UHASSELT

KNOWLEDGE IN ACTION

Faculteit Bedrijfseconomische Wetenschappen

master in de toegepaste economische wetenschappen: handelsingenieur in de beleidsinformatica

Masterthesis

Extractie van een sociaal netwerk uit een event log

Daan Roosen

Scriptie ingediend tot het behalen van de graad van master in de toegepaste economische wetenschappen: handelsingenieur in de beleidsinformatica

PROMOTOR :

Prof. dr. Mieke JANS

2018
2019

UHASSELT

KNOWLEDGE IN ACTION

www.uhasselt.be
Universiteit Hasselt
Campus Hasselt:
Martelarenlaan 42 | 3500 Hasselt
Campus Diepenbeek:
Agoralaan Gebouw D | 3590 Diepenbeek

UHASSELT

KNOWLEDGE IN ACTION

Faculteit Bedrijfseconomische Wetenschappen

master in de toegepaste economische wetenschappen: handelsingenieur in de beleidsinformatica

Masterthesis

Extractie van een sociaal netwerk uit een event log

Daan Roosen

Scriptie ingediend tot het behalen van de graad van master in de toegepaste economische wetenschappen: handelsingenieur in de beleidsinformatica

PROMOTOR :

Prof. dr. Mieke JANS

Social Networks from Event Logs: Comparing Existing Methods through Social Network Analysis

Daan Roosen, Mathijs Creemers and Mieke Jans

Hasselt University, Martelarenlaan 42, 3500 Hasselt, Belgium

Abstract. Event logs mostly contain information about the employees in a business process, which can potentially be valuable for social network analysis. At the moment, it is not clear how to extract a social network from an event log. Some methods for different purposes have been proposed. However, it is not known to what extent the methods differ. In this work, the methods are implemented and social networks are extracted from real-life event logs. The methods are compared conceptually, on network-level and through social network analysis. These analyses show that the methods are fundamentally different. Consequently, the results of a social network analysis will depend on the choice of extraction method. To conclude, some situations are presented in which some of the extraction methods do not perform well.

Keywords: Process mining · Social network analysis · Community analysis.

1 Introduction

Social network analysis (SNA) comprises a number of techniques for analyzing social networks. A social network is formed by social entities and the relationships among them. Social network analysis techniques examine the structure of those relationships [9]. The data, needed for these studies, mostly comes from surveys [37] where people are asked about their social relations within some context. Data gathered from electronic sources, such as email traffic and chatboxes, is also used quite often. [34], for example, used messages, posted in a chatbox, to conduct a social network analysis study. These approaches are less useful for the analysis of social networks around business processes, because they are based on unstructured information [3]. When analyzing email correspondence, for example, it is difficult to distinguish which information belongs to the process and which messages are less relevant [3]. Therefore, more structured data is recommended. Companies nowadays use information systems, such as enterprise resource planning (ERP), customer relationship management (CRM), workflow management systems, etc. to support their processes. Typically, these systems log information about transactions in log files, also known as *event logs*.

Table 1: An event log.

Case ID	Activity ID	Resource	Timestamp
1	A	Susan	09-01-2019 14:32
1	B	Mike	10-01-2019 11:45
1	C	Susan	10-01-2019 13:50
1	D	Charlotte	10-01-2019 18:32
2	A	Susan	11-01-2019 17:16
2	B	Mike	12-01-2019 12:13
2	C	Charlotte	13-01-2019 10:11
2	D	Anna	14-01-2019 09:56

The idea of process mining is to discover, monitor and improve real processes by extracting knowledge from event logs readily available in today's (information) systems [1]. Process mining includes among others process discovery, conformance checking, case prediction, social network/organizational mining, etc. For each instance of the process, event logs capture all events that take place. Moreover, numerous attributes can be included in an event log, for example timestamps, resources, etc. Table 1 shows a simple example of an event log.

When people are involved in the process, the event log will typically contain a *resource* or *originator* attribute, which represents who started or executed the activity [4]. This information can be used to extract a social network from the event log. However, the research concerning this topic is quite scarce. Only a handful of papers are available at the time of writing. The most important contribution can be found in [3, 4]. In their work, the authors propose techniques for extracting social networks from event logs, namely handover of work, subcontracting, joint cases, joint activities and special event types. For each of these categories, they have established metrics for computing the strength of the relations between individuals. [51] built further on these techniques by using them in some methods for organizational mining and [23] present a technique for community analysis in event logs. In Section 5, the existing techniques will be discussed in more detail.

Although all presented metrics in [3] - to which we refer as extraction methods - relate to the concept of social networks, the extraction methods differ greatly on a semantical basis. Handover of work, for example, searches for transfers of work from one resource to another. On the other hand, joint cases counts the number of times resources have worked together on cases. It is clear that the definitions of these metrics are very different. This means that each of these extraction methods might produce a unique social network with different weights. In fact, without taking the choice of parameter settings into account, seven social networks can be calculated from the same event log. At the moment, it is not clear whether these social networks are similar or fundamentally different. In the latter case, validation issues can arise because the results of social network analysis might depend on which extraction method is used. As a consequence,

one might have multiple different outcomes for a social network analysis from the same event log. Which results are then the most truthful? Currently, this is difficult to say, because no information is available to decide which technique is suitable in a particular situation.

The aim of this paper is to investigate the (dis)similarities between the existing extraction methods. The research question we address is the following: does the choice of social network extraction method have an impact on the later results of social network analysis? Within SNA, no use cases are available that use event logs as input for SNA. We chose to focus on community analysis, which is one of the most popular applications of SNA. To find an answer to our research question, we investigate the subgroups that are represented in social networks from the same event log. In order to do so, we developed a tool which allows to extract social networks from event logs using the proposed methods [3]. We apply the seven methods on six event logs which results in 42 social networks. These socials networks are compared on some general characteristics, namely density and mutual edges, and on the embedded subgroups. The findings of our analysis indicate that the extraction methods not only yield different social networks, but also different insights.

In Section 2, we give an overview of related work that has investigated the application of social network analysis to event logs. Next, our methodology is described in Section 3. An introduction to SNA is given in Section 4. The extraction methods are discussed in more detail in Section 5. Section 6 explains how the experiments were conducted. The results of our experiments are shown in Section 7. To conclude, we summarize the results obtained in this work and present challenges for future research in sections 8 and 9, respectively.

2 Related work

As already stated, research concerning the combination of process mining and social network analysis is very scarce. Apart from [4], who proposed techniques for extracting social networks from event logs, two papers have covered this topic. The extraction methods will be discussed in more detail in section 5. [51] built further on the techniques, presented by [4], by using them in some methods for organizational mining. They distinguish three types of organizational mining, namely organizational model mining, social network analysis and information flows between organizational entities. Organizational model mining is aimed at finding groups of originators that have similar characteristics, which are called organizational entities. Task-based team and case-based team are both categories of organizational entities [51], who are, respectively, based on the joint activities and joint cases metrics. For the task-based team, [51] also propose a method using Agglomerative Hierarchical Clustering. At last, [23] present a technique for community analysis in event logs. Whereas [51] calculated subgroups based on similar tasks, [23] use a working together metric for extracting the social network. They argue that the discovery of user communities in a network, based on actual interactions between individuals, can also hold valuable information.

Thereafter, they apply a community analysis algorithm based on modularity to find subgroups. The aforementioned research articles are, to the best of our knowledge, the only literature that is available concerning social networks in event logs.

3 Methodology

In order to investigate the (dis)similarities between the extraction methods, we applied the different extraction methods to real-life event logs. We used six different publicly available datasets, wherein multiple resources are involved. There are four extraction methods, namely handover of work, subcontracting, joint cases and joint activities. Additionally, joint activities comprises four submethods. As a result, seven social networks were retrieved from each event log. These networks form the basis of our research. To compare the extracted social networks in a structural way, we examined the use of established SNA techniques. Therefore, we identified the prerequisites social networks need to adhere to in order to apply these techniques.

The remainder of our work can be divided in three parts. First of all, some background information on SNA is given, together with the discussion of the prerequisites. Secondly, an overview is given of the existing extraction methods, along with a conceptual analysis. At last, we performed an empirical comparative analysis of the seven extraction methods, applied to six event logs. Two aspects were analyzed. Firstly, the topology of the networks was compared to find a preliminary insight in the similarity of the extraction methods. Therefore, some general characteristics and measures were used to compare the social networks. Furthermore, the number of mutual edges in the networks was analyzed. The results of this analysis are reported in Section 7.1. Secondly, we applied community analysis techniques to the networks. The resulting community division were compared with each other. Our goal was to find out to what extent the community partitions of different social networks from the same event log resembled. Based on this comparison, summarized in Section 7.2, we can draw conclusions about the similarity of the social structure embedded in the social networks.

4 Social network analysis and prerequisites

In this section, an overview of social network analysis is given. Furthermore, the prerequisites of social networks for community analysis are introduced.

4.1 Social network analysis

Social network analysis is a technique for examining the structure of relationships among social entities [9]. Individuals, companies and organizations are examples of social entities. The relationships mostly have a social character. However,

they can also represent interactions, flow of information, flow of money, similarities, etc. Two main forms of social network analysis can be distinguished: the ego-network analysis, and the global network analysis [39]. An ego-network is a network of one entity with all related neighbors, whereas a global network analysis is aimed at finding all relations between all entities.

A social network consists of actors and links (edges), where actors represent social entities and links correspond to the relations among them. Networks can be undirected or directed. An undirected network only has relations where no distinction can be made between the sender and receiver of the relation [9]. On the other side, networks of which all relations clearly have a sender and a receiver, are said to be directed. Examples include relations representing flow of knowledge or messages between individuals. Another subdivision can be made based on the links. Numbers can be assigned to the links in a network in order to include information about the frequency or strength of the relations. In that case, the network is called weighted. An unweighted network does not have this information.

Social networks can be represented in a variety of ways, of which adjacency matrices and graph theoretical notation are the most common [9]. An adjacency matrix is a matrix with n rows and n columns (n is the number of actors in the network). In case of an unweighted network, a cell can only have two possible values: 1 if there exists a relation between the two actors and 0 if there is no relation. In a weighted network, the values in the cells can be, for example, the number of messages one actor has sent to the other. Graph notation is similar, but more visual. A graph consist of nodes that are connected by edges. Each node corresponds to a specific actor and each edge depicts a relation between the nodes it connects. The strength or frequency of the relation can be placed on the edge.

Density and centrality are two important measures for the analysis of a social network. The density is an indicator for the general level of connectedness of the graph [39]. If every node is connected to all other nodes in the network, density will be 1. Centrality indicates the importance of a node in a graph relative to the other nodes. Degree, closeness and betweenness are the most used metrics for centrality. The degree of a node is the number of direct connections it has [26]. In-degree is the number of incoming arcs, whereas out-degree is the number of outgoing relations. Closeness measures the mean geodesic distance of one node to all the other nodes [26]. Low closeness indicates that the node has a central position. Betweenness measures the extent to which a node lies on the shortest paths between nodes [26]. For more information on these metrics, we refer to [25, 26].

Social network analysis consists of a variety of techniques, each for a specific purpose. We distinguish seven subgroups within these methods. First, interaction analysis aims at analyzing the relations between individuals, using the aforementioned measures [34]. Second, attribute-based analysis examines social networks with attributes attached to the social entities or the relationships [12]. Third, the knowledge flows in an organization can be represented in a special type of

social network, namely one wherein the relationships represent the transfer of knowledge flows [15]. Fourth, routing problems are concerned with the determination of the most efficient route between two points [17]. Fifth, clustering or community analysis searches for subgroups in the network. A subgroup can be defined as a group of highly connected individuals who are loosely connected with the rest of the network [47]. Sixth, hierarchical analysis tries to find the social hierarchy in a network [14]. Seventh, optimal information diffusion tries to find the most optimal way to spread information across the actors [30].

In the remainder of this section, some characteristics, that social network data needs to have, are presented. The goal is to verify whether event logs meet the prerequisites to be used as input for a social network analysis.

4.2 Prerequisites

Social network analysis techniques require data to have certain characteristics. Based on the inputs that the different social network analysis algorithms need, we identified five possible prerequisites, who will be discussed in this section. It is important to mention that not all of these prerequisites need to be fulfilled for each application. Everything depends on the situation and the goal of the study.

Actor The first prerequisite is the presence of a social entity in the data, also referred to as an actor. As already discussed, many different things can serve as a social entity, an individual being the most intuitive. Furthermore, companies, organizations and academic articles are all examples of social entities. Actually, anything among which a relationship can exist. Additionally, it is possible to have more than one type of actor. For instance, a network of football transfers would contain clubs and players.

It follows from the definition [9] of social network analysis that an actor needs to be included in the data, otherwise a social network analysis is impossible. Of course, this prerequisite might seem trivial, but we need it to assess if data is suited for a social network analysis.

Link The second prerequisite - that needs to be fulfilled - is the existence of relationships (links) between actors in the data. Not only social relations but also associations, similarities and flow of information are examples of those relationships. In theory, a collection of actors without links can be represented by a social network with zero connectedness, i.e., a density value of 0. Although this is possible, an analysis of the problem is only meaningful when information about the relationships among the actors is available. Indeed, the definition of social network [9] analysis stipulates that it is the analysis of the relations among actors.

Weight Weighted links are the third prerequisite. Weights increase the information value of a social network because they indicate the strength of the relations

or associations. Therefore, they facilitate a more detailed study of the relations among the actors. Especially when the goal of the study is to investigate behaviour in the network on actor-level, weights can be useful. In the next section, we will show that not all applications of social network analysis require weights.

Directed links The type of links can also be important for the analysis. While undirected links mostly represent associations or similarities, directed links can be more informative. Because the sender can be distinguished from the receiver, directed networks have more options for an in-depth analysis. For example, knowledge flows, email traffic and telephone conversations can be included in such a network. Furthermore, the type of links can constitute certain difficulties. In the case of a directed network, two links are possible between each pair of entities. Actor A can have an outgoing and an incoming relation with actor B. Therefore, the number of possible links is two times greater in comparison to an undirected network with the same number of entities. As we will see in the next section, some algorithms can not handle directed networks. Consequently, the directed data needs to be transformed into undirected data, if possible.

Attributes Some methods need additional information about actors. In those cases, the data must contain the necessary attributes. Of course, this prerequisite is very specific and will be different for each problem. Examples of attributes include: age, height, timestamps, etc.

4.3 Prerequisites community analysis

Community analysis is the most frequently used technique within social network analysis. In this section, we give an overview of the prerequisites for community analysis. Therefore, the existing methods and algorithms were analyzed in order to discover the inputs they need.

Clustering or community analysis is a wide domain within network analysis. A variety of methods and algorithms has been presented. In this paper, we examined 22 common approaches, mainly based on overviews presented by [18, 29]. For each of those algorithms, we analyzed the characteristics of the social networks that were used for the community analysis. In Table 2 the results are summarized. As expected, actors and links are required by each algorithm.

Table 2: Prerequisites of community detection algorithms[1].

Paper	Actor	Links	Weights	Directed links
1 Girvan and Newman [27]	X	X		
2 Palla et al. [41]	X	X		
3 Bagrow and Bollt [5]	X	X		
4 Radicchi et al. [44]	X	X		
5 Clauset et al. [13]	X	X		
6 Capocci et al. [10]	X	X		O
7 De Meo et al. [20]	X	X		O
8 Chen and Saad [11]	X	X		O
9 Reichardt and Bornholdt [48]	X	X	O	
10 Duch and Arenas [22]	X	X	O	
11 Riedy et al. [49]	X	X	O	
12 Wu and Huberman [55]	X	X	O	
13 Fortunato et al. [24]	X	X	O	O
14 Raghavan et al. [45]	X	X	O	O
15 Guimerà and Amaral [28]	X	X	O	O
16 Donetti and Muñoz[21]	X	X	X	
17 Pons and Latapy [42]	X	X	X	
18 Zhou and Lipowsky [56]	X	X	X	
19 Ovelgönne and Geyer-Schulz [40]	X	X	X	
20 Jiang and Singh [33]	X	X	X	
21 Blondel et al. [6]	X	X	X	O
22 Jain and Dubes [32]	X	X	X	O

The algorithms are ordered from least demanding to most demanding, with the least demanding algorithms at the top. Actually, the least demanding algorithms are also the least flexible. The first five algorithms, for example, only require the presence of actors and links, which is the most elementary form of a social network. They only need a simple network as input. However, they can not handle weights and directed links, which makes them less flexible. Although, it is, in some cases, possible to transform a weighted directed network into an unweighted undirected network, it requires an extra step in the analysis. Algorithms 6, 7 and 8 are the only ones that can not handle weights, but are able to work with directed links. The remaining algorithms, 9 till 22, can all manage weighted social networks. 7 of them have the option to include weights, while the rest, 16 till 22, really needs them. Furthermore, we can see that none of the algorithms require directed links. It is always optional.

[1] X: required, O: optional, for weights and directed links an empty cell means that the algorithm can not handle them

In summary, Table 2 shows us that the minimal prerequisites to do a community analysis of a social network, are the presence of both actors and links. As we will see in Section 5, the social networks retrieved from event logs always have actors and links. Additionally, weights and directed links are also included. This means that, in theory, it should be possible to apply community detection algorithms on these social networks. This does not mean that the algorithms will produce a good partitioning of the network since this depends on the data.

5 Extraction methods and conceptual analysis

In this section, an overview is given of the existing process mining techniques for extracting social networks from event logs. Research on this topic is very scarce. Only one extensive work is available [3]. They present five types of extraction methods, namely handover of work, subcontracting, joint cases, joint activities and special event types. All those methods are ways of computing the weights on the relations among actors in the social network. The authors [3] presented the methods for specific purposes. However, it is not known how the methods relate to one another. In the remainder of this section, we will explain the methods in a concise way. For a detailed, formal description, we refer to [3].

The first two types of extraction methods monitor for individual cases how work moves among performers [3]. A handover of work is present within a case when there are two subsequent activities performed by resource i and resource j, where i completed the first activity and j the second one. Subcontracting occurs when an activity performed by resource j is nested between two activities performed by resource i. In a sense, i subcontracted a piece of work to j. Thus, subcontracting is more strict than handover of work, because it requires the work to come back to the original resource. Handover of work does not. Hence, all edges present in a subcontracting network, are by definition also contained in a handover of work network.

For both extraction methods, three kinds of refinements can be applied. The first one takes into account the length of the handover. Not only direct succession, but also indirect succession is considered. When three subsequent activities are performed by resources i, j and k, an indirect succession exists between resource i and k. However, an indirect succession will not get the same importance as a direct succession. Therefore, the *causality fall factor* β^n is introduced. β ranges between 0 and 1. For handover of work, the importance of a succession is denoted by β^{n-1}, whereas for subcontracting the importance equals β^{n-2}. A value of 0.5 means that the importance of a succession of degree n is twice the importance of a succession of degree $n+1$. Note that the choice of β does not have an impact on the topology of the social networks. It only effects the values of the weights. Conversely, the structure does depend on the value of n.

The second refinement is whether or not to ignore multiple transfers within one case. It is possible that more than one succession between resource i and resource j occurs within one case. When the choice is made to ignore multiple transfers, a succession will be counted only once per case.

Thirdly, we can consider all transfers of work or we can only consider those where there is a causal dependency. To verify if there is causal dependency, the process model of the event log is needed. For more information on these refinements, we refer to [3]. In all cases, handover of work and subcontracting produce directed social networks.

The third extraction method is more general than the previous ones. Joint cases counts the times resource i and j worked on the same case and divides it by the number of cases resource i appeared in. The division is needed to take into account the relative importance. For example, if resource i and j worked together on three cases and resource i was active in only three cases, the weight should be higher than when resource i had worked on thirty cases. Thus, the weight for joint cases from i to j will be different from the weight from j to i. Therefore, the resulting social network will be directed. Handover of work and subcontracting also require the resources to appear in the same case. However, some additional constraints also need to be satisfied. As a result, a joint cases network will always contain at least the same number of edges as a handover of work or subcontracting network. Any edge that is present in the latter two, will also be in a joint cases network.

Fourth, joint activities takes a completely different approach. It makes the assumption that resources, who perform similar tasks, have a stronger connection than people doing different things. The extraction method uses a profile approach, meaning that a profile is made for each resource based on how often the resource conducts specific activities. The results are summarized in a resource by activity matrix. The weight from resource i to resource j is measured by the distance between them. Therefore, [3] propose four different distance metrics, namely Minkowksi distance ($n = 1$ or $n = 2$), Hamming distance and Pearson's correlation coefficient. Hence, four social networks can be extracted by using joint activities. In these networks, the edges represent the extent to which resources resemble each other. Therefore, only one relationship exists between two resources. Consequently, these social networks are undirected, which makes them unique.

The last category of extraction methods are those based on special event types. Examples of event types include *assign, reassign, start, schedule, end, withdraw, suspend, resume, ...* A *reassign* event, for example, gives an indication of a handover of work. Therefore, these special event types can be interesting for SNA. The methods based on special event types are not applicable to all datasets, because not all event logs contain these events. Therefore, this technique was not considered in our analysis. In the remainder of the paper, we analyze handover of work, subcontracting, joint cases and the four joint activities networks in more detail.

6 Experiment design

The extraction methods are integrated in the ProM framework [2], which is a tool for process mining. Although it allows to calculate the social networks from

event logs, the data can not be extracted from the tool. Therefore, the techniques were implemented in Python. To validate our implementations, the running example of [3] was used. By replicating this analysis, we obtained identical results. Additionally, we verified - for some of the data sets used in this study - randomly some weights by calculating the social networks in ProM and comparing them with our results. Exactly the same values were found for all techniques, except for joint activities with Pearson and joint activities with Hamming. This is contradictory because the running example could be replicated precisely. Therefore, we used ProM to calculate joint activities Pearson and joint activities Hamming for the event log of the running example, which gave conflicting results in comparison with the results in [3]. Accordingly, the implementation, validated using the results in [3], was retained.

In Section 3, we already mentioned that we used six event logs, which are listed in Table 3. From each of these event logs, we extracted 7 social networks. β was set to 0.5 and n to 5 for the handover of work and subcontracting metrics. With n set to 5, the successions, that may be missed, have an importance that is 16 (0.5^0 vs. 0.5^4) times smaller than the importance of direct succession for handover of work and 8 (0.5^0 vs. 0.5^3) times smaller for subcontracting. Hence, we assume them to be negligible.

Table 3: Research instruments.

BPI 2012 [52]	GN-method [38]	
BPI 2017 [53]	Fastgreedy [13]	nmi [18]
EPA [7]	Label propagation [45]	vi [46]
Production data [35]	Louvain [6]	Rand [46]
Receipt phase EPA process [8]	Walktrap [42]	Adj. Rand [31]
Sepsis cases [36]	Infomap [50]	Split-join [54]

(a) Event logs used. (b) Community detection. (c) Comparison metrics.

The SNA part of the experiment is conducted in R [43], by using the functions provided by the iGraph package [16]. In this package, implementations can be found of the most common SNA techniques. However, neither of the community analysis algorithms in iGraph uses the direction of the edges. Therefore, we chose to transform the extracted networks to undirected networks by replacing multiple edges between resources by one edge and summing the weights. To reduce the individual effects of the algorithms as much as possible, we applied 6 community detection methods to each social network. They are summarized in Table 3. Thus, 42 community divisions were obtained per event log. A community division denotes for each resource to which group it belongs. The number of subgroups is not fixed. It can vary across the social networks and the community detection algorithms. Furthermore, a resource can only be part of one subgroup.

In order to investigate the similarity of the different social networks, we compared the community divisions for each pair of community detection method and event log. Because the dataset and the community detection algorithm are the same, the variation in these results can only be caused by the choice of extracting method. Hence, this allows us to draw conclusions about the similarity of the social networks. In order to compare the community divisions, we used the metrics in Table 3. These metrics compare exactly how the individual resources were divided in subgroups. Additionally, the composition of the subgroups is also taken into account. The normalized mutual information (nmi) measure ranges between 0 and 1. Identical partitions are indicated by a value of 1, whereas a measure of 0 is found when the partitions are completely different. The Rand index measures the percentage of pairs of resources the two community divisions agree on. A derivative of the Rand index is the adjusted Rand index, which is corrected for chance. In contrast to the Rand index, this index can also be negative. This happens when the index is smaller than the expected value. Split-join distance represents the number of necessary moves to go from one community division to the other. At last, variation of information (vi) gives an indication of the variation in the community division. For perfectly similar divisions, this metric will be 0.

To get an overall result across all datasets and community detection algorithms, we averaged the results per comparison measure into one table. Therefore, we had to do an adjustment to the split-join distance. We divided it, for each eventlog, by the total number of resources in the network to make it relative. Thus, it currently represents the percentage of resources that need to be moved to go from one community division to the other.

7 Results

The results section comprises three subsections, namely comparative analysis, community analysis and encountered shortcomings of the extraction methods.

7.1 Comparative analysis social networks

In this section, the results of the comparative analysis of the extracted networks are summarized. Therefore, the structure of the social networks was compared on some general characteristics, namely density and mutual edges.

In Table 4, we can see the average density of the social networks across the different event logs. As already stated, density measures the connectedness in a social network. These values indicate big differences among the social networks. Subcontracting, for example, has a density of 0.27, which means that 27% of all possible edges are included in the network. On the other hand, the joint activities networks are, except for Hamming, all fully connected. The high densities are most likely caused by the definition of the distance metrics. If the resources have very few or none activities in common, the distance will be very big. Actually, this should be a very weak relationship because the resources differ a lot. As a

result, an edge is created for almost every pair of resources, which explains the high connectedness.

Table 4: Average density per extraction method [2].

	SC	HOW	JC	JAP	JAH	JAM1	JAM2
Average density	0.27	0.52	0.57	0.91	1	1	1

Based on these differences, we suspect that these networks do not hold similar information. Actually, we can say that the networks in Table 4 are ordered from most strict to least strict. A handover of work network consists on average of more edges than a subcontracting network.

Table 5: Common edges across the networks averaged over the event logs.

	HOW	SC	JC	JAP	JAH	JAM1	JAM2
HOW	1	0.41	1	1	1	1	1
SC	1	1	1	1	1	1	1
JC	0.74	0.31	1	1	1	1	1
JAP	0.42	0.18	0.57	1	1	1	1
JAH	0.42	0.18	0.57	1	1	1	1
JAM1	0.42	0.18	0.57	1	1	1	1
JAM2	0.42	0.18	0.57	1	1	1	1

In order to further verify this, an analysis was done of the mutual edges across the networks by counting for each pair of networks the edges they have in common. Table 5 summarizes the results of this analysis. On average 41% of the edges present in the handover of work network were also present in the subcontracting network. Apart from the joint activities networks, the measures are very low, indicating big differences between the networks.

The results in this section give us some preliminary insights in the extraction methods. It appears that they do generate different social networks from the same event log. Although we suspect that the underlying social structure in these networks will also be different, we can not conclude this on these results alone. Therefore, community analysis techniques were applied to compare the underlying communities that reside in the extracted networks.

[2] HOW: handover of work, SC: subcontracting, JC: joint cases, JAP: joint activities Pearson, JAH: joint activities Hamming, JAM1: joint activities Minkowski 1, JAM2: joint activities Minkowski 2

7.2 Comparison community divisions

As mentioned before, for each event log, a community detection algorithm was applied on each of the 7 social networks. The resulting community divisions were then compared per event log with the metrics described in Table 3. We executed this procedure for six different community detection algorithms. All the results are averaged to get a general view of the (dis)similarities between the community divisions. This leads to one value per pair of extraction methods, which gives an indication of the extent to which the community divisions based on the methods resemble, on average.

Table 6 shows the results obtained by using the split-join distance. The best resemblance is found between joint activities Minkowski 1 and joint activities Minkowski 2. Since these weights are based on a similar distance metric, this result is quite logical. However, still 8% of the resources was allocated differently. In general, the mutual results of the different joint activities networks are very poor. The lowest value is 0.36, between joint activities Minkowski 1 and joint activities Hamming. Based on split-join distance, this means that the choice of distance metric has a very big impact on the social structure in the network. The other methods also seem to differ heavily. Handover of work and subcontracting have a value of 0.42, which is one of the lowest. The values between handover of work, subcontracting and joint cases range between 0.32 and 0.50, whereas the values between these three and the other networks vary from 0.48 until 0.78. Therefore, joint activities seems to be very different from the other ones.

Table 6: Comparison with split-join distance.

	HOW	SC	JC	JAP	JAH	JAM1	JAM2
HOW	0	0.44	0.42	0.40	0.74	0.57	0.57
SC		0	0.50	0.47	0.70	0.61	0.62
JC			0	0.46	0.77	0.64	0.63
JAP				0	0.75	0.59	0.60
JAH					0	0.37	0.37
JAM1						0	0.08
JAM2							0

As can be seen in Table 7, the normalized mutual information measure does not confirm all the results found by the split-join distance. Joint activities does not seem to differ more than average from the other networks, which was the case in Table 6. On the other hand, it does agree on the fact that the mutual differences between the joint activities networks are not to be neglected. Especially, joint activities Pearson has low values, ranging between 0.26 and 0.33. Because the other joint activities networks show reciprocally descent results, the nmi measure indicates that the Pearson network differs substantially from Hamming and Minkowski.

Table 7: Comparison with normalized mutual information measure.

	HOW	SC	JC	JAP	JAH	JAM1	JAM2
HOW	1	0.30	0.45	0.40	0.33	0.35	0.32
SC		1	0.24	0.19	0.38	0.31	0.30
JC			1	0.36	0.30	0.32	0.30
JAP				1	0.26	0.33	0.27
JAH					1	0.62	0.61
JAM1						1	0.87
JAM2							1

In Table 8, the results of the Rand index are summarized. This index is far more positive than the previous ones. Most values are above 0.50, which means that the community divisions agree on at least 50% of the resource pairs. Handover of work, subcontracting and joint cases resemble each other quite good, with values between 0.69 and 0.71. Furthermore, joint activities Pearson has again lower values than the other joint activities networks.

Table 8: Comparison with Rand index.

	HOW	SC	JC	JAP	JAH	JAM1	JAM2
HOW	1	0.69	0.71	0.68	0.52	0.62	0.61
SC		1	0.66	0.61	0.55	0.59	0.57
JC			1	0.66	0.49	0.57	0.57
JAP				1	0.45	0.57	0.55
JAH					1	0.72	0.72
JAM1						1	0.94
JAM2							1

The adjusted Rand index has, in comparison to the Rand index, a very negative view of the similarities between the social networks. Ignoring the mutual values of the joint activities networks, the index (0.36) between handover of work and joint cases is by far the highest one. Even values below 0.1 are noted, which means that the community divisions disagree on more than 90% of the resource pairs. In regard to the Pearson network, the adjusted Rand index confirms the findings of the previous metrics. Furthermore, the comparisons of the joint activities networks and the other networks denoted very few similarities, which was also indicated by the split-joint distance in Table 6.

Table 9: Comparison with adjusted Rand index.

	HOW	SC	JC	JAP	JAH	JAM1	JAM2
HOW	1	0.15	0.36	0.35	0.09	0.15	0.12
SC		1	0.09	0.11	0.03	0.03	0.02
JC			1	0.31	0.06	0.14	0.12
JAP				1	0.06	0.18	0.13
JAH					1	0.53	0.52
JAM1						1	0.84
JAM2							1

The last comparison metric, namely variation of information, shows similar results in Table 10. Again, handover of work, subcontracting and joint cases differ more from joint activities than from each other. In correspondence with the previous results, the best resemblance is found between handover of work and joint cases. Additionally, the Pearson network is once more very different from the other joint activities networks.

Table 10: Comparison with variation of information.

	HOW	SC	JC	JAP	JAH	JAM1	JAM2
HOW	0	1.20	0.98	1.02	2.46	1.54	1.55
SC		0	1.32	1.36	2.2	1.70	1.74
JC			0	1.11	2.57	1.78	1.78
JAP				0	2.59	1.71	1.78
JAH					0	1.34	1.35
JAM1						0	0.22
JAM2							0

This section was devoted to the comparison of the social structure in the different social networks. Some patterns were found in the results of various comparison metrics. At first, the most obvious finding is probably the fact that the joint activities networks differ a lot, depending on which distance metric is chosen. Especially, the Pearson networks were very different from Hamming and Minkowski (1 and 2). Although the choice of distance metric does not have a large impact on the edges in the resulting social network, it appears that the results of community analysis are still very much affected by it. Therefore, we can conclude that the weights on the edges must be very different. Secondly, there is some evidence for the distinction between handover of work, subcontracting and joint case on the one hand, and joint activities on the other hand. The

mutual differences between the first three were, in general, smaller than the dissimilarities with joint activities.

To conclude, we can say that the results - presented in this section - show big differences between the community divisions of the different social networks. As previously mentioned, the social networks from the various extraction methods can be the only explanation for these outcomes. Therefore, our research suggests that each of the extraction methods results in a unique social network, that holds a unique social structure. Consequently, the choice of extraction method will have a significant impact on the results in the further course of the analysis. This raises difficulties for researchers willing to perform a SNA on event log data. The validation of the results will be non-trivial because it is not known which extraction method is the most truthful.

7.3 Encountered shortcomings of the extraction methods

In our research, we faced some shortcomings of the extraction methods, which are presented in this section.

Type of event log The shape of the event log under study can have an influence on the choice of extraction method. There exists a situation in which not all methods are suitable, namely when the event log has the form of Table 11. In this event log, every case is executed by only one resource. Because handover of work, subcontracting and joint cases are case-based, they are not applicable on this type of event log. Handover of work and subcontracting search for successions within a case, which are not present in this event log. Joint cases counts the numbers of times resources have worked together on a case. Therefore, these methods will yield a fully unconnected network, i.e. all weights are zero. For an example of a real-life event log, we refer to the road traffic fine management process [19].

Table 11: Event log with 1 resource per case.

Case ID	Activity ID	Resource	Timestamp
1	A	Susan	09-01-2019 14:32
1	B	Susan	10-01-2019 11:45
1	C	Susan	10-01-2019 13:50
1	D	Susan	10-01-2019 18:32
2	A	Mike	11-01-2019 17:16
2	B	Mike	12-01-2019 12:13
2	C	Mike	13-01-2019 10:11
2	D	Mike	14-01-2019 09:56

Joint activities social networks are based on the activities resources execute. Individual cases are not taken into account by the algorithm. Hence, joint activities is the only extraction method that can handle this type of event log.

Joint activities The previous analysis has shown that joint activities extracts social networks that are very different from the other methods. Something that is most definitely caused by the definition of this metric. Whereas the other methods are mostly based on resources working together, joint activities measures the similarity of the resources. Furthermore, the interpretation of the results of SNA is also very different in comparison to the other networks. In our case, a community extracted from a joint activities network is a group of similar resources. Communities in handover of work, subcontracting and joint cases networks consist of resources that work frequently together.

Table 12: Analysis failures.

	SC	HOW	JC	JAP	JAM1	JAM2	JAH
No communities found	0	10	12	14	5	5	2
Separate groups	1	2	0	0	15	15	25
Total	36	36	36	36	36	36	36
% Failures	0.03	0.33	0.33	0.39	0.56	0.56	0.75

In our research, we came across some difficulties with the joint activities networks. It appeared that the community detection algorithms had issues finding subgroups in these networks. Repeatedly, community divisions of only one subgroup or community divisions with one subgroup per resource were obtained. Table 12 shows, for all extraction methods, an overview of the frequency of these failures. It appears that the joint activities networks are very prone to this issue. Especially, the Hamming distance scores very bad. In 75% of the experiments, no communities were found. On the other hand, Pearson seems to be in line with handover of work and joint cases, with 39%. This finding is probably the cause for our conclusion in Section 7.2, where we found that the Pearson correlation coefficient was very different from the other distance metrics.

In Table 4, the average densities across the various event logs are summarized, per extraction method. The joint activities networks have very high values, which means that these networks are almost fully connected. A subgroup is defined as a group of resources that are strongly connected with each other, but loosely with the rest of the network. Therefore, we suspect that the high connectedness of the joint activities networks creates difficulties for the community detection algorithms.

Based on these results, we can state that the Pearson distance is probably the most suited for community analysis.

Joint cases Joint cases is the most general extraction method. It measures in how much percent of the cases resource A and resource B worked together. Because the proximity of resources is not taken into account, this metric is heavily dependent on the definition of the case. Suppose that we have an event log of a company-wide sales process, that comprises request for quotation, production of goods, payment, etc. In this situation, the metric between someone working in the administrative department and someone on the production department, will get the same importance as the metric between two resources working together at the production department. Therefore, community analysis algorithms will have difficulties finding a distinction between the resources in the administrative department and resources in the production department.

Table 13: Random case of artificial event log.

Case ID	Activity ID	Resource	Department
4	activity A	resource C	department 1
4	activity B	resource B	department 1
4	activity C	resource B	department 1
4	activity F	resource H	department 2
4	activity G	resource F	department 2
4	activity E	resource G	department 2
4	activity L	resource D	department 1
4	activity K	resource B	department 1

Because no event logs with the related social structure are available, we tested this by making a simple artificial event log of a process conducted in two departments. Resources A, B, C and D form department 1 and resources E, F, G and H work in department 2. An important assumption we made, is that a case always has the same structure, namely a random number of activities conducted in department 1 is followed by a random number of activities carried out by department 2. The case ends with some random activities of department 1. A total of 1000 cases were generated.

On this event log, we repeated the methodology described in Section 6. The results, for one of the community analysis algorithms, are summarized in Table 14. As we can see, no communities were found in the joint cases network and the joint activities Hamming network. All 6 community detection methods showed the same results for these two networks. In the other networks, the proper communities were found, by at least 50% of the algorithms.

Table 14: Community analysis on artificial event log.

Resource	HOW	SC	JC	JAP	JAH	JAM1	JAM2
Number communities	2	2	1	2	8	2	2

These results indicate that the joint cases metric is probably not suited for community analysis. Especially, if the cases consist of activities conducted sequentially by different departments, as in our example. Nevertheless, we can not generalize these results to real-life event logs because they are based on a random artificial environment. Additional research on real-life event logs is needed to verify these findings.

8 Conclusion

In this work, we analyzed the possibilities of using event logs as input for SNA, with the emphasis on community analysis. The techniques for extracting social networks from event logs, proposed by [4], were analyzed and applied. The results denoted that these methods yield social networks that are very different from each other. Handover of work, subcontracting and joint cases network are directed, whereas joint activities networks are undirected. Furthermore, great differences were found in the topology of these networks. Based on the density measure, we found that the networks have a different level of strictness. Subcontracting networks, for example, comprised far less relations than joint activities networks. As a result, different relations were found between the same set of resources, depending on the extraction method.

Community detection techniques were applied to different social networks from the same event logs. Our goal was to investigate whether the results of community analysis are dependent on the choice of extraction method. Therefore, we compared the community divisions found in the social networks. As to be expected, little resemblance between the community partitions was noted. Especially, the community divisions from joint activities networks were unique. Within the joint activities social networks, the community partitions of the Pearson distance metric differed a lot from the other distance metrics. These results show that the results of SNA will vary based on the extraction method, which creates validation problems. However, we identified one scenario in which the choices are limited, namely when each case in the event log is executed by just one resource. Handover of work, subcontracting and joint cases are not applicable in this situation. Therefore, joint activities should be chosen. In all other cases, the choice remains open.

In our experiment, the community detection algorithms had issues finding subgroups within joint activities networks. Only Pearson distance approached handover of work, subcontracting and joint cases. Minkowski ($n = 1$ and $n = 2$) and Hamming, performed poor. The chance of not finding a community structure was on average greater than the chance of finding one.

To conclude, the answer to our research question is definitely positive. The results of SNA will most likely vary based on the extraction method that is chosen.

9 Future work

This paper has given a first insight in the comparison of the different process mining techniques for social network extraction. Future research could focus on the application of these techniques to real-life event logs, where the underlying social structure is known. The comparison of the results of SNA with the ground truth, can yield additional insights. In this study, we used community analysis techniques to compare the different social networks. Besides community detection, SNA comprises many more aspects, such as hierarchical analysis, detection of important actors and routing problems. Therefore, future work could investigate the application of these techniques on social networks retrieved from event logs.

References

1. van der Aalst, W., Adriansyah, A., de Medeiros, A.K.A., Arcieri, F., Baier, T., Blickle, T., Bose, J.C., van den Brand, P., Brandtjen, R., Buijs, J., Burattin, A., Carmona, J., Castellanos, M., Claes, J., Cook, J., Costantini, N., Curbera, F., Damiani, E., de Leoni, M., Delias, P., van Dongen, B.F., Dumas, M., Dustdar, S., Fahland, D., Ferreira, D.R., Gaaloul, W., van Geffen, F., Goel, S., Günther, C., Guzzo, A., Harmon, P., ter Hofstede, A., Hoogland, J., Ingvaldsen, J.E., Kato, K., Kuhn, R., Kumar, A., La Rosa, M., Maggi, F., Malerba, D., Mans, R.S., Manuel, A., McCreesh, M., Mello, P., Mendling, J., Montali, M., Motahari-Nezhad, H.R., zur Muehlen, M., Munoz-Gama, J., Pontieri, L., Ribeiro, J., Rozinat, A., Seguel Pérez, H., Seguel Pérez, R., Sepúlveda, M., Sinur, J., Soffer, P., Song, M., Sperduti, A., Stilo, G., Stoel, C., Swenson, K., Talamo, M., Tan, W., Turner, C., Vanthienen, J., Varvaressos, G., Verbeek, E., Verdonk, M., Vigo, R., Wang, J., Weber, B., Weidlich, M., Weijters, T., Wen, L., Westergaard, M., Wynn, M.: Process mining manifesto. Business Process Management Workshops pp. 169–194 (2012)
2. Aalst, W.M.P., F. van Dongen, B., W. Günther, C., Rozinat, A., Verbeek, E., Weijters, A.: ProM: The Process Mining Toolkit (2009)
3. van der Aalst, W.M.P., Reijers, H.A., Song, M.: Discovering social networks from event logs. Computer Supported Cooperative Work (CSCW) **14**(6), 549–593 (2005)
4. van der Aalst, W.M.P., Song, M.: Mining social networks: Uncovering interaction patterns in business processes. Business Process Management pp. 244–260 (2004)
5. Bagrow, J.P., Bollt, E.M.: Local method for detecting communities. Physical Review E **72**(4), 046108 (2005)
6. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment **2008**(10), P10008 (2008)
7. Buijs, J.C.A.M.: Environmental permit application process (ǵ1wabo ǵ2), coselog project [data set]. eindhoven university of technology (2014)

8. Buijs, J.C.A.M.: Receipt phase of an environmental permit application process ('wabo'), coselog project [data set]. eindhoven university of technology (2014)
9. Butts, C.T.: Social network analysis: A methodological introduction. Asian Journal of Social Psychology **11**(1), 13–41 (2008)
10. Capocci, A., Servedio, V.D.P., Caldarelli, G., Colaiori, F.: Communities detection in large networks. Algorithms and Models for the Web-Graph pp. 181–187 (2004)
11. Chen, J., Saad, Y.: Dense subgraph extraction with application to community detection. IEEE Transactions on Knowledge and Data Engineering **24**(7), 1216–1230 (2012)
12. Christakis, N.A., Fowler, J.H.: The spread of obesity in a large social network over 32 years. New England Journal of Medicine **357**(4), 370–379 (2007)
13. Clauset, A., Newman, M.E.J., Moore, C.: Finding community structure in very large networks. Physical Review E **70**(6), 066111 (2004)
14. Creamer, G., Rowe, R., Hershkop, S., Stolfo, S.J.: Segmentation and automated social hierarchy detection through email network analysis, pp. 40–58. Springer (2009)
15. Cross, R., Parker, A., Borgatti, S.P.: A bird's-eye view: Using social network analysis to improve knowledge creation and sharing. IBM Institute for Business Value pp. 1669–1600 (2002)
16. Csardi, G., Nepusz, T.: The igraph software package for complex network research. InterJournal **Complex Systems**, 1695 (2006), http://igraph.org
17. Daly, E., Haahr, M.: Social network analysis for routing in disconnected delay-tolerant MANETS (2007)
18. Danon, L., Díaz-Guilera, A., Duch, J., Arenas, A.: Comparing community structure identification. Journal of Statistical Mechanics: Theory and Experiment **2005**(09), P09008 (2005)
19. De Leoni, M., Mannhardt, F.: Road traffic fine management process [data set]. eindhoven university of technology (2015)
20. De Meo, P., Ferrara, E., Fiumara, G., Provetti, A.: Mixing local and global information for community detection in large networks. Journal of Computer and System Sciences **80**(1), 72–87 (2014)
21. Donetti, L., Muñoz, M.A.: Improved spectral algorithm for the detection of network communities. AIP Conference Proceedings **779**(1), 104–107 (2005)
22. Duch, J., Arenas, A.: Community detection in complex networks using extremal optimization. Physical review E **72**(2), 027104 (2005)
23. Ferreira, D.R., Alves, C.: Discovering user communities in large event logs. Business Process Management Workshops pp. 123–134 (2012)
24. Fortunato, S., Latora, V., Marchiori, M.: Method to find community structures based on information centrality. Physical review E **70**(5), 056104 (2004)
25. Freeman, L.C.: A set of measures of centrality based on betweenness. Sociometry pp. 35–41 (1977)
26. Freeman, L.C.: Centrality in social networks conceptual clarification. Social networks **1**(3), 215–239 (1978)
27. Girvan, M., Newman, M.E.J.: Community structure in social and biological networks. Proceedings of the National Academy of Sciences **99**(12), 7821 (2002)
28. Guimerà, R., Amaral, L.A.N.: Cartography of complex networks: modules and universal roles. Journal of statistical mechanics (Online) **2005**(P02001), nihpa35573–nihpa35573 (2005)
29. Harenberg, S., Bello, G., Gjeltema, L., Ranshous, S., Harlalka, J., Seay, R., Padmanabhan, K., Samatova, N.: Community detection in large-scale networks: a

survey and empirical evaluation. Wiley Interdisciplinary Reviews: Computational Statistics **6**(6), 426–439 (2014)
30. Haythornthwaite, C.: Social network analysis: An approach and technique for the study of information exchange. Library Information Science Research **18**(4), 323–342 (1996)
31. Hubert, L., Arabie, P.: Comparing partitions. Journal of Classification **2**(1), 193–218 (1985)
32. Jain, A., Dubes, R.: Algorithms for clustering data. Prentice-Hall, Inc. (1988)
33. Jiang, P., Singh, M.: Spici: a fast clustering algorithm for large biological networks. Bioinformatics **26**(8), 1105–1111 (2010)
34. de Laat, M., Lally, V., Lipponen, L., Simons, R.J.: Investigating patterns of interaction in networked learning and computer-supported collaborative learning: A role for social network analysis. International Journal of Computer-Supported Collaborative Learning **2**(1), 87–103 (2007)
35. Levy, D.: Production analysis with process mining technology [data set]. nool - integrating people solutions (2014)
36. Mannhardt, F.: Sepsis cases - event log [dataset]. eindhoven university of technology (2016)
37. Marsden, P.V.: Network data and measurement. Annual Review of Sociology **16**(1), 435–463 (1990)
38. Newman, M.E., Girvan, M.: Finding and evaluating community structure in networks. Physical review E **69**(2), 026113 (2004)
39. Otte, E., Rousseau, R.: Social network analysis: a powerful strategy, also for the information sciences. Journal of Information Science **28**(6), 441–453 (2002)
40. Ovelgönne, M., Geyer-Schulz, A.: An ensemble learning strategy for graph clustering, vol. 588 (2013)
41. Palla, G., Derényi, I., Farkas, I., Vicsek, T.: Uncovering the overlapping community structure of complex networks in nature and society. Nature **435**, 814 (2005)
42. Pons, P., Latapy, M.: Computing communities in large networks using random walks. International symposium on computer and information sciences pp. 284–293 (2005)
43. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2016), https://www.R-project.org/
44. Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., Parisi, D.: Defining and identifying communities in networks. Proceedings of the National Academy of Sciences **101**(9), 2658–2663 (2004)
45. Raghavan, U.N., Albert, R., Kumara, S.: Near linear time algorithm to detect community structures in large-scale networks. Physical Review E **76**(3), 036106 (2007)
46. Rand, W.M.: Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association **66**(336), 846–850 (1971)
47. Reffay, C., Chanier, T.: How social network analysis can help to measure cohesion in collaborative distance-learning (2003)
48. Reichardt, J., Bornholdt, S.: Detecting fuzzy community structures in complex networks with a potts model. Physical Review Letters **93**(21), 218701 (2004)
49. Riedy, J., Bader, D.A., Meyerhenke, H.: Scalable multi-threaded community detection in social networks. 2012 IEEE 26th International Parallel and Distributed Processing Symposium Workshops PhD Forum pp. 1619–1628 (2012)

50. Rosvall, M., Bergstrom, C.T.: Maps of random walks on complex networks reveal community structure. Proceedings of the National Academy of Sciences **105**(4), 1118 (2008)
51. Song, M., van der Aalst, W.M.P.: Towards comprehensive support for organizational mining. Decision Support Systems **46**(1), 300–317 (2008)
52. Van Dongen, B.F.: Bpi challenge 2012 [data set]. eindhoven university of technology (2012)
53. Van Dongen, B.F.: Bpi challenge 2017 [data set]. eindhoven university of technology (2017)
54. Van Dongen, S.: Performance criteria for graph clustering and markov cluster experiments. Report, CWI (Centre for Mathematics and Computer Science) (2000)
55. Wu, F., Huberman, B.A.: Finding communities in linear time: a physics approach. The European Physical Journal B **38**(2), 331–338 (2004)
56. Zhou, H., Lipowsky, R.: Network brownian motion: A new method to measure vertex-vertex proximity and to identify communities and subcommunities. International conference on computational science pp. 1062–1069 (2004)