first two categories are based on specific database structure representations: the PRMs are corresponding to the *Relational Model* and the *BRMs,* defined on the *Directed Acyclic Probabilistic Entity Relationship (DAPER)* framework, are corresponding to the *Entity-Relationship Model.*

Probabilistic Relational Models, different than the classical graph models, propose a set of three graphical models for representing the network relational data: *data graph* ($G_D = (V_D, E_D)$), *model graph* ($G_M = (V_M, E_M)$) and *inference graph* ($G_I = (V_I, E_I)$). The original application of these models was in the problem of attribute prediction for relational data. The probabilistic relational models reduce the link prediction problem to the task of prediction of the existence of attributes for potential new network links. Therefore, with the PRM framework, the link prediction problem requires setting up an <*exist*> attribute.

The *data graph* $(G_D = (V_D, E_D))$ contains the network information as the set of nodes, $v_i \in V_D$, and the set of links defined between these nodes, $e_i \in E_D$. Each node and link have associated a type $t_i \in T$: $\mathcal{T}(v_i) = t_{v_i}$ and $\mathcal{T}(e_j) = t_{e_j}$ and implicitly by a set of attributes corresponding to this type, $Z^{t_i} = (Z_1^{t_i}, \dots, Z_{m_{t_i}}^{t_i})$. As the PRMs consider a joint probability distribution over the network data information (attributes), in the given context, this can be formulated then as:

$$z = \left\{ z_{v_i}^{t_{v_i}} : v_i \in V_D, \mathcal{T}(v_i) = t_{v_i} \right\} \bigcup \left\{ z_{e_j}^{t_{e_j}} : e_j \in E_D, \mathcal{T}(e_j) = t_{e_j} \right\}$$

The *model graph* $(G_M = (V_M, E_M))$ has the purpose to present the dependencies between the type attributes Z characterizing the set of network nodes V_D. There can be probabilistic dependencies between attributes of the same type or different types. The model graph ties together the network entities with the same type as well as the attributes of these entities. In this way, a decomposition of the data graph per type can be achieved, this leading to a joint model of type attributes dependencies. Aside the structure of dependencies between the defined type attributes, a second component of the model graph is the *Conditional Probability Distributions (CPD)* associated with the network nodes *.*

The *inference graph* $(G_I = (V_I, E_I))$ is generated based on the prior two models G_D and G_M through a process similar to the one used by the *Hidden Markov Models (HMM)* to instantiate sequence models. In this process, the structure of G_I is defined based on the G_D and G_M, with the particularity that for each node-attribute pair in G_D a local copy of the correspondent CPD from G_M is made in G_I.

The PRMs differ among them mainly in the definition of the model graph G_M, the learning models and inference procedures. A number of *Probabilistic Relational Models* are next introduced:

- **Relational Bayesian Networks (RBN)** use the object oriented approach for extending the Bayesian networks concept. The model graph ($G_M^D = (V_M, E_M^D)$) in this case is a *Directed Acyclic Graph (DAG)* representing the joint distribution over the network entity type attributes by a set of CPDs. A CPD corresponding to an attribute *Z* is specified by the likelihood $P(Z|pa(Z))$, where $pa(Z)$ represents the value of the parents of *Z.* In general though, a network object is characterized by a set of attributes $(Z_1, Z_2, \dots, Z_n)$, the DAG and CPT

specifying the Bayesian network and representing the distribution for the *n*- dimensional random attributes as*:*

$$P(Z_1, Z_2, \dots, Z_n) = \prod_{i=1}^{n} P\,(Z_i | pa(Z_i))$$

Corresponding to the dependencies in the DAG structure, the joint probabilistic distribution can be expressed as a factorization of the following form:

$$p(z) = \prod_{t \in T} \prod_{z_i^t \in Z^t} \prod_{v:T(v)=t} p(z_{v_i}^t | pa_{z_{v_i}^t}) \prod_{e:T(e)=t} p(z_{e_i}^t | pa_{z_{e_i}^t})$$

where $v_i \in V_D$are the network nodes, $e_i \in E_D$ are the set of links defined between these nodes, $t_i \in T$ are the set of types associated to the network nodes and links: $\mathcal{T}(v_i) = t_{v_i}$and $\mathcal{T}(e_j) = t_{e_j}$. Each $t_i \in T$ is defined by a set of attributes $Z^{t_i} = (Z_1^{t_i}, \dots, Z_{m_{t_i}}^{t_i})$.

The structure learning problem in a Bayesian network is similar to searching the optimum in the space of all DAGs. RBNs use closed-form parameter estimation techniques, helping the structure learning. The learning methods for RBN are similar to the ones used for Bayesian networks, the efficiency of such parameter learning techniques representing the strength of this approach.

For reasons of simplicity, accuracy and efficiency, the *Relational Bayesian Networks* propose a *belief propagation* inference.

Relational Markov Networks (RMN) extend the concepts of conditional Markov Networks for relational data. The model graph in this case is an *undirected graph* ($G_M^U = (V_M, E_M^U)$) and represents the joint distribution over the attribute *z* as a set of potential functions $\boldsymbol{\phi} = \{\phi_C \mid C \in \boldsymbol{C}\}$, where $C_i \in \boldsymbol{C}$ is a set of templates of relational cliques specified by a RMN model for defining all cliques. For a graph *G*, a *clique* is a set of nodes $\mathbf{V}_c$ in *G*, not necessarily maximal (can be also one single node), such that each $V_i, V_j \in \mathbf{V}_c$ is connected by an edge in *G*. The combined probabilistic model for a set of variables *Z* is:

$$p(Z) = \frac{1}{N} \prod_{C_i \in \boldsymbol{C}} \prod_{c_j \in C_j} \phi_{C_i}(z_{c_j})$$

where *N* is a normalization constant and C_j represents all instantiations of the set of clique templates, *C.*

The RMN models extend the learning techniques of the Markov networks with an approach of parameter estimation "*maximum-a-posteriori"* using Gaussian priors. The approach considers predefined clique templates, reducing the prediction problem to optimizing the potential functions $\boldsymbol{\phi} = \{\phi_C \mid C \in \boldsymbol{C}\}$. With RMN models, the learning efficiency is not as high as in the case of RBN (the structure is not defined nor improved by learning) but this category of models presents flexible and detailed representations.

Similar to RBN models, in this case too, a *belief propagation* approach is used as inference procedure.

- **Relational Dependency Networks (RDN)** propose an extension of the dependency networks for relational data. The model graph in this case is a *bi-directed graph* ($G_M^B = (V_M, E_M^B)$), presenting a set of CPDs. RDN models try to maximize the pseudo-likelihood for each variable *z* independently. For a considered graph data G_D, the pseudo-likelihood *PL* is formulated as the product over network item types $t \epsilon T$, the set of type attributes Z^t and the nodes v_i and the links e_i of the considered type:

$$PL(G_D; \partial) = \prod_{t \epsilon T} \prod_{z_i^t \epsilon Z^t} \prod_{v:T(v)=t} p\left(z_{v_i}^t \middle| pa_{z_{v_i}^t}; \partial\right) \prod_{e:T(e)=t} p\left(z_{e_i}^t \middle| pa_{z_{e_i}^t}; \partial\right)$$

where ∂ is global the set of parameters of the learned model and pa_z represents the value of the parents of *z*.

In this approach, there are used specific queries to define the relational neighborhoods. The learning algorithm used by the RDN models takes in consideration these queries, on one hand for structuring the learning and on the other hand for the parameter estimation. Different than RBN and RMN models, the CDPs of RDN models do not need factoring over the data model, being considered that for an attribute $z_{v_i}^t$ the parent values are conditioned $pa_{z_{v_i}^t}$, independent of the fact that the parent values might have been conditioned by the considered attribute in their CPD estimation. The downside of the approach of independent CPD learning is that it does not lead with certainty to a consistent joint distribution.

In what concerns the inference approach, the RDN models propose the *Gibbs sampling* technique.

Bayesian Relational Models are based on the *Directed Acyclic Probabilistic Entity Relationship (DAPER)* framework, a probabilistic framework defined for the Entity-Relationship database model (Heckerman et. al [2004]). The framework proposes the modeling of data in specific classes: *entities, relationships, arcs, attributes, constraints and local distribution*. Classes are connected by dashed lines. For link prediction, the entities and relationship classes are given equal importance. In real-world it is often encountered that in the defined relationships, one part is defined with certainty and the other part presents uncertainty. In these cases, uncertainty referencing is used.

A Bayesian approach is applicable to relational modeling as it proposes a clear representation of parameters and hyper parameters, not at global level, but at network component level (nodes and relationships). This approach supports the *Hierarchical Bayesian Framework (HB),* structure that centers the parameterization of the prior distribution on the consideration that the prior distribution should represent both the *prior belief* and *learned prior.* The *DAPER framework* is most often considered in the context of a *Hierarchical Bayesian Framework,* in either a parametric or a non-parametric form.

The parametric form, *Parametric Hierarchical Bayesian Relational Mode,* is applicable in cases when the individual parameterization of network entities can be assumed to derive from a common prior distribution which can be learned and shared globally by the network entities.

Often the parameterization of *prior belief* and *learned prior* are different distribution types and therefore a non-parametric prior distribution presents more flexibility. This model is knows as the *Non Parametric Hierarchical Bayesian Relational Model* and is based on specifying the prior distribution as a sample from a *Dirichlet Process (DP)*, seen as a generalization of the *Dirichlet* distribution, infinitely-dimensioned.

Xu [2005] formulated the *Dirichlet Enhanced Relational Learning Model (DERL)* as: $G_{pc} \sim DP(G_0, \alpha_0)$, a sample from a DP where the base distribution G_0 presents *uncertain prior belief* and $\alpha_0 \geq 0$ represents the parameter reflecting the prior belief certainty. The flexibility of this approach lies in the fact that a multinomial parameter $\partial_{.|pc,pa}$ can be expressed as samples from the G_{pc} prior, when this is rich: $\partial_{.|pc,pa} \sim G_{pc}$.

A relational learning model is expected to predict new entities and relationship attributes based on the already defined relationship attributes. In non-parametric models, learning is based on a *sampling approach* such as *Gibbs, Polya urn*, *Chinese restaurant* etc.

A generalization of the nonparametric DERL model is the *Infinite Hidden Relational Model (IHRM),* introduced as well by XU et al. [2005], which combines the *Hidden Relational Model* with a *DP Mixture Model*. The *DP Mixture Model* aims to determine in an organized manner the appropriate number of latent states by embedding an infinite number of DP mixture models, which based on the considered data, are limited automatically to a finite number of mixture components.

A challenge in the relational learning the large number of features that might characterize an attribute. A solution is capturing information in latent variables so that information can be distributed at global level in the network and the need of extensive structural learning is reduced. From this perspective, the *Hidden Relational Model* can be considered as a generalization of *Hidden Markov Models (HMM)* using *hidden Markov random fields*.

A second particular nonparametric model is the *Infinite Relational Model (IRM)* proposed by Kemp et al. [2006], very much alike *IHRM*, though independently formulated. The main difference between the two models is that *IHRM* is able to define a CPD for an attribute based on structural consideration (considering its structural parents) and *IRM*, by modeling attributes as unary predicates, represents the CPD in a logical binary from.

Stochastic Relational Models propose a *Gaussian Process (GP)* framework based on the consideration that, for prediction tasks, the training models using a discriminative approach perform better that the generative models. The pioneers of this framework are Yu and Chu. The principal of *Stochastic Relational Models* is defining a GP for each entity type and then using a tensor composed by the set of such defined GPs for modeling the stochastic network link structure. The approach considers that the candidate links are local derivates of a latent relational function: $\tau : U\,X\,V \rightarrow E$. A candidate link $l_{i,j}$ is dependent on its correspondent latent value $\tau_{i,j}$ and is modeled by the probability $p(l_{i,j}|\tau_{i,j})$. The candidate links introduce a set of *Stochastic Relational Processes* (SRP) defined on $U\,X\,V$, generating the function τ via the tensor interaction of two GP kernel functions, one defined on *U* and one defined on *V (U, V* could have infinite number of network entities). The SRPs are described by a set of two hyper parameters $\sigma = \{\sigma_{\varepsilon,}\sigma_{\vartheta}\}$, corresponding to the GP kernel functions on *U,* respectively *V*.

In this context, the *Stochastic Relational Models (SRM)* define a *Bayesian-prior* for latent variables τ, denoted $p(\tau|\sigma)$. For a set of candidate links *C*, the marginal probability is then formulated as:

$$p(L_I|\sigma) = \int \prod_{(i,j)\epsilon\, C} p(l_{i,j}\,|\tau_{i,j})\, p(\tau|\sigma)\, d\tau, \qquad \sigma = \{\sigma_{\varepsilon,}\sigma_{\vartheta}\} \text{ and } L_I = \{l\}_{(i,j)\epsilon C}$$

By estimating the hyper parameters $\sigma = \{\sigma_{\varepsilon,}\sigma_{\vartheta}\}$ with the maximum marginal probability, the link prediction problem is realized by marginalization: $p(\tau|E_I,\sigma)$. This type of prediction is similar to general GP regressions, with the difference that the GP approach makes use of a set of hyper parameters. With the same constraint, the GP approach can be compared to a classification task.

A challenge of the approach is the scaling of GP inferences. Such attempts, due to the cubic complexity of GP inference, present computational risks even for networks of reduced size. If considering a network graph *G = (V, E)*, where *V* represents the set of network nodes and *E* represents the set of links between these nodes, the size of observations of missing links scales in $\theta(VE)$. GP inference has the computational complexity cubic to the missing data size, $\theta(V^3E^3)$, an extremely complex computation.

A solution for this problem is given by the *Stochastic Relational Process (SRP).* This approach starts from the probabilistic model which considers that the link candidate solution is generated by the latent function $\tau : U\,X\,V \rightarrow E$ following the GP process $GP(u,K)$ where *u* is the mean function and *K* is the kernel function between network links. Considering two network links: $(v_{i,}v_j)$ and $(v'_{i,}v'_j)$, the *K* covariance function can be expressed depending on the other two kernel functions, ε , ϑ, defined on *U* and *V:* $K\big((v_{i,}v_j),(v'_{i,}v'_j)\big) = \varepsilon\,(v_{i,}v'_i)\,\vartheta\,(v_{j,}v'_j)$.

The link structure dependency can be expressed by node dependency. In this way, based on a similarity notion ensured by the kernel function, if considering two pair of similar nodes: v_i with v'_i and v_j with v'_j, then also $\tau(i,j)$ is similar with $\tau(i',j')$. The edge descriptive function τ can be defined thus by a factorization of two node descriptive functions which are samples of the priors: $GP(0,\varepsilon)$ and $GP(0,\vartheta)$. In this way the computational complexity of the GP is of range $\theta(V^3+E^3)$, a significant complexity reduction.

A second approach of improving the GP scaling complexity is based on a link descriptive covariance:

$$K\big((v_{i,}v_j),(v'_{i,}v'_j)\big) = \frac{1}{\sqrt{2}}\big(C(v_{i,}v'_i)C(v_{j,}v'_j) + C(v_iv'_j)\,C(v_{j,}v'_i)\big),\ \ where\ C(v_{i,}v_j) = < z_i, z_j >$$

This approach is very similar to the previous one presented, based on a node descriptive covariance. With this approach the computational complexity of the GP is of range $\theta(\rho^3+\rho^2|\mathbb{O}|)$, where $|\mathbb{O}|$ represents the input network links and ρ represents a small value.

4.2.5 Conclusion

The problem of *Link Prediction* can be approached in from various perspectives. This paper reviews the most significant link prediction models according to three perspectives: *node similarity, topological patterns* and *probabilistic models.*

Prediction models based on node similarity propose measurements or learning techniques of the similarity of two network entities. The prediction task is in this case based on the similarity distance.

Prediction models based on topological patterns focus on local or global patterns in the network data. In this approach, the prediction task depends on the determined or learned patterns.

Eventually, the prediction models extending the probabilistic model intend to capture the network data in a compact object oriented structure. There are two categories of models within this approach: graphical models such as the frameworks: *Probabilistic Relational Models (PRM)* and *Directed Acyclic Probabilistic Entity Relationship (DAPER)* or discriminative models such as: *Stochastic Relational Models (SRM).* In this case, the prediction task is based on the probabilistic distribution (i.e. joint, CPD, posterior) determined with the help of a learning model. These models are robust and accurate, but present the disadvantage of extensive computation. Recently the most attention is given to relational modeling. The main reason is that this approach provides a mean to capture any network informational data. More, node and link regeneration is possible using a learned model.

An overview of the advantages and disadvantages of the presented models is given in Table 2, below.

Approach		**Advantage**	**Disadvantage**
Node Similarity	Predefined	Simple	Exclude of network context
	Supervised learning	Supervised learning	Classification imbalance
Topological Patterns	Node based	Simple	Mainly local considerations
	Path based	More detailed	Mainly neighbourhood considerations
	Graph based	More detailed	Mainly global considerations
Probabilistic Model	PRM	Factorization	Parameterization
	Bayesian DAPER	Nonparametric	Computational complexity
	SRM	Nonparametric	Computational complexity

Table 2 Comparison of the LP problem approaches

Within the presented approaches, a variety of techniques are applicable for the link prediction problem. The nature of these techniques are disciplines such as machine learning, stochastic optimization, graph theory, probabilistic models etc. Often mixed models are formed; combining these techniques as well as, due to domain correlation, a technique might apply within different approaches.

Unfortunately the study of social networks and their evolutions is community dependent and is done in isolation. Therefore a major interest in this field is the study of link prediction in a dynamic social network as well as mechanism of knowledge transfer between different social networks.

Considering the various models presented, the choice of models supporting the link prediction problem is wide. A current research interest is to identify the best fitting models and their correspondent tuning for certain types of analysis. For this purpose, the case study presented in Chapter 6, proposes an accuracy comparison of three algorithms: *Support Vector Machine (SVM), K-Nearest Neighbor* and *Naïve Bayes,* for the co-authorship link prediction problem, based on the DBLP dataset. The study is specific for the particular set of features applied to the three algorithms: *sum of papers, weighted sum of neighbors, weighted sum of secondary neighbors* and w*eighted shortest distance.*

5. CASE STUDY 1: Influence in Virtual Communities

Due to impressive technological progress, virtual communities involve today millions of individuals. Most of these individuals integrating successfully such activity into their daily practices. Aside their significant role in exceeding geographical boundaries in the individual interaction, virtual communities are the most valuable source of data on human behavior. The accessibility of such data supports the technological and scientific progress. Knowledge is gathered by explorations of individual profiling, relationship nature, patterns of interaction, arising interests, trends, influence, patterns of technology usage, etc.

In the competitive business and social contexts we currently take part of, influence and leadership play an imperative role. Leadership is no longer resumed to the traditional face-to-face interaction, this concept includes nowadays elements of influencing others at a distance, on various virtual platforms. Both traditional and modern interaction share though the same concept: a leader is an active agent interpreting content for the lower-end users and influence is accomplished via special techniques, knowledge, personality or/and other uniqueness. Such agents are seen trustworthy and non-purposive on both platforms and today they have considerable more influence than media.

For any business field, it is therefore critical to grow and identify such players. There are three aspects that need to be considered when identifying an influential actor in a social network: *who one is* (identification); *what one knows* (his competence); *who one knows (*his positioning in the global network). Logistically it is not hard to advance some standards for such leaders, the challenge lies though in the fact that power and influence vary continuously within the members of a community.

In very large networks, due to such high dynamics, it is impossible to monitor, track and measure influence players and patterns at all times. Today there is a fair large choice of software tools for social networking analysis and analytics. Most of these software packages facilitate quantitative or qualitative analysis of any type of network, through numerical and/or visual representation. Network features defined at various levels (i.e. node, dyads, triads, node links, groups or globally) are supporting such analysis.

This paper presents an influence study based on the UCINET dataset *Bernard & Killworth Fraternity*, according to the fundamental concepts of centrality and power in social networking: *degree centrality, betweenness centrality* and *closeness centrality.* The technology used in this study is the UCINET software network analysis software, a program developed by Steve Borgatti, Martin Everett and Lin Freeman. The study proposes the exercise and comparison of six measures: *Freeman Degree Centrality, Bonacich Degree Centrality, Freeman Betweenness Centrality, Flow Betweenness Centrality, Path Distance Closeness Centrality* and *Eigenvector of Geodesic Distances Closeness Centrality.*

The UCINET software package works in tandem with the freeware program NETDRAW for graphical network representations. NETDRAW is installed automatically with UCINET.

5.1 Data

The proposed influence study is based on the UCINET dataset - *BFRAT (Bernard & Killworth Fraternity).* The data describes interactions among the students of a fraternity at a college in West Virginia, over a period of five days. The dataset gathers both observed behavior information (BKFRAC matrix) and cognitive (recall) information (BKFRAC matrix).

Dataset	BFRAT	58 nodes
Two 58 X 58 matrices	BKFRAB symmetric, valued	1934 ties
	BKFRAC non-symmetric, valued (rankings)	3306 ties

Table 3 Description of the BFRAT dataset

BKFRAB matrix presents the number of times an external observer notice a pair of students in a conversation. For collecting the data, the observer agent walked through the main public areas of the community every quarter of an hour, twenty one hours a day. As the observation was made by an external agent, the information is symmetrical: subject 1 interacting with subject 2 implies subject 2 interacting with subject 1. The link of two nodes represents at least one interaction between two subjects. The network links are weighted with the number of times the two subjects connected have been noticed interacting.

Figure 24 below presents an overview of the gathered external observations:

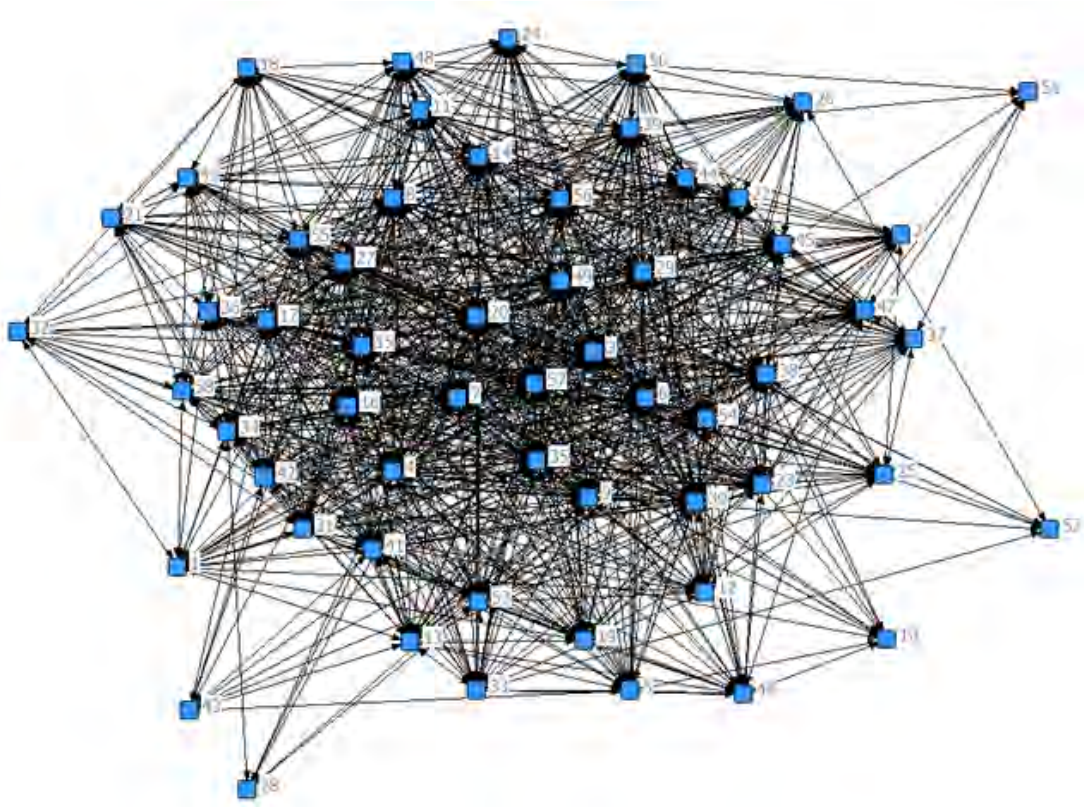

Fig. 24 BKFRAB Network Graph

BKFRAC matrix presents the post-statements of the subjects of their interaction with other subjects during the observed five days. As this information is based on recall, the resulted matrix is non-symmetric: subject 1 stating an interaction with subject 2 does not imply that subject 2 remembers interacting with subject 1. The individual interactions have a range from 0 to 5, 0 representing no interaction during the surveyed period. These individual interactions represent the ranking of the node.

The surveyed members could re-call more interactions than the interactions collected by the external observer: the BKFRAC graph has 3306 ties versus 1934 ties in the BKFRAB graph.

The overview of the recalled interactions by the subjects is presented as a network graph in Figure 25.

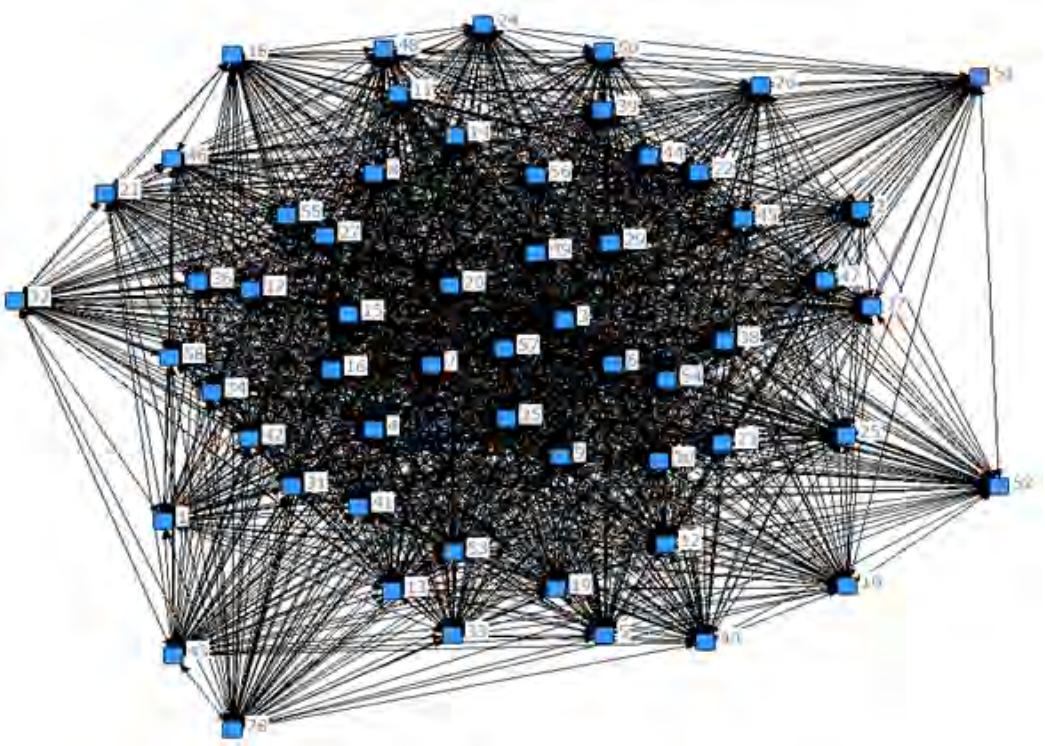

Fig. 25 BKFRAC Network Graph

5.2 Approach

The most comprehensive method of evaluating influence in a graph is the network theory approach. This is used extensively in consumer analysis in closed networks (i.e. telecom), organizational consulting, terrorism analysis etc. The method incorporates three main objective measures of influence:

- *Degree Centrality* – measures the number of adjacent links to a node of the network. Though simple, this measure is most often very effective in the measurement of centrality and power of a subject in the network. The underlying consideration is that the more ties a subject has, the better connected and therefore more influential he is. Subjects with a strong connectivity are also less dependent on others. In case of non-symmetric graphs, a distinction has to be made between in-degree and out-degree (concept introduced in Chapter 3).

- *Betweenness Centrality* – measures the number of subjects that an individual is connecting indirectly, through their direct links. The underlying concept is that the more subjects depend on a certain individual to establish new connections, the more power that individuals has. In the case that two subjects can establish connection via more than one geodesic path, not all passing through the considered individual, the power of this individual is affected negatively. The implied computations are not easily made without computer assistance.

- *Closeness Centrality* – is the measure inverse to *farness* - the sum of the shortest distance (geodesic path) between each individual and every other subject in the considered network. The measure can be seen as a synthesis of the previous two measures. The underlying consideration is that the connectivity of an individual should be considered broader than the size of the immediate neighborhood or the bridge role the individual, but as the number of subjects the individual can reach (both directly and indirectly) and the minimum number of steps required. The more subjects an individual can reach, in the least steps, the more power he has.

These measures describe how a subject is embedded in the considered network structure. The resulted information helps the understanding of the individual influential potential within the network. Aside the individual considerations, sub-structures of the network such as groups and cliques are other mediators of understanding the network dependencies, behavior and dynamics. A *group* or a *clique* is a sub-set of subjects connected closer or more specific between one another than with other nodes in the network. A clique is specifically referring a dyad.

UCINET offers a rich library of centrality measurements supporting influence analysis in various network types. The study considers a total of six measures, two for each concept: *degree centrality, betweenness centrality* and *closeness centrality.* The specific measures are presented in the next section as well as the discussions based on measurements and observations, both at concept and content level.

In the network theory, non-symmetrical connections do not count. Therefore, the proposed measurement comparison study will be based on the symmetric matrix BKFRAB. The BKFRAC matrix will be initially considered only for a content comparison of the information resulted from regular observations and the information reflecting the subjects recall.

5.3 Results and Discussions

The start point of any social networking analysis is the structure of a network. For small networks like the one used in this study, certain conclusions can be already drawn from the network structural statistics. Table 4 below presents a collection of basic ego network measurements for each actor in the BKFRAT network.

Density Measures

	1 Size	2 Ties	3 Pairs	4 Densit	5 AvgDis	6 Diamet	7 nWeakC	8 pWeakC	9 2StepR	10 ReachE	11 Broker	12 nBroke	13 EgoBet
1	23.00	400.00	506.00	79.05	1.21	2.00	1.00	4.35	98.25	6.17	53.00	0.21	5.29
2	24.00	462.00	552.00	83.70	1.16	2.00	1.00	4.17	100.00	5.78	45.00	0.16	3.14
3	50.00	1654.002	450.00	67.51	1.32	2.00	1.00	2.00	100.00	3.18	398.00	0.32	27.25
4	43.00	1214.001	806.00	67.22	1.33	2.00	1.00	2.33	100.00	3.68	296.00	0.33	23.98
5	24.00	428.00	552.00	77.54	1.22	2.00	1.00	4.17	100.00	6.20	62.00	0.22	7.10
6	47.00	1514.002	162.00	70.03	1.30	2.00	1.00	2.13	100.00	3.32	324.00	0.30	22.21
7	52.00	1674.002	652.00	63.12	1.37	3.00	1.00	1.92	100.00	3.16	489.00	0.37	39.53
8	40.00	1102.001	560.00	70.64	1.29	3.00	1.00	2.50	100.00	3.85	229.00	0.29	19.94
9	46.00	1350.002	070.00	65.22	1.35	3.00	1.00	2.17	100.00	3.51	360.00	0.35	37.52
10	12.00	108.00	132.00	81.82	1.18	2.00	1.00	8.33	98.25	12.23	12.00	0.18	1.73
11	36.00	974.001	260.00	77.30	1.23	2.00	1.00	2.78	100.00	4.11	143.00	0.23	8.37
12	29.00	610.00	812.00	75.12	1.25	2.00	1.00	3.45	100.00	5.13	101.00	0.25	7.65
13	31.00	644.00	930.00	69.25	1.31	2.00	1.00	3.23	100.00	5.02	143.00	0.31	13.53
14	38.00	1018.001	406.00	72.40	1.28	2.00	1.00	2.63	100.00	4.01	194.00	0.28	12.31
15	41.00	1196.001	640.00	72.93	1.27	2.00	1.00	2.44	100.00	3.71	222.00	0.27	15.06
16	44.00	1268.001	892.00	67.02	1.33	2.00	1.00	2.27	100.00	3.62	312.00	0.33	24.29
17	39.00	1096.001	482.00	73.95	1.26	2.00	1.00	2.56	100.00	3.86	193.00	0.26	12.58
18	21.00	314.00	420.00	74.76	1.25	2.00	1.00	4.76	100.00	7.24	53.00	0.25	5.87
19	30.00	666.00	870.00	76.55	1.23	2.00	1.00	3.33	100.00	4.96	102.00	0.23	7.35
20	50.00	1618.002	450.00	66.04	1.34	2.00	1.00	2.00	100.00	3.21	416.00	0.34	30.88
21	23.00	432.00	506.00	85.38	1.15	2.00	1.00	4.35	100.00	6.15	37.00	0.15	2.79
22	34.00	858.001	122.00	76.47	1.24	2.00	1.00	2.94	100.00	4.40	132.00	0.24	8.23
23	38.00	990.001	406.00	70.41	1.30	3.00	1.00	2.63	100.00	4.09	208.00	0.30	21.27
24	26.00	522.00	650.00	80.31	1.20	2.00	1.00	3.85	98.25	5.43	64.00	0.20	4.56
25	25.00	504.00	600.00	84.00	1.16	2.00	1.00	4.00	100.00	5.74	48.00	0.16	3.12
26	22.00	374.00	462.00	80.95	1.19	2.00	1.00	4.55	100.00	6.47	44.00	0.19	4.92
27	42.00	1212.001	722.00	70.38	1.30	2.00	1.00	2.38	100.00	3.69	255.00	0.30	19.87
28	6.00	30.00	30.00	100.00	1.00	1.00	1.00	16.67	98.25	20.74	0.00	0.00	0.00
29	46.00	1402.002	070.00	67.73	1.32	2.00	1.00	2.17	100.00	3.44	334.00	0.32	22.57
30	42.00	1170.001	722.00	67.94	1.32	2.00	1.00	2.38	100.00	3.76	276.00	0.32	20.35
31	39.00	1078.001	482.00	72.74	1.27	2.00	1.00	2.56	100.00	3.91	202.00	0.27	14.60
32	16.00	194.00	240.00	80.83	1.19	2.00	1.00	6.25	98.25	9.18	23.00	0.19	2.71
33	27.00	552.00	702.00	78.63	1.21	2.00	1.00	3.70	100.00	5.35	75.00	0.21	8.53
34	42.00	1166.001	722.00	67.71	1.32	2.00	1.00	2.38	100.00	3.78	278.00	0.32	22.90
35	47.00	1454.002	162.00	67.25	1.33	2.00	1.00	2.13	100.00	3.39	354.00	0.33	24.83
36	38.00	1060.001	406.00	75.39	1.25	2.00	1.00	2.63	100.00	3.96	173.00	0.25	10.06
37	27.00	526.00	702.00	74.93	1.25	2.00	1.00	3.70	100.00	5.53	88.00	0.25	11.61
38	38.00	990.001	406.00	70.41	1.30	2.00	1.00	2.63	100.00	4.07	208.00	0.30	15.68
39	38.00	976.001	406.00	69.42	1.31	3.00	1.00	2.63	100.00	4.11	215.00	0.31	23.66
40	21.00	322.00	420.00	76.67	1.23	2.00	1.00	4.76	100.00	7.12	49.00	0.23	5.16
41	40.00	1138.001	560.00	72.95	1.27	2.00	1.00	2.50	100.00	3.82	211.00	0.27	15.23
42	37.00	1026.001	332.00	77.03	1.23	2.00	1.00	2.70	100.00	4.00	153.00	0.23	10.52
43	10.00	80.00	90.00	88.89	1.11	2.00	1.00	10.00	98.25	14.21	5.00	0.11	0.69
44	35.00	898.001	190.00	75.46	1.25	2.00	1.00	2.86	100.00	4.27	146.00	0.25	9.49
45	34.00	812.001	122.00	72.37	1.28	2.00	1.00	2.94	100.00	4.51	155.00	0.28	15.91
46	29.00	584.00	812.00	71.92	1.28	2.00	1.00	3.45	100.00	5.33	114.00	0.28	11.63
47	30.00	666.00	870.00	76.55	1.23	2.00	1.00	3.33	100.00	4.90	102.00	0.23	11.70
48	31.00	684.00	930.00	73.55	1.26	2.00	1.00	3.23	100.00	4.94	123.00	0.26	10.06
49	44.00	1310.001	892.00	69.24	1.31	2.00	1.00	2.27	100.00	3.56	291.00	0.31	21.48
50	28.00	600.00	756.00	79.37	1.21	2.00	1.00	3.57	100.00	5.10	78.00	0.21	10.81
51	6.00	26.00	30.00	86.67	1.13	2.00	1.00	16.67	89.47	26.15	2.00	0.13	0.50
52	7.00	38.00	42.00	90.48	1.10	2.00	1.00	14.29	96.49	19.93	2.00	0.10	0.40
53	37.00	996.001	332.00	74.77	1.25	2.00	1.00	2.70	100.00	4.08	168.00	0.25	13.50
54	44.00	1312.001	892.00	69.34	1.31	2.00	1.00	2.27	100.00	3.56	290.00	0.31	19.19
55	41.00	1194.001	640.00	72.80	1.27	2.00	1.00	2.44	100.00	3.72	223.00	0.27	15.24
56	42.00	1214.001	722.00	70.50	1.30	2.00	1.00	2.38	100.00	3.70	254.00	0.30	15.27
57	48.00	1570.002	256.00	69.59	1.30	2.00	1.00	2.08	100.00	3.26	343.00	0.30	18.64
58	34.00	836.001	122.00	74.51	1.25	2.00	1.00	2.94	100.00	4.44	143.00	0.25	10.15

Table 4 BKFRAT Ego Network Basic Measurements

From left to right, the measures in the table above represent:

Nr.	Measure	Description
1.	Size	Number of direct neighbors
2.	Ties	Number of ties in the ego-network
3.	Pairs	Number of ordered pairs in the ego network
4.	Density	*Ties* divided by *Pairs* (%)
5.	AvgDist	Average geodesic distance between pairs
6.	Diameter	Longest distance within the ego network
7.	nWeakComp	Number of weak components in the ego network
8.	pWeakComp	*nWeakComp* divided by *Size* (%)
9.	2StepReach	Number of nodes reached within 2 steps within the ego network
10.	ReachEffic	*2StepReach* divided by *Size* (%)
11.	Broker	Number of pairs not directly connected
12.	nBroker	*Broker* divided by *Pairs* (%)
13.	EgoBetweenness	Betweenness of ego in own network

Table 5 Description of UCINET Density Measures

In many cases though, a large number of direct neighbors or many direct and indirect links of a node does not necessarily represent a benefit. It is important to consider the structural measurements of a network in the context of the type of interactions in the community and the purpose of the analysis (competitive analysis, marketing analysis, sales analysis, relational analysis etc.).

Besides the right context, ego-centric measurements are a good start for defining a set of focal nodes (egos) for an analysis. Such measurements are focused on the individual rather than on the network as a whole and provide information over the *local networks* or *neighborhoods* of this individual. The information helps understanding how the network affects the individual subjects and it describes in somewhat the general texture of the network.

Considering the statistics presented in Table 5, a number of nodes stand out:

- *Nodes with high connectivity:* { #7, #3, #20, #57 }
- *Nodes with low connectivity:* { #51, #28, #52 }

For the proposed study of influence, the nodes with high connectivity can be seen as central or popular individuals interacting frequently with many of the fellow students.

1. Degree Centrality

This simple measure sums the number of adjacent links of an individual in a network. In case of directed data, it is important to distinguish between the centrality based on *in-degree* from the centrality based on *out-degree*.

If an individual receives many connections, has a high *in-degree,* he is considered *prominent* or with a *high prestige*. The fact that many other subjects want to connect to this individual indicates his importance. Individuals who have a high *out-degree* are usually interested in exchanging knowledge with others or they want to bring awareness to the community of their opinions. Such individuals are considered *influential* actors. In case of symmetric graphs, the out-degree equals the in-degree of each node.

Table 5 below presents the analysis on in-degrees and out-degrees of the BKFRAB (symmetric) and BKFRAC (un-symmetric) matrices.

FREEMAN'S DEGREE CENTRALITY MEASURES

Relation 1: BKFRAB

	1 OutDegree	2 InDegree	3 NrmOutDeg	4 NrmInDeg
1	32.000	32.000	1.101	1.101
2	72.000	72.000	2.477	2.477
3	315.000	315.000	10.836	10.836
4	109.000	109.000	3.750	3.750
5	101.000	101.000	3.474	3.474
6	280.000	280.000	9.632	9.632
7	402.000	402.000	13.829	13.829
8	119.000	119.000	4.094	4.094
9	156.000	156.000	5.366	5.366
10	21.000	21.000	0.722	0.722
11	128.000	128.000	4.403	4.403
12	66.000	66.000	2.270	2.270
13	60.000	60.000	2.064	2.064
14	128.000	128.000	4.403	4.403
15	159.000	159.000	5.470	5.470
16	208.000	208.000	7.155	7.155
17	191.000	191.000	6.570	6.570
18	36.000	36.000	1.238	1.238
19	68.000	68.000	2.339	2.339
20	379.000	379.000	13.037	13.037
21	43.000	43.000	1.479	1.479
22	88.000	88.000	3.027	3.027
23	81.000	81.000	2.786	2.786
24	41.000	41.000	1.410	1.410
25	55.000	55.000	1.892	1.892
26	30.000	30.000	1.032	1.032
27	161.000	161.000	5.538	5.538
28	6.000	6.000	0.206	0.206
29	162.000	162.000	5.573	5.573
30	135.000	135.000	4.644	4.644
31	99.000	99.000	3.406	3.406
32	37.000	37.000	1.273	1.273
33	85.000	85.000	2.924	2.924
34	111.000	111.000	3.818	3.818
35	163.000	163.000	5.607	5.607
36	77.000	77.000	2.649	2.649
37	50.000	50.000	1.720	1.720
38	87.000	87.000	2.993	2.993
39	103.000	103.000	3.543	3.543
40	35.000	35.000	1.204	1.204
41	118.000	118.000	4.059	4.059
42	69.000	69.000	2.374	2.374
43	12.000	12.000	0.413	0.413
44	79.000	79.000	2.718	2.718
45	99.000	99.000	3.406	3.406
46	64.000	64.000	2.202	2.202
47	65.000	65.000	2.236	2.236
48	75.000	75.000	2.580	2.580
49	122.000	122.000	4.197	4.197
50	52.000	52.000	1.789	1.789
51	9.000	9.000	0.310	0.310
52	16.000	16.000	0.550	0.550
53	114.000	114.000	3.922	3.922
54	179.000	179.000	6.158	6.158
55	115.000	115.000	3.956	3.956
56	158.000	158.000	5.435	5.435
57	276.000	276.000	9.494	9.494
58	67.000	67.000	2.305	2.305

Relation 2: BKFRAC

	1 OutDegree	2 InDegree	3 NrmOutDeg	4 NrmInDeg
1	259.000	165.000	90.877	57.895
2	172.000	150.000	60.351	52.632
3	171.000	206.000	60.000	72.281
4	237.000	212.000	83.158	74.386
5	146.000	169.000	51.228	59.298
6	170.000	193.000	59.649	67.719
7	159.000	196.000	55.789	68.772
8	166.000	139.000	58.246	48.772
9	158.000	179.000	55.439	62.807
10	165.000	157.000	57.895	55.088
11	180.000	195.000	63.158	68.421
12	150.000	173.000	52.632	60.702
13	164.000	150.000	57.544	52.632
14	194.000	191.000	68.070	67.018
15	155.000	197.000	54.386	69.123
16	184.000	193.000	64.561	67.719
17	174.000	176.000	61.053	61.754
18	200.000	129.000	70.175	45.263
19	189.000	185.000	66.316	64.912
20	156.000	208.000	54.737	72.982
21	115.000	141.000	40.351	49.474
22	174.000	175.000	61.053	61.404
23	170.000	180.000	59.649	63.158
24	151.000	135.000	52.982	47.368
25	141.000	172.000	49.474	60.351
26	156.000	142.000	54.737	49.825
27	166.000	170.000	58.246	59.649
28	171.000	108.000	60.000	37.895
29	164.000	186.000	57.544	65.263
30	146.000	173.000	51.228	60.702
31	192.000	180.000	67.368	63.158
32	210.000	137.000	73.684	48.070
33	183.000	177.000	64.211	62.105
34	208.000	181.000	72.982	63.509
35	179.000	202.000	62.807	70.877
36	171.000	158.000	60.000	55.439
37	179.000	137.000	62.807	48.070
38	172.000	155.000	60.351	54.386
39	193.000	166.000	67.719	58.246
40	134.000	137.000	47.018	48.070
41	180.000	169.000	63.158	59.298
42	129.000	170.000	45.263	59.649
43	134.000	160.000	47.018	56.140
44	167.000	179.000	58.596	62.807
45	115.000	157.000	40.351	55.088
46	148.000	149.000	51.930	52.281
47	165.000	169.000	57.895	59.298
48	168.000	151.000	58.947	52.982
49	191.000	187.000	67.018	65.614
50	192.000	169.000	67.368	59.298
51	126.000	105.000	44.211	36.842
52	157.000	133.000	55.088	46.667
53	146.000	171.000	51.228	60.000
54	175.000	187.000	61.404	65.614
55	141.000	153.000	49.474	53.684
56	203.000	185.000	71.228	64.912
57	157.000	223.000	55.088	78.246
58	152.000	178.000	53.333	62.456

DESCRIPTIVE STATISTICS

		1 OutDegree	2 InDegree	3 NrmOutDeg	4 NrmInDeg
1	Mean	168.448	168.448	59.105	59.105
2	Std Dev	26.129	24.435	9.168	8.574
3	Sum	9770.000	9770.000	3428.070	3428.070
4	Variance	682.730	597.075	84.054	73.509
5	SSQ	1685338.000	1680370.000	207490.063	206878.422
6	MCSSQ	39598.344	34630.344	4875.143	4263.508
7	Euc Norm	1298.206	1296.291	455.511	454.839
8	Minimum	115.000	105.000	40.351	36.842
9	Maximum	259.000	223.000	90.877	78.246
10	N of Obs	58.000	58.000	58.000	58.000

Network Centralization (Outdegree) = 32.330%
Network Centralization (Indegree) = 19.477%

DESCRIPTIVE STATISTICS

		1 OutDegree	2 InDegree	3 NrmOutDeg	4 NrmInDeg
1	Mean	109.793	109.793	3.777	3.777
2	Std Dev	84.133	84.133	2.894	2.894
3	Sum	6368.000	6368.000	219.057	219.057
4	Variance	7078.302	7078.302	8.376	8.376
5	SSQ	1109704.000	1109704.000	1313.158	1313.158
6	MCSSQ	410541.531	410541.531	485.811	485.811
7	Euc Norm	1053.425	1053.425	36.238	36.238
8	Minimum	6.000	6.000	0.206	0.206
9	Maximum	402.000	402.000	13.829	13.829
10	N of Obs	58.000	58.000	58.000	58.000

Network Centralization (Outdegree) = 10.228%
Network Centralization (Indegree) = 10.228%

Table 5 Degree Centrality BKFRAB versus BKFRAC

Considering first the measurement on the BKFRAB matrix (Relation 1), the set of influential individuals - as having the highest out-degree – is presented in Table 6 below. As this graph is symmetric, the in-degree of the nodes equals the out-degree.

Node	Nodal-out / Nodal-in degree
7	$go(n_7) = gi(n_7) = 402$
20	$go(n_{20}) = gi(n_{20}) = 379$
3	$go(n_3) = gi(n_3) = 315$
6	$go(n_6) = gi(n_6) = 280$
57	$go(n_{57}) = gi(n_{57}) = 276$

Table 6 BKFRAB Nodal-out / Nodal-in Degree

With NETDRAW, UCINET offers the possibility to visualize the degree centrality of a network. In Figure 26 the node ranking according to the nodal-out / nodal-in degree is visualized by the size of the nodes.

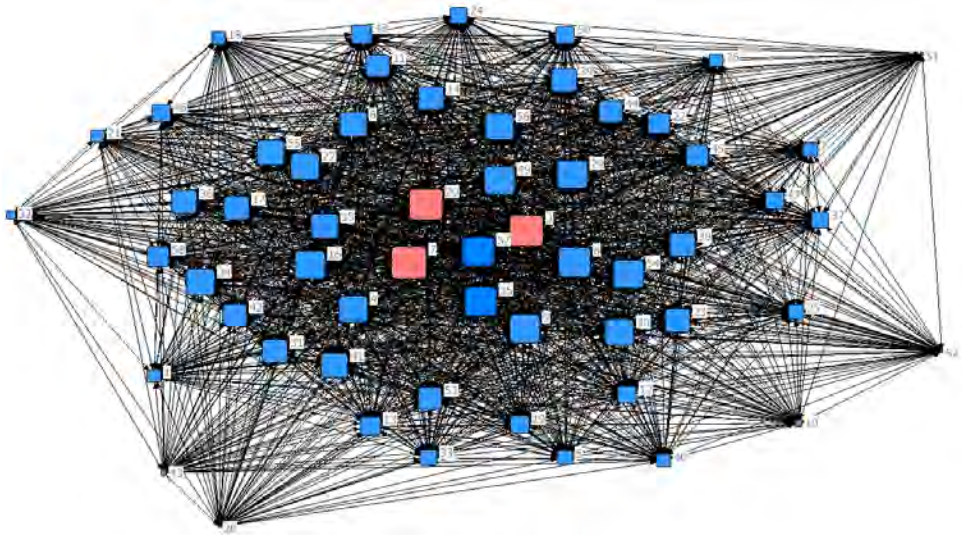

Fig. 26 BKFRAB Degree Centrality Ranking

In the student ranking graph, BKFRAC (Relation 2), though not symmetric, the picture of the interaction, influence and prestige of the subjects within the community is more balanced than in the BKFRAB graph. Also on individual level, the *out-degree* and in-*degree* measurements are in the same range and high. There is a high connectedness of the BKFRAC network – all nodes have both the in-degree and out-degree greater than 100.

Node	Nodal-out degree	Nodal-in degree
1	$go(n_1) = 259$	$gi(n_1) = 165$
4	$go(n_4) = 237$	$gi(n_4) = 212$
32	$go(n_{32}) = 210$	$gi(n_{32}) = 137$
57	$go(n_{57}) = 157$	$gi(n_{57}) = 223$
20	$go(n_{20}) = 156$	$gi(n_{20}) = 208$
3	$go(n_3) = 171$	$gi(n_3) = 206$

Table 7 BKFRAC Nodal-out and Nodal-in Degree

Table 7 presents the nodes with the highest out-degree and the highest in-degree in the BKFRAC network but the conclusions of influence should not be drawn based only on maximum values. The analysis should consider both measurements, in rapport to the average values of in-degree and out-degree in the network. The last two columns of the first panels in Table 5 presents the node degree counts expressed as percentages of the largest out and in-degree count in the dataset: *go(n_1) = 259* and respectively *gi(n_{57}) = 223.* Considering the overall context, a set of actors with influence potential in the BKFRAC network is {4, 1, 56, 57, 34, 35, 49, 54, 6}.

The second panels in Table 5, *Descriptive Statistics*, present the *mass* level of the degree centrality analysis in the two networks. In other words, the score distribution of the actor's degree centrality. In the BKFRAB graph, actors have on average a degree of 109.79, while in the ranking BKFRAC graph, the average degree is of 168.45. Both are rather high given the fact that there are only 58 actors in the surveyed community. When looking at the range of degrees between the two matrices (minimum and maximum values), the BKFRAB matrix has a considerable larger range than the BKFRAC. The same is reflected in the standard deviation and variance measurements. These indicators tell that the BKFRAB graph has the central nodes more clearly defined, while in the BKFRAC graph the interaction between subjects is more balanced.

The last information presented by the UCINET Freeman's degree centrality measures is the *Network centralization according to the in-degree and out-degree measurements.* These measures express the *degree of inequality or variance* in the global network as a percentage of the centralization of a perfect star network of the same size. A star network is considered reference as, independent of the number of nodes, it is the most centralized or the most unequal possible network.

BKFRAB is a symmetric matrix and therefore the out-degree and in-degree graph centralizations in this network are equal, of value 10.22%. In the BKFRAC network the out-degree graph centralization is 32.33% versus the in-degree graph centralization of 19.47%. Both graphs present a low *concentration* or *centralization.* The measures of *node centrality* and overall *network* or *graph centralization* should not be confused. *Node centrality* regards the connectivity of individual nodes while the *network centralization* refers to the overall cohesion or integration of the graph. In the considered case, BKFRAC presents a higher network centralization than BKFRAB due to a more balanced interaction, influence and prestige between the nodes in the network.

The degree centrality analysis compared the connectedness and centrality of the two graphs, BKFRAB and BKFRAC. The study focuses further on the comparison of various centrality measurements based on the BKFRAB network.

2. Betweenness centrality

The measure is based on the consideration that a node has power if it falls on the geodesic paths of other pairs in the network. The more subjects depend on one individual to make connections, the more power the individual has. A disadvantage is when a pair of subjects is connected by more geodesic paths and the individual does not lie on all these paths, *between* the two subjects. The measure can be normed as the percentage of the maximum possible *betweenness* a node could have. Table 8 presents the UCINET measurements of the betweenness centrality in the BKFRAB community.

Un-normalized centralization: 1604.565

	1 Betweenness	2 nBetweenness
7	39.613	2.482
9	36.787	2.305
20	30.542	1.914
3	27.504	1.723
35	23.168	1.452
39	22.822	1.430
16	22.630	1.418
4	22.408	1.404
6	22.247	1.394
34	21.121	1.323
29	20.763	1.301
49	20.357	1.275
23	20.101	1.259
27	18.439	1.155
57	18.102	1.134
8	18.097	1.134
30	17.953	1.125
54	17.900	1.122
45	14.701	0.921
41	14.174	0.888
55	14.133	0.886
15	13.681	0.857
56	13.378	0.838
31	12.928	0.810
38	12.808	0.803
53	12.363	0.775
17	10.867	0.681
47	10.208	0.640
14	9.807	0.615
37	9.540	0.598
42	9.375	0.587
13	9.118	0.571
36	8.498	0.532
50	8.472	0.531
58	7.888	0.494
46	7.743	0.485
44	7.304	0.458
48	7.241	0.454
11	6.940	0.435
22	6.576	0.412
33	6.461	0.405
5	5.238	0.328
19	5.020	0.315
12	4.889	0.306
26	3.927	0.246
40	3.053	0.191
18	2.906	0.182
24	2.799	0.175
1	2.748	0.172
25	1.957	0.123
2	1.775	0.111
21	1.668	0.105
32	1.167	0.073
10	0.663	0.042
43	0.219	0.014
52	0.113	0.007
51	0.101	0.006
28	0.000	0.000

DESCRIPTIVE STATISTICS FOR EACH MEASURE

		1 Betweenness	2 nBetweenness
1	Mean	11.948	0.749
2	Std Dev	9.211	0.577
3	Sum	693.000	43.421
4	Variance	84.847	0.333
5	SSQ	13201.304	51.826
6	MCSSQ	4921.148	19.320
7	Euc Norm	114.897	7.199
8	Minimum	0.000	0.000
9	Maximum	39.613	2.482
10	N of Obs	58.000	58.000

Network Centralization Index = 1.76%

Table 8 BKFRAB Betweenness Centrality

From the statistics it can be observed that the *betweenness* of the individual network nodes varies on a range between 0 to 39.61 and that the overall network centralization is rather low, of value 1.76%. This can be justified by the fact that the network presents a high connectedness, with a large number of direct connections. Though there is little power in the network, the nodes # 7, # 9, # 20 and #3 stand out. A similar conclusion was drawn also on structural basis, with the Freeman's degree centrality measurement.

Figure 27 presents graphically the BKFRAB node ranking according to node betweenness centrality.

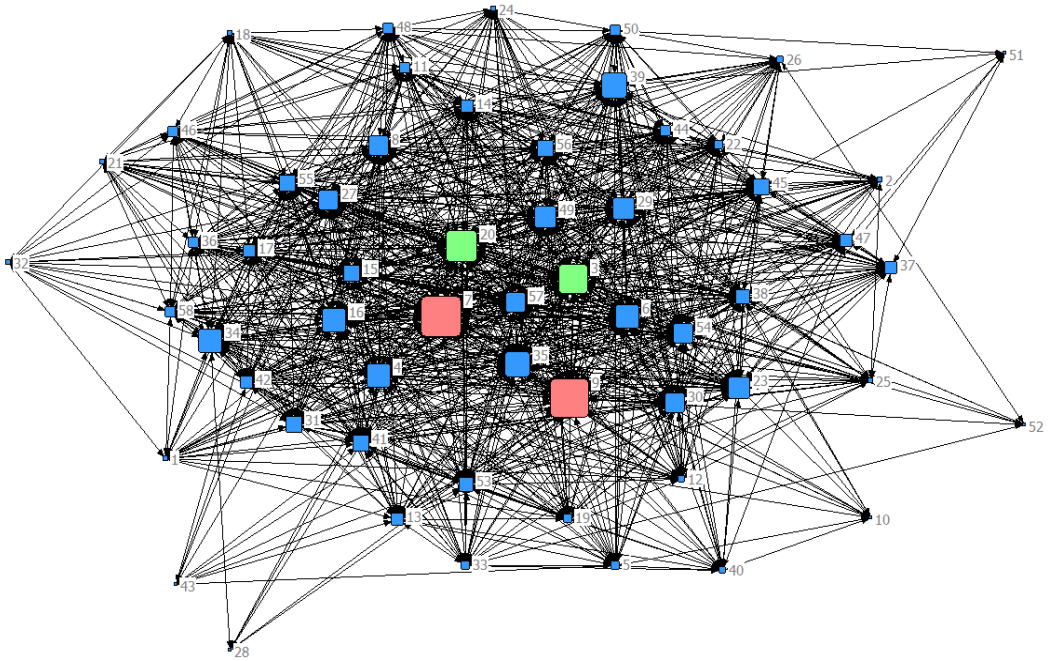

Fig. 27 BKFRAB Betweeneess Centrality Ranking

3. Closeness centrality

The degree centrality measurements have the disadvantage that take into account only the direct links a node has, rather than also the indirect links to all other nodes in the network. Differently, *closeness centrality* emphasizes this aspect and considers the geodesic distance of a node to all other nodes in the network. The sum of the geodesic distances of a node is considered the *farness* of the node from all other network nodes. The reciprocal of farness (that is one divided by the farness) represents the measure of *nearness* or *closeness* centrality. This can be normed relatively to the most central node of the actor. The *closeness centrality* UCINET results for the considered information exchange data, BKFRAB, is presented in Table 9.

Closeness Centrality Measures

	1 Farness	2 nCloseness
7	62.000	91.935
3	64.000	89.063
20	64.000	89.063
57	66.000	86.364
6	67.000	85.075
35	67.000	85.075
9	68.000	83.824
29	68.000	83.824
54	70.000	81.429
16	70.000	81.429
49	70.000	81.429
4	71.000	80.282
34	72.000	79.167
30	72.000	79.167
27	72.000	79.167
56	72.000	79.167
55	73.000	78.082
15	73.000	78.082
41	74.000	77.027
8	74.000	77.027
17	75.000	76.000
31	75.000	76.000
38	76.000	75.000
39	76.000	75.000
14	76.000	75.000
23	76.000	75.000
36	76.000	75.000
42	77.000	74.026
53	77.000	74.026
11	78.000	73.077
44	79.000	72.152
58	80.000	71.250
45	80.000	71.250
22	80.000	71.250
13	83.000	68.675
48	83.000	68.675
47	84.000	67.857
19	84.000	67.857
46	85.000	67.059
12	85.000	67.059
50	86.000	66.279
33	87.000	65.517
37	87.000	65.517
24	89.000	64.045
25	89.000	64.045
2	90.000	63.333
5	90.000	63.333
21	91.000	62.637
26	92.000	61.957
1	92.000	61.957
18	93.000	61.290
40	93.000	61.290
32	99.000	57.576
10	103.000	55.340
43	105.000	54.286
52	109.000	52.294
28	109.000	52.294
51	114.000	50.000

Statistics

		1 Farness	2 nCloseness
1	Mean	80.897	71.894
2	Std Dev	11.939	9.859
3	Sum	4692.000	4169.875
4	Variance	142.541	97.205
5	SSQ	387834.000	305428.500
6	MCSSQ	8267.379	5637.905
7	Euc Norm	622.763	552.656
8	Minimum	62.000	50.000
9	Maximum	114.000	91.935
10	N of Obs	58.000	58.000

Network Centralization = 41.15%

Table 9 BKFRAB Closeness Centrality

The table above presents the results of individual closeness centrality ordered on *nCloseness,* the *normalized closeness*. Similar to the results of the prior centrality measurements, in this case too, nodes #7; #3 and #20 are most central.

Node #7, connected to 52 nodes with 1674 ties, is the *closest* or *most central* node in the network, meaning that the sum of its geodesic distances to all other nodes in the network is the smallest. Within the BKFRAB graph, node # 51 has the greatest farness. Figure 28 presents graphically the ranking of the node farness in the BKFRAB network.

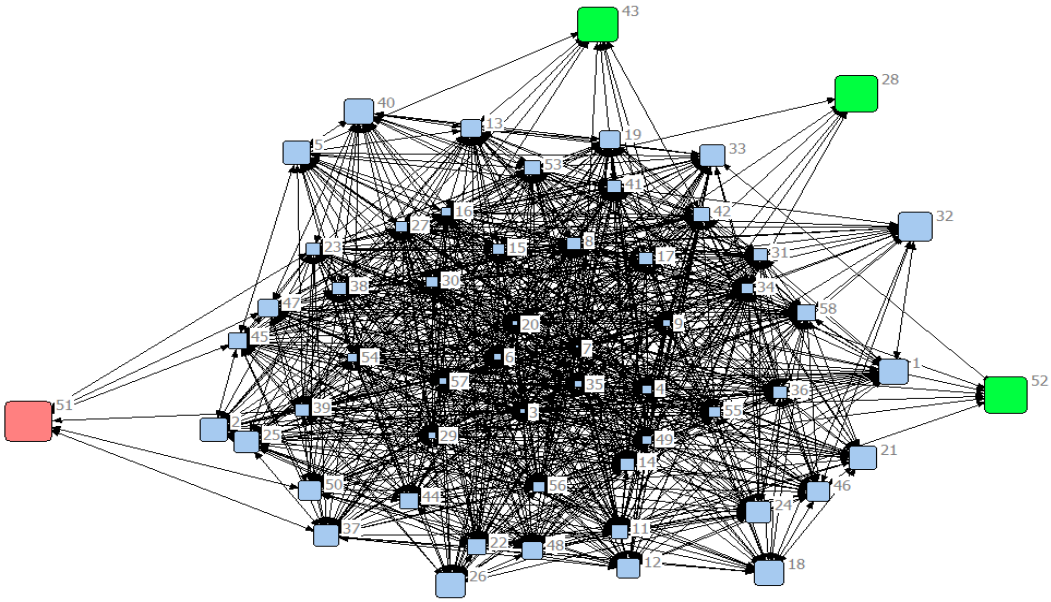

Fig. 28 BKFRAB Farness Centrality Ranking

In a small network with high density as the one used in this study, the geodesic distance based centrality is usually similar to the centrality based on connectivity adjacency. This is justified by the fact that many geodesic distances in such a network are adjacencies. This does not apply in larger or less dense networks.

As the other centrality measures, closeness centrality can also characterize the centralization of the entire network. In this case, the *network centralization* presents a better concentration in the network, 41.15%. The measure presents how unequal is the distribution of centrality across the network nodes compared to the variance in a star network of the same size. In a star network the distribution of farness of actors presents the maximum possible concentration, with one node being maximally close to all others and all other nodes being maximally distant from one another.

Each of the three basic centrality measurements – *degree centrality*, *betweenness centrality* and *closeness centrality* are capturing specific properties of the network based on which conclusions of individual node power and influence can be drawn. Further in the study, a set of other three centrality and power measures are considered, as alternatives to the ones already discussed.

4. The Bonacich Power index

Phillip Bonacich suggested an extension of the *degree centrality* measure. The original degree centrality approach considers that the more connections and individual has, the more power he presents. Bonacich reasoned that the centrality of an individual depends on the number of direct connections he has as well as on the number of connections his neighbors have. In other words, being connected to many others in the network might imply that one is centrally positioned but not necessarily that he has a significant power in the network. An individual connected to others not too well connected has a great power on the consideration that those subjects depend more on him for information exchange then if they would have been better connected.

Table 10 presents the *Bonacich power index* measurements based on the BKFRAB graph.

Bonacich Power

	1 Power	2 Normali
1	32.000	1.762
2	72.000	3.964
3	315.000	17.343
4	109.000	6.001
5	101.000	5.561
6	280.000	15.416
7	402.000	22.134
8	119.000	6.552
9	156.000	8.589
10	21.000	1.156
11	128.000	7.047
12	66.000	3.634
13	60.000	3.304
14	128.000	7.047
15	159.000	8.754
16	208.000	11.452
17	191.000	10.516
18	36.000	1.982
19	68.000	3.744
20	379.000	20.867
21	43.000	2.368
22	88.000	4.845
23	81.000	4.460
24	41.000	2.257
25	55.000	3.028
26	30.000	1.652
27	161.000	8.864
28	6.000	0.330
29	162.000	8.919
30	135.000	7.433
31	99.000	5.451
32	37.000	2.037
33	85.000	4.680
34	111.000	6.111
35	163.000	8.975
36	77.000	4.240
37	50.000	2.753
38	87.000	4.790
39	103.000	5.671
40	35.000	1.927
41	118.000	6.497
42	69.000	3.799
43	12.000	0.661
44	79.000	4.350
45	99.000	5.451
46	64.000	3.524
47	65.000	3.579
48	75.000	4.129
49	122.000	6.717
50	52.000	2.863
51	9.000	0.496
52	16.000	0.881
53	114.000	6.277
54	179.000	9.855
55	115.000	6.332
56	158.000	8.699
57	276.000	15.196
58	67.000	3.689

Table 10 BKFRAB Bonacich Power Index

The second column in the table above presents the absolute value of the Bonachich power index scores. With this measure too, for the reasons presented above, nodes # 7, # 20 and # 3 are seen most central and powerful.

5. Flow centrality

The *betweenness centrality* measure prior discussed considers individuals in a positional advantage if they often fall on the shortest geodesic path between others in the network. The *Flow Centrality approach* expands this concept by assuming that subjects will make use of these indirect paths proportionally to the path length. With this consideration, the measure sums for each network node his involvement in all flows between all other pairs of the network. As the magnitude of such measurement increases with the sheer size of the network or the network density, the *flow betweenness* of a network node is calculated in ratio to the *total flow betweeness* in the network that does not involve the considered node.

FLOW BETWEENNESS CENTRALITY MEASURES

	1 FlowBet	2 nFlowBet
1	15.477	0.485
2	35.489	1.112
3	137.427	4.305
4	77.698	2.434
5	53.048	1.662
6	92.828	2.908
7	195.516	6.125
8	62.955	1.972
9	148.879	4.664
10	10.892	0.341
11	67.356	2.110
12	31.984	1.002
13	41.743	1.308
14	43.049	1.349
15	58.508	1.833
16	95.032	2.977
17	66.194	2.074
18	27.550	0.863
19	35.721	1.119
20	128.507	4.026
21	25.175	0.789
22	57.492	1.801
23	53.074	1.663
24	18.509	0.580
25	22.213	0.696
26	23.067	0.723
27	89.520	2.804
28	1.359	0.043
29	70.039	2.194
30	65.521	2.053
31	57.889	1.814
32	26.473	0.829
33	34.868	1.092
34	80.534	2.523
35	86.806	2.719
36	37.910	1.188
37	34.888	1.093
38	45.719	1.432
39	103.902	3.255
40	20.882	0.654
41	67.818	2.125
42	31.942	1.001
43	4.571	0.143
44	28.543	0.894
45	56.026	1.755
46	51.514	1.614
47	50.398	1.579
48	54.647	1.712
49	75.623	2.369
50	31.202	0.978
51	8.714	0.273
52	9.995	0.313
53	59.954	1.878
54	66.923	2.097
55	49.747	1.558
56	84.530	2.648
57	100.428	3.146
58	39.985	1.253

Network Centralization Index = 4.444%

DESCRIPTIVE STATISTICS FOR EACH MEASURE

		1 FlowBet	2 nFlowBet
1	Mean	56.108	1.758
2	Std Dev	36.883	1.155
3	Sum	3254.254	101.950
4	Variance	1360.368	1.335
5	SSQ	261490.438	256.643
6	MCSSQ	78901.328	77.439
7	Euc Norm	511.361	16.020
8	Minimum	1.359	0.043
9	Maximum	195.516	6.125
10	N of Obs	58.000	58.000

Table 11 BKFRAB Flow Centrality

Also when using such elaborated betweenness centrality measure (Table 11), node # 7 is seen as the most influential and powerful individual. In fact, the overall picture of individual centrality and power does not change a great deal – the set of individuals that stand out in this case is: {#7; #9; #3; #20}.

Looking at the *descriptive statistics,* the relative variability in *flow betweenness* for BKFRAB is rather high, of approximately 58%. Despite this aspect, the degree of concentration in the distribution of flow betweenness centrality among the nodes of the network, network centralization, is fairly low (4.44%) when compared with that of a star network. Still, this is greater than the network centralization index obtained in the case of geodesic distance based betweenness measurement (1.76%).

6. Eigenvector of the geodesic distances.

The *closeness centrality* measure prior presented is based on the sum of all geodesic distances from one individual to all others in the network (farness). This measure can be though misleading in larger and more complex networks. The reasoning is that if considering two nodes: node 1- close to a small and closed sub-group of the network but far away from many others in the network and node 2 - at a moderate distance from all nodes in the network, the farness measures for the two nodes might be of similar magnitude. In the presented example, node 2 is though fairly more central placed that node 1 as it is able to reach more nodes in the network, with a similar effort. This concept is fundamental in the *eigenvector of the geodesic distances* measure.

This measure represents the effort to find the individuals with the smallest farness from all others in the entire network (the most central nodes) and not only locally. For this, the method used is *factor analysis*, which identifies *dimensions* of the distances between network nodes. The positioning of a node with respect to each defined dimension is the *eigenvalue*. The collection of eigenvalues is called *eigenvector*. In general, *global* aspects of distances between nodes are described by the first dimension and more specific or more local properties and described secondary dimensions. This approach is more suitable for an analysis of information *exchange* relationships, rather than an analysis over the senders and receivers of information within a network.

UCINET incorporates the *eigenvector of geodesic distance* measurement in the *Bonacich Centrality.* The calculated eigenvectors of the distance matrix for the BKFRAB is presented in Table 12.

Bonacich Eigenvector Centralities

	1 Eigenvec	2 nEigenvec
1	0.024	3.386
2	0.055	7.832
3	0.273	38.678
4	0.094	13.313
5	0.086	12.143
6	0.328	46.366
7	0.370	52.292
8	0.101	14.294
9	0.124	17.570
10	0.017	2.344
11	0.115	16.270
12	0.057	8.053
13	0.038	5.416
14	0.131	18.572
15	0.144	20.358
16	0.196	27.737
17	0.186	26.259
18	0.023	3.232
19	0.053	7.473
20	0.395	55.858
21	0.034	4.804
22	0.070	9.958
23	0.060	8.502
24	0.033	4.617
25	0.043	6.086
26	0.023	3.278
27	0.149	21.028
28	0.007	0.936
29	0.153	21.615
30	0.137	19.374
31	0.082	11.574
32	0.026	3.732
33	0.113	15.983
34	0.102	14.402
35	0.145	20.525
36	0.060	8.548
37	0.036	5.143
38	0.076	10.699
39	0.073	10.296
40	0.035	4.940
41	0.098	13.800
42	0.063	8.906
43	0.008	1.123
44	0.069	9.725
45	0.073	10.291
46	0.049	6.937
47	0.057	8.027
48	0.057	8.065
49	0.093	13.111
50	0.041	5.743
51	0.003	0.423
52	0.013	1.883
53	0.115	16.241
54	0.176	24.831
55	0.122	17.198
56	0.141	19.957
57	0.276	39.054
58	0.053	7.566

Descriptive Statistics

		1 Eigenvec	2 nEigenvec
1	Mean	0.100	14.075
2	Std Dev	0.086	12.113
3	Sum	5.773	816.365
4	Variance	0.007	146.715
5	SSQ	1.000	20000.004
6	MCSSQ	0.425	8509.449
7	Euc Norm	1.000	141.421
8	Minimum	0.003	0.423
9	Maximum	0.395	55.858
10	N of Obs	58.000	58.000
11	N Missing	0.000	0.000

Network centralization index = 49.01%

Table 12 BKFRAB Bonacich Eignevector Centrality

In the statistics above, the first measurements, the *eigenvector*, indicates how much of the overall pattern of paths between the network nodes can be considered *global patterns*. The rest of the patterns are considered *additional* or *local patterns*. The higher the score, the *more central* an individual is with respect to the main pattern of node pair distances in the network. The nodes with low scores are seen *peripheral*. For the BKFRAB graph, with this measure too, the nodes # 7, # 20, # 57 and # 3 are most central and nodes # 51 and # 28 are most peripheral.

The *descriptive statistics* third set of results presented in Table 9 examines the individual centralities distribution and the overall centralization of the network. The results indicate little variability in node centralities: the standard deviation is *0.08* relative to a mean of *0.1*. This means that the centrality and power in the network is balanced.

At global level, the degree of inequality or concentration of the BKFRAB network is 49% when compared with a star network of the same size. The result is slightly higher when compared with the *network centralization* based on *closeness centrality*, of 41%.

Figure 29 below presents graphically the *eigenvector* measurement on the BKFRAB network.

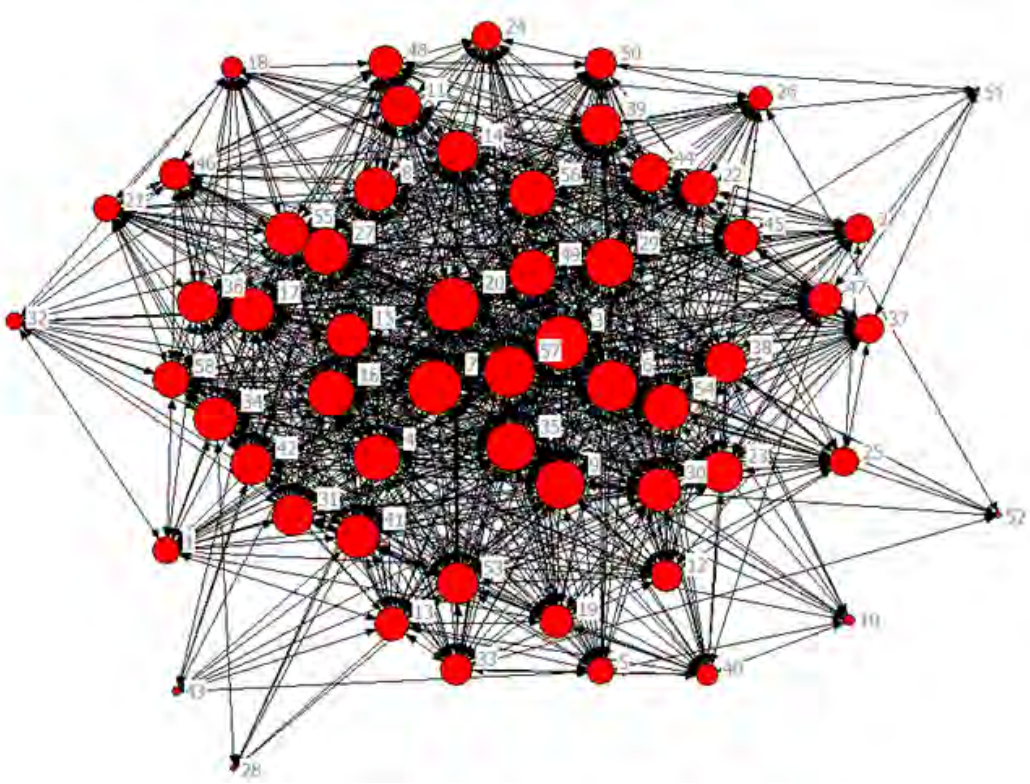

Fig 29 BKFRAB Bonacich Eignevector Centrality Ranking

The factor analysis approach could be applied also in the degree or betweenness measurements. In general, *geodesic distances* between the network nodes are reasonable indicators of centrality. The indication of centrality might though refer more local or more global contexts.

5.4 Conclusions

The social network analysis software UCINET provides a set of useful measures for some of the most important aspects of a social community, the sources and the distribution of influence. From a network approach, influence is seen as centrality and power, both at individual level as well as at group or global level. The centrality and power of individuals can be determined on network structural basis and arises from relations with others in the network. In general, the individual power results from an advantageous relational positioning of a node in rapport with others in the network. At global level, the power of a social structure may result from variations in connectivity patterns between the individual nodes.

In the social networking theory, there are three fundamental sources of centrality and power: *high degree*, *high betweenness* and *high closeness.* The proposed case study compared a set of six centrality and power measures, applied to the UCINET dataset *BFRAT (Bernard & Killworth Fraternity)*: *Freeman Degree Centrality, Bonacich Degree Centrality, Freeman Betweenness Centrality, Flow Betweenness Centrality, Path Distance Closeness Centrality* and *Eigenvector of Geodesic Distances Closeness Centrality.* The first three measures are the basic measures of centrality, the other three measures are more elaborated formulations of centrality and power concepts. The purpose of the study was presenting and comparing the main centrality and power measurements and exercising them with UCINET 6.0.

The different definitions and measures considered capture different concepts of sources of centrality and power and therefore present specific insights in the social structure. Applied to the valued, symmetric graph - BFRAB, a rather small but dense network, the six measures concluded in unanimity a set of influential nodes: { #7; #3; #20}. The main conclusion of the study is that there is no *right* or *wrong* approach in the centrality and power measurement. Such measurements need to be considered in the context of the studied problem, the type of relationships in the network and according to the specifics of the considered measures.

The question of how structural position confers influence or power in social networks, remains a topic of interest in the SNA research. Understanding how an entities (company, person, product) interact within virtual communities is a fundamental aspect for professional and social successes.

The *International Network for Social Network Analysis (*http://www.insna.org/software/index.html*)* community collects and maintains information related to the Social Networking Analytics discipline: research papers, documentation, available data and the list of accredited SNA software packages. Furthermore, a number of social analysis tools are available today online, many free of charge:

- PostRank Connect - in-depth analytics with social layer
- Twitalyzer - in-depth analytics with Twitter layer
- Crowdbooster - ranks own best Tweets and the optimal time to send these
- Tweriod – measures active twitter followers
- Yottaa - monitors and tracks website performance
- Google Analytics - full traditional website analytics
- Klout - social influence rating based on Twitter, Facebook, and LinkedIn
- PeerIndex - social score based on Twitter, Facebook, LinkedIn, Quora, and personal blog
- BuzzFeed - social traffic tracking
- Google Page Speed - tests and offers solutions to speed up personal site

6. CASE STUDY 2: Co-authorship Link Prediction

The *Link Prediction* problem treats the probability with which two unconnected nodes will be connected in the future by a direct link. The prediction task is in fact a proximity measure between two network nodes. If knowing this, the objective over the entire network can be optimized. This problem is one of the main interests within social networking analytics as it facilitates the understanding of social groups and their behavior.

Achieving this understanding can help the development of analytical applications specialized on identification of hidden patterns, behavior prediction, missing link detection etc. These types of data analysis are today applicable in most domains (i.e. marketing, health, security, sociology, criminal investigation, education etc.)

The domain choice of this study is the scientific co-authorship and the considered problem is the prediction of prospective collaborations (links). This problem is identical with the link prediction task in many other social networks, from both the structural and conceptual point of view. The authors in the co-authorship community establish collaboration links with a mutual purpose. Such network respects a fundamental social networking property, the *power-law distribution.*

The proposed study concerns the future collaborations between the author members of the DBLP community. The goal of the study is a comparative analysis of accuracy of a set of three classification algorithms: *Support Vector Machine (SVM), K-Nearest Neighbor* and *Naïve Bayes* applying a particular set of features: *sum of papers, weighted sum of neighbors, weighted sum of secondary neighbors* and w*eighted shortest distance.*

6.1 Data

The DBLP (http://www.informatik.uni-trier.de/~ley/db/) is the largest bibliographic dataset of Computer Science publications available on Internet. The dataset has 442,886 members and 678,296 papers. The dataset gathers information over Computer Science publications, journals and conferences over the period 1936 – 2009. The DBLP dataset was built by manual entry, either by the authors themselves or by DBLP members.

Number of Authors	696360
Number of Papers	1158648
Avg. authors per paper	2.11
Avg. papers per author	3.26

Table 13 DBLP statistics

On average, the publication contribution in the network is of 3.26 per author. Looking at the connectedness of the network, the author in this network have frequently interacted and collaborated on their publications. The indicator of this behavior is the high average of authors per paper reported to the number of publications per author, 2.11 versus 3.26. The network is clearly a well formed research community.

Also, different than other datasets (i.e. CiteSeer) the DBLP registers the full name of the authors per publication, this making the author identification easier and more accurate.

The reasons for selecting the DBLP dataset for the proposed study are its size, accessibility, structure and the high connectedness degree of the captured community. For the same reasons, a large number of various other SNA studies have been based on this dataset.

6.2 Approach

The link prediction problem is approached as a classification problem of candidate versus non-candidate links. Today there are available numerous classification algorithms suitable for supervised learning. Many of them have a comparable performance, some though suiting better than others for specific types of data analysis. This study will investigate the accuracy and fit of three classification algorithms for the co-authorship problem, considering the DBLP data set:

1. **A Support Vector Machine (SVM)** is a classification algorithm using an N-dimensional hyper plane and is best applicable for separating network data in two categories. In this case the classification regards exactly two categories: candidate and non-candidate. Using a sigmoid kernel function, the model relates somewhat to a two layer perceptron neural networks. The main difference is that the SVM teaching model uses linear constraints for solving a quadratic programming problem. The model uses *attributes* as predictor variables, *features* to define the hyper plane and *vectors* as the feature sets describing the classification classes. The SVM algorithm used in this study was the one provided by the SVM-Light implementation (http://svmlight.joachims.org/). The software offers two kernel functions: a linear and a nonlinear Kernel (also known as Radial Basis Function Kernel). Following prior recommendations, for this study the RBF kernel is used.

2. **K-Nearest Neighbor** is a fundamental though rather simple classification algorithm. K-NN is a type of instance-based learning with the principle of classifying objects based on closest training examples in the feature space. Typically, the *k* indicator is a small positive integer. For this study, the K-NN algorithm was developed in MATLAB®. The tool offering a high-level programming language and enabling intensive computational tasks.

3. **Naive Bayes** classifier is another fundamental classifier, though of probabilistic nature. The applied concept is the Bayes' theorem, the algorithm considering that all features contribute independently to the calculated probability for the classification. In general, Naïve Bayes algorithms can efficiently be trained for specific supervised learning. In this study the supporting technology for this algorithm was Weka. Weka being a collection of machine learning algorithms, mainly applicable for data mining investigations.

The core of any machine learning algorithm is though the applied feature set. Such features depend on the type of relationships in the considered network, the domain and the goal of the proposed analysis. For the link prediction problem, Mohammad et. al [2006] defined three categories of features: *proximity features, aggregated features* and *topological features*.

Proximity features regard similarity aspects between network entities. Considering the case of a co-authorship network, a suitable proximity feature could be the closeness of two authors based on the mapping of a set of identical keywords representing their research work. Authors with similar interests and expertise are more likely to collaborate. *Aggregated features* are mainly focusing on the individual characteristics of a node and are then combined for the evaluation of a candidate pair. An exemplification for this domain could be the joint indication of how prolific a pair of two authors is. The measure of prediction being defined as: if at least one author of a considered pair is prolific then it is more likely that the two authors will collaborate in the future. A *topological feature* could be the shortest path between two authors. The underlying consideration is that the shorter the distance between the authors, the better the chance that in the future these authors will collaborate.

As the considered dataset DBLP does not provide a matching indicator, be it keywords, classification or grouping dimensions, this study considers a particular set composed by only *aggregated* and *topological* features:

1. **Sum of papers** – an *aggregated feature* summing the number of published papers by the authors of a candidate author-pair. An author having a large number of publications is considered prolific. The feature is supported by the consideration that if at least one of the considered pair is prolific, these authors will collaborate in the future with a higher probability.

2. **Weighted sum of neighbors** – an *aggregated feature* indicated the social connectedness of a pair of authors by summing the weighted connections of these authors with their first neighbors. For this study, the weight of a connection is considered to be the number of publications between two actors. The feature is supported by the consideration that the higher the connectedness of the candidate pair, the higher the probability that the two authors will collaborate in the future. This feature can be considered also of topological nature as the total number of neighbors of a node (in this case weighted) is in fact the degree of the node.

3. **Weighted sum of secondary neighbors** – an *aggregated feature* indicating the second level of social connectedness of an author by summing the weighted connections between his direct neighbors and their first neighbors. Also in this case, the weight represents the number of shared collaboration between two specific authors. The consideration supporting this feature is strongly domain related: in the field of scientific collaboration, an author having established collaboration with a high connected other author might collaborate with high probability with the co-author's highest ranked neighbor(s). In certain types of social networks, this feature has the disadvantage of large computations. Similar to the previous feature, this too can be considered a topological feature.

4. **Weighted shortest distance** – a *topological feature* indicating the shortest path between two author-nodes. The proposed measurement is the weighted network hop count. The same weighting of the network connection is considered: the number of collaborations of two authors. The consideration supporting the feature is that the closer two nodes are positioned, the greater the probability that they will establish a collaboration in the future.

Against this feature set, the performance and fit of the three algorithms: *Support Vector Machine (SVM), K-Nearest Neighbor (K-NN)* and *Naïve Bayes* will be measured by a set of standard indicators: *accuracy*, *precision*, *recall* and *F-value*. It is important to mention to point that alone, the *precision* and *recall* measurements are not strong indicators of performance. With respect to a classification class, *Recall* represents the percentage of objects that are relevant for classification and *Precision* represents the percentage of relevant objects that are classified. Therefore, these two measures alone are not good enough indicators of the performance of recognition on a specific class. The *F-value* measurement is the harmonic mean on the *recall* and *precision*: $F = \frac{2RP}{R+P}$ and is a more suitable indicator of the power of a learning algorithm on a specific class.

In conclusion, the most relevant indicators for the proposed comparison are in fact *Accuracy* and *F-value.* The next section discusses the results and observations of the proposed performance testing.

6.2 Results and Discussions

Initially, the three algorithms were measured independently for investigating any additional required tuning. For the considered dataset, the default parameter values of WEKA and SVM light performed satisfactory. The only sensitive improvement needed was for the K-NN algorithm, K = 25 performed best.

The tests were run three times for consolidation of the obtained results. The graphics in Figure 30 below present the average performance indication of the three classification algorithms:

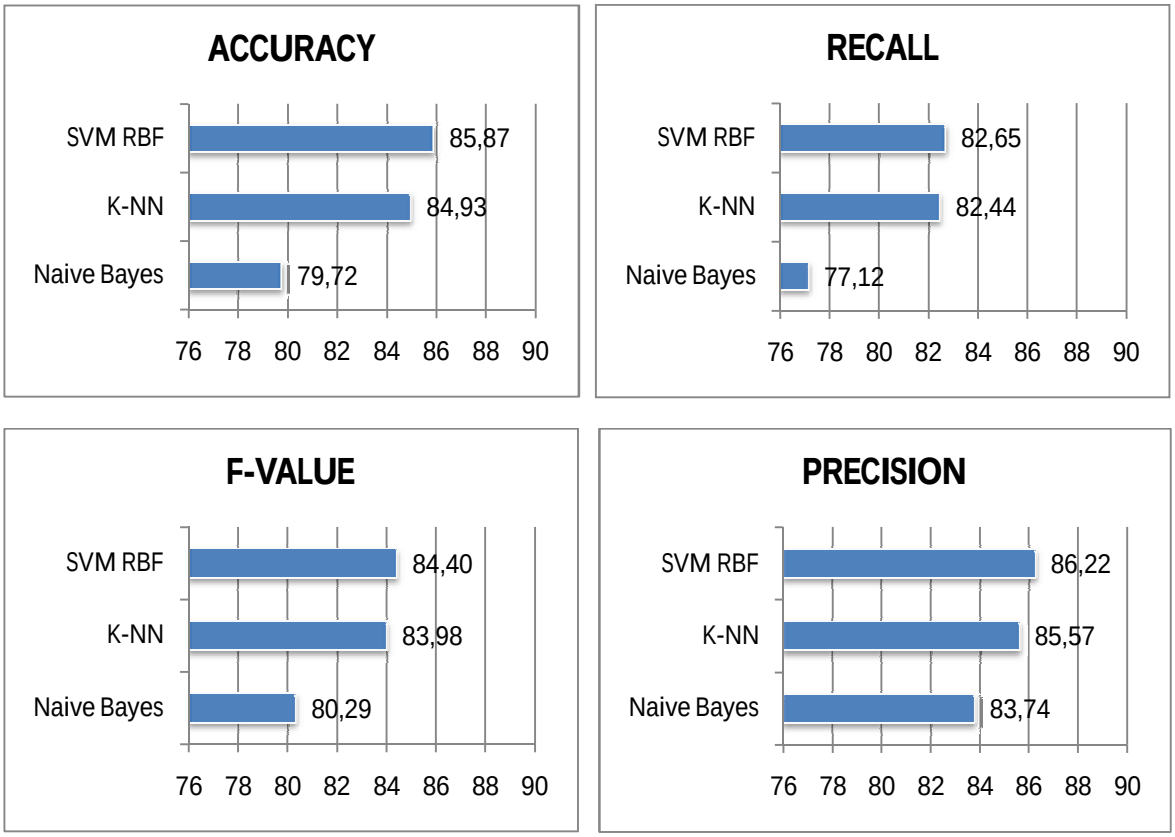

Fig. 30 Algorithm Performance Statistics on DBLP per Measurement

Considering the averaged *accuracy* indicator, SVM with RBF kernel performed best, with an accuracy of 85.87%. The K-NN classifier had a satisfactory performance as well, with an averaged *accuracy* of 84.93%, a difference of less than 1% when compared to SVM RBF performance. Such small difference is insignificant statistically, meaning that the two algorithms performed comparably with respect to accuracy. The *precision-recall* and *F-value* measurements indicate too the similar capabilities of the SVM RBF and K-NN algorithms in detecting patterns for candidate classification of future author collaborations within the DBLP community.

According to all metrics, the Naive Bayes algorithm performed considerable less. The averaged accuracy was only 79.42% and the averaged *F-Value* was 80,29%. This means a difference of about 5% when compared with the performance of SVM RBF. Clearly, the Naïve Bayes algorithm, used with this particular feature set, is rather weak in detecting patterns for link prediction classification in the DBLP dataset.

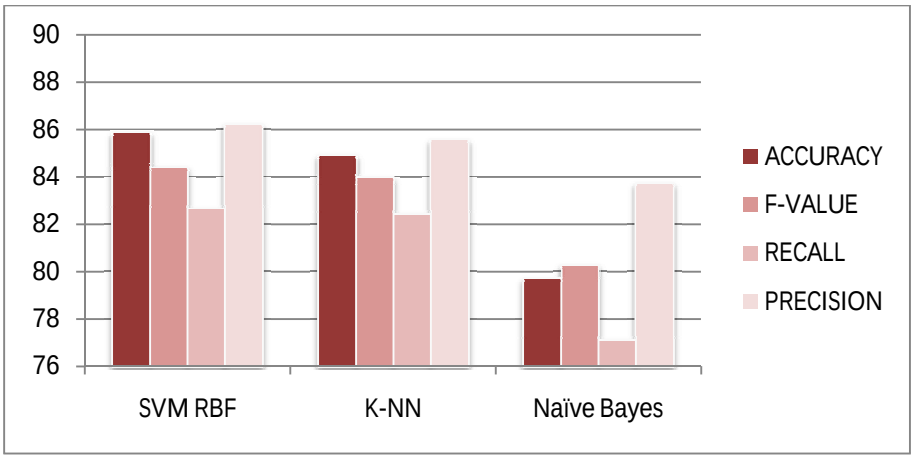

Fig. 31 Algorithm Performance Statistics on DBLP

When comparing the *precision* and *recall* results, it can be noticed that all classifiers have the precision higher than the recall. This indicates that the models are missing actual links, more than they are predicting false links. In the domain of co-authorship this can be explained by the fact that some author pairs might co-authorship accidentally. This can be explained by the fact that in the research world it often happens that researchers of a the same branch name each other in various materials. Another reason might be that two different authors share the same name. In this dataset the author name is the identification criteria and clearly this might not always be sufficient.

Looking then at the way the three algorithms ranked the four features, as presented in Table 14 below, the SVM RBF proposes a different approach than K-NN and Naïve Bayes:

	SVM	K-NN	Naive Bayes	AVG. RANK
Sum of papers	2	4	4	4
Weighted sum of neighbors	3	2	2	2
Weighted sum of secondary neighbors	4	3	3	3
Weighted shortest distance	1	1	1	1

Table 14 Feature ranking per algorithm

All three algorithms ranked the feature *weighted shortest distance* as main classification feature. It is interesting to notice that SVM is the only algorithm ranking as second important feature *sum of papers.* The other two algorithms give the second importance to the *weighted sum of neighbors* feature, third to *weighted sum of secondary neighbors* feature and only in the last instance to the *sum of papers* is considered.

The comparable performance indications of the SVM and K-NN algorithms lead to the conclusion that the features *sum of papers* and *weighted sum of neighbors* have comparable weight. This is a realistic conclusion considering the high frequency of collaboration in this community: the average number of authors per paper of 2.40 with respect to the number of publications per author 3.67. Both the SVM and K-NN algorithms and their proposed ranking schema suit this problem. The K-NN algorithm presents though a considerable less complexity from the point of view of implementation.

6.3 Conclusions

The problem of link prediction can be effectively handled by various modeling approaches. The suggested approach in this study is modeling the link prediction as a classification problem. The core of any classification algorithm is the defined set of features. This needs to suit the dataset and the type of problem analyzed.

For the link prediction problem there can be considered three categories of features: *proximity features, aggregated features* and *topological features*. In this study a set of four features were proposed, three of aggregate nature: *sum of papers, weighted sum of neighbors, weighted sum of secondary neighbors* and one of topological nature: *weighted shortest distance.*

The study investigated the highest prediction accuracy obtained with three different classification algorithms: *Support Vector Machine RBF, K-Nearest Neighbor* and *Naïve Bayes*. Even though the algorithm *Support Vector Machine RBF* leads in accuracy, the *K-Nearest Neighbor* algorithm, which is one of the most simple machine learning algorithms, proves a comparable accuracy and performance results. This is an important conclusion when considering that the modeling efforts required by the *K-NN* algorithm are considerable less than in the case of the *SVM RBF*.

Looking then at the feature ranking of the used algorithms, *weighted shortest distance* was chosen, in unanimity, the main feature. The two aggregated features: *sum of papers* and *weighted sum of neighbor,* have been ranked second by the best performing two algorithms. As the algorithm performance very similar, proposed feature ranking by the SVM and K-NN algorithms resulted in similar performance, an interesting extension of the study could be the performance of these algorithms with fixed feature ranking.

The goal of the proposed co-authorship study was a link prediction exercise, using different models and techniques. Though this problem has been prior studies, the used set of features for the co-authorship problem is unique. This could be considered as modeling alternative for the link prediction problem in contexts similar to the co-authorship problem. For best performance, small or medium datasets should be considered.

A common weakness in the link prediction studies, is the fact that social structures and their evolutions are studied separately. Therefore, some of the major interests in the domain are the link prediction problem in dynamic social networks and the knowledge exchange between heterogeneous social networks.

7. Conclusions

Social networks are a popular way to model the individual interaction within an organized group or community. Such social structure can be visualized as a network or graph, where an actor represents a group member and a link represents the form of association between two members of the group. Social Network Analytics combines the concept of the *sociogram* with elements of *graph theory* to analyze patterns of interaction among the group members, allowing quantitative comparisons between different network structures.

Due to the recent globalization of the commercial environment and the impact of the new technologies, the analysis of social networks represents a major interest**.** This rather new area of research grew out of social and exact sciences, computers supporting today modeling and complex mathematical calculations, previously impossible. The analysis of social networks is driven by business and social interests, combining various academic fields.

The current paper introduced the fundamental concepts and metrics in Social Network Analytics and proposed a set of mathematical models that can be applied for the problem of link prediction. Two case studies place in practice some of the presented concepts. The first study treats the problem of influential behavior in a *Bernard & Killworth Fraternity (BKFRAT)* social structure by measurements of sources and distribution of centrality and power in the network. The second study proposes a comparison of accuracy of three learning algorithms: *Support Vector Machine, K-Nearest Neighbor* and *Naïve Bayes* for the link prediction problem in the DBLP co-authorship community.

The question of how structural position confers influence or power in social networks, remains a topic of interest in the SNA research. Understanding how an entities (company, person, product) interact within virtual communities is a fundamental aspect for professional and social successes. Today a considerable number of social network analysis software packages are available, offering various tools for visualization, analysis and analytics. Considering the type of data and analysis, certain software packages might be more suitable than others.

Since new nodes and links are constantly added to a social structure, such entities are very dynamics. Understanding the behavioral concepts and the dynamics that drives the evolution of social networks is an important but a rather complex problem due to a large number of variable parameters.

Link prediction is a measure of social proximity between two individuals in a community that can be used to optimize an objective function over the entire social network. The link prediction problem implies modeling the way an information, a trend, a piece of knowledge etc. propagates via a social network. Such knowledge supports the development of tools for detection of hidden, missing or potential new links within a group. These type of problems are critical in many domains: security and criminal investigation, biology, marketing and sales, CRM, knowledge management systems and so on. A common weakness in the link prediction studies, is the fact that social structures and their evolutions are studied separately. Therefore, some of the major interests in the domain are the link prediction problem in dynamic social networks and the knowledge exchange between heterogeneous social networks.

Social networking provides clear advancements in communication and self expression. Businesses uses social networking to promote products, concepts and services. But if not understood and managed properly, social networking could cost the reputation of business and individuals.

References

1. Asur, S., Parthasarathy, S., and Ucar, D. *An event based framework for characterizing the evolutionary behavior of interaction graphs.* In KDD, pp. 913–921, 2007.

2. Bilenko, M. and Mooney, R. J. *Adaptive duplicate detection using learnable string similarity measures*. In KDD, pp. 39–48, 2003.

3. Bilgic, M., Namata, G., and Getoor, L. *Combining collective classification and link prediction.* In ICDM Workshops, pp. 381–386, 2007.

4. Borgatti, S. P., and Everett, M. G. *A Graph-theoretic perspective on centrality.* In Social Networks 28: 466-484, 2006.

5. Brandes, U., and Erlebach, T. *Network analysis : methodological foundations.* In Springer, Lecture notes in computer science, pp 471. Springer, 2005.

6. Breiger, R. L. *The analysis of social networks. In: Hardy MA, Bryman A Handbook of Data Analysis.* In Sage, London, pp 505–526, 2004.

7. Carrington, P. J., Scott, J., and Wasserman, S. *Models and Methods in Social Network Analysis.* In Cambridge University Press, Cambridge, 2005.

8. Chu, W., Sindhwani, V., Ghahramani, Z., and Keerthi, S. S. *Relational learning with gaussian processes.* In NIPS, pp 289–296, 2006.

9. Clauset, A., Moore, C., and Newman, M. E. J. *Hierarchical structure and the prediction of missing links in networks*. PMID 453(7191):98-101, 2008.

10. Degenne, A., and Forsé, M. *Introducing social networks.* In SAGE, London, 1999.

11. Easley, D., Kleinberg, J. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. In Cambridge University Press, 2010.

12. Frakes, W. B., and Baeza-Yates, R. *Information Retrieval: Data Structures and Algorithms*. Prentice Hall PTR, June 1992.

13. Freeman, L. C. *Centrality in social networks conceptual clarification.* In Social Networks 1: 215-239, 1978.

14. Friedkin, N. E. *Theoretical Foundations for Centrality Measures.* The American Journal of Sociology 96: 1478-1504, 1991.

15. Friedman, N., Getoor, L., Koller, D., and Pfeffer, A. *Learning probabilistic relational models.* In IJCAI, pp. 1300–1309, 1999.

16. Gelman, A., Carlin, J. B., Stern, H., and Rubin, D. B. *Bayesian Data Analysis, Second Edition*. Chapman & Hall/CRC, July 2003.

17. Getoor, L. *Learning Statistical Models from Relational Data.* PhD thesis, Standford, 2001.

18. Getoor, L., Friedman, N., Koller, D., and Taskar, B. *Probabilistic models of relational structure.* In Proc. ICML, 2001.

19. Getoor, L., and Diehl, C. P. *Link mining: a survey.* In SIGKDD Explorations, 7(2):3–12, 2005.

20. Getoor, L., and Taskar, B. *Introduction to Statistical Relational Learning (Adaptive Computation and Machine Learning)*. The MIT Press, August 2007.

21. Hanneman, R. A., and Riddle, M. *Introduction to Social Network Methods.* Publisher: University of California, Riverside, 2005.

22. Hasan, M. A., Chaoji, V., Salem, S., and Zaki, M. *Link prediction using supervised learning.* In LinkKDD, 2005.

23. Heckerman, D., Chickering, D. M., Meek, C., Rounthwaite, R., and Kadie, C. M. *Dependency networks for inference, collaborative filtering, and data visualization.* Journal of Machine Learning Research, 1:49–75, 2000.

24. Jin, E. M., Girvan, M., and Newman, M. E. J. *Structure of growing social networks.* Physical Review E, 64(4):046132, 2001.

25. Joachims, T. *Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms.* Kluwer Academic Publishers, 2003.

26. Kashima, H., and Abe, N. *A Parameterized Probabilistic Model of Network Evolution for Supervised Link Prediction*, Sixth IEEE International Conference on Data Mining, pp.340-349, 2006.

27. Kashima, H., Kato, T., Yamanishi, Y., Sugiyama, M., and Tsuda, K. *Link Propagation: A Fast Semi-supervised Learning Algorithm for Link Prediction.* In Proceedings of SDM, pp.1099~1110, 2009.

28. Lawrence, S., Giles, C. L., and Bollacker, K. D. *Autonomous citation matching.* In Agents, pp. 392–393, 1999.

29. Leskovec, J., Kleinberg, J. M., and Faloutsos, C. *Graphs over time: densification laws, shrinking diameters and possible explanations.* In KDD, pp 177–187, 2005.

30. Li, M., Chen, X., Li, X., Ma, B., and Vitanyi, P. M. B. *The similarity metric*. In IEEE Transactions on Information Theory, 50(12):3250–3264, 2004.

31. Liben-Nowell, D., and Kleinberg, J. M. *The link-prediction problem for social networks.* In JASIST, 58(7):1019–1031, 2007.

32. Lin, Y. R., Chi. Y., Zhu, S., Sundaram, H., and Tseng, B. L. *Facetnet: a framework for analyzing communities and their evolutions in dynamic networks.* In WWW, pp. 685–694, 2008.

33. Neville, J., and Jensen, D. *Relational Dependency Networks.* In Journal of Machine Learning Research, 8:653–692, 2007.

34. Newman, M. E. J. *Clustering and preferential attachment in growing networks,* PMID:11497639, 2001.

35. Newman, M. E. J. *The structure and function of complex networks.* SIAM Review, 45:167–256, 2003.

36. Newman, M. E. J. *A measure of betweenness centrality based on random walks.* In Social Networks 27: 39-54, 2005.

37. Oyama, S., and Manning, C., D. *Using Feature Conjunctions Across Examples for Learning Pairwise Classifiers*. In ECML, pp.322-333, 2004.

38. Popescul, R., and Ungar, L. H. *Statistical relational learning for link prediction*. In IJCAI03 Workshop on Learning Statistical Models from Relational Data, 2003.

39. Romero, D., Meeder, B., Barash, V., and Kleinberg, J. *Maintaining Ties on Social Media Sites: The Competing Effects of Balance, Exchange, and Betweenness.* In Proc. 5th International AAAI Conference on Weblogs and Social Media, 2011.

40. Schlkopf, B., and Smola, A. J. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond (Adaptive Computation and Machine Learning)*. The MIT Press, 2001.

41. Shawe-Taylor, J., and Cristianini, N. *Kernel Methods for Pattern Analysis*. Cambridge University Press, June 2004.

42. Stephenson, K., and Zelen, M. *Rethinking centrality: Methods and examples.* In Social Networks 11: 1-37, 1989.

43. Sun, J., Tao, D., and Faloutsos, C. *Beyond streams and graphs: dynamic tensor analysis.* In KDD, pp. 374–383, 2006.

44. Taskar, B., Abbeel, P., and Koller, D*. Discriminative probabilistic models for relational data*. In Proc. UAI, pp. 485-492, 2002.

45. Taskar, B., Wong, M. F., Abbeel, P., and Koller, D. *Link prediction in relational data.* In NIPS, 2003.

46. Tizghadam, A., and Leon-Garcia, A. *Betweenness centrality and resistance distance in communication networks*. In IEEE, ISSN: 0890-8044, pp: 10 - 16 , 2010.

47. Tresp, V., and Yu K. *An introduction to nonparametric hierarchical bayesian modelling with a focus on multi-agent learning.* In European Summer School on Multi-AgentControl, pp. 290–312, 2003.

48. Wasserman, S, and Faust, K. *Social Network Analysis: Methods and Applications.* Cambridge University Press, 1994.

49. White, D. R., and Borgatti, S. P. *Betweenness centrality measures for directed graphs.* In Social Networks 16: 335-346, 1994.

50. Xiang, E. W. *A survey on link prediction models for social network data.* Student paper, Hong Kong University of Science and Technology, 2008.

51. Xu, Z., Tresp, V., Yu, K., Yu, S., and Kriegel, H. P. *Dirichlet enhanced relational learning.* In ICML, pp. 1004–1011, 2005.

52. Xu, Z., Tresp, V., Yu, K., and Kriegel, H. P. *Infinite hidden relational models.* In UAI, 2006.

53. Xu, Z., Tresp, V., Yu, S., and Yu, K. *Nonparametric relational learning for social network analysis.* In 2nd ACM Workshop on Social Network Mining and Analysis (SNA-KDD), 2008.

54. Yu, K., Chu, W., Yu, S., Tresp, V., and Xu, Z. *Stochastic relational models for discriminative link prediction.* In NIPS, pp. 1553–1560, 2006.

55. Yu, K., and Chu, W. *Gaussian process models for link analysis and transfer learning.* In NIPS, pp. 1657–1664, 2007.