Kuikka *Comput Soc Netw* (2021) 8:13
https://doi.org/10.1186/s40649-021-00094-z

Computational Social Networks

RESEARCH Open Access

Modelling community structure and temporal spreading on complex networks

Vesa Kuikka*

*Correspondence:
vesa.kuikka@mil.fi
Finnish Defence Research Agency, Tykkikentäntie 1, PO BOX 10, 11311 Riihimäki, Finland

Abstract

We present methods for analysing hierarchical and overlapping community structure and spreading phenomena on complex networks. Different models can be developed for describing static connectivity or dynamical processes on a network topology. In this study, classical network connectivity and influence spreading models are used as examples for network models. Analysis of results is based on a probability matrix describing interactions between all pairs of nodes in the network. One popular research area has been detecting communities and their structure in complex networks. The community detection method of this study is based on optimising a quality function calculated from the probability matrix. The same method is proposed for detecting underlying groups of nodes that are building blocks of different sub-communities in the network structure. We present different quantitative measures for comparing and ranking solutions of the community detection algorithm. These measures describe properties of sub-communities: strength of a community, probability of formation and robustness of composition. The main contribution of this study is proposing a common methodology for analysing network structure and dynamics on complex networks. We illustrate the community detection methods with two small network topologies. In the case of network spreading models, time development of spreading in the network can be studied. Two different temporal spreading distributions demonstrate the methods with three real-world social networks of different sizes. The Poisson distribution describes a random response time and the e-mail forwarding distribution describes a process of receiving and forwarding messages.

Keywords: Complex networks, Community detection, Building blocks of community, Influence spreading model, Network connectivity, Probability matrix, E-mail forwarding process, Temporal spreading on networks

Introduction

Network science is a field that studies a wide range of properties and phenomena on complex networks [1]. Applications are on social networks, biomedical networks, technical networks and many other fields. Some practical examples of research questions are studying more realistic community detection methods, investigating processes on networks, and defining different quantitative measures for comparing and ranking network structures.

This study proposes an algorithm for detection of communities and their sub-structures in networks. The proposed algorithm is based on optimisation of a quality function whose input is a matrix that captures the strength of 'interaction' between each pair of nodes in the network. The nature of the interaction can vary depending on the application and the model that is used to describe the network. This study focuses on two such models: the influence spreading model and the network connection model. In the case of the influence spreading model, the matrix that is used in the quality function is the influence spreading matrix whose elements capture the influence of nodes on one another at an equilibrium state (i.e. at an infinite time horizon). In the case of the network connection model, the network connectivity matrix is used which captures the reliability of connection between each pair of nodes in the network. This approach differs from other community detection algorithms in the literature that use the adjacency matrix to detect communities. The study also numerically implements the proposed algorithm for two network structures: the Les Misérables network and Zachary's karate club. As another exercise, in the context of the influence spreading model, we assume that the network is not necessarily at an equilibrium state (i.e. time is finite), and show how one can repeat the analysis for any finite time. We also provide methods that can help us check robustness and strength of the communities detected by this proposed algorithm.

Properties and characteristics of a network can be analysed with several network metrics. An example of an important class of network measures is that of centrality measures. There are a variety of mathematical measures of centrality that focus on different concepts and definitions of what it means to be important or central in the network. The generalised centrality measure proposed in this study is a form of node influence metrics that rank or quantify the influence of nodes.

Conventional methods of detecting communities are not directly based on a particular network measure. Because centrality and community structure are related concepts, a reasonable conclusion is that community detection algorithms and definitions of centrality should be based on a common methodology. The same argument can be stated for community detection algorithms used in time-dependent models. We propose one underlying methodology for studying network characteristics, community structure and temporal spreading on complex networks. The methodology of this study uses a probability matrix or an influence spreading matrix as a common basis for analysing community structure and defining network measures.

In section 'Centrality measures', we introduce basic properties of conventional centrality measures in order to provide background information for comparison with the definitions proposed in section 'Influence spreading model'. In section 'Community detection methods', we provide a short introduction to community detection methods in the literature, especially those that use the adjacency matrix of a network. The community detection method proposed in this study analyses the network structure more deeply than conventional methods aiming at uncovering the building blocks of potential communities and sub-communities. We demonstrate the method with a static and a dynamic network model in order to compare the community structures of the two models and also to validate the community detection method itself.

After section 'Related literature' the body of the text is divided into four main parts. The idea and general methodology of this study are explained in section 'General methodology'. After that, we present the detailed methods and models in sections 'Influence spreading model', 'Community detection method in the context of the influence spreading model' and 'Community detection in the context of the network connection model'. Applications with examples of modelling static network structure are present in section 'Application of the community detection model'. Finally, the method of modelling spreading on network structure is described and demonstrated with two example temporal distributions in section 'Temporal spreading on networks in the influence spreading model'. The contributions and applications are further discussed in section 'Conclusions'.

Related literature

Concepts of network science have been presented comprehensively in books written by Albet-László Barabási [1] and M.E.J. Newman [2]. Before we continue with related literature, we refer to some basic concepts. A network can be defined as a graph in which nodes and links have attributes. Networks are characterised as complex when nodes are connected with non-trivial topological features. In [3] a complex network is defined as a network that is substantially distinct from a regular network or uniformly random network. In a regular network, if one knows the degree of one node, all the other degrees will also be revealed. Degree is the number of edges that are incident to the node [1]. Example network structures used in this study are complex in line with the definition in [3]. However, the models in this study can also be applied to simple networks, such as lattices or random networks.

Standard centrality and betweenness measures are introduced in the next section together with some comments on some less commonly used centrality measures related to the definitions of centrality proposed in this study. Brief general introductions to community detection methods, processes on networks and the standard network connection model are provided in sections 'Community detection methods', 'Processes on complex networks' and 'Network connection model'.

Centrality measures

Nodes in a network can have different roles as central influencers, mediators or peripheral nodes. Centrality quantifies how important nodes are in the networked system. Degree is the simplest closeness centrality measure. Degree centrality of a node is defined as the number of nodes connected to it. It is a local measure and does not take into account the node's position in the network. Closeness centrality of a node measures how central or influential the node is with respect to other nodes. Betweenness centrality measures the role of a node as a proxy between other nodes and measures the ability to mediate influence in the network. Closeness and betweenness centrality measures have many variants and they depend on the research question of a particular application [4]. In-centrality and out-centrality can be defined both for directed and undirected networks.

Standard closeness centrality is based on the inverse sum of the shortest distances to the other nodes of the network. Standard betweenness centrality of a node is based on counting how often it falls on connecting paths between pairs of nodes. Nodes having high betweenness centrality values can control the flow of information. In addition to the standard measures, many other definitions have been proposed in the literature [4]. Some of them are related to the centrality and betweenness measures proposed in section 'Influence spreading model' of this study. One of the centrality measures is the Katz centrality measure [5]. The Katz centrality measure generalises the degree centrality and the closeness centrality by taking into account not only the immediate neighbours or not only the shortest paths from a node to other nodes. The Katz centrality for a node i is defined as

$$C_i^{(Katz)} = \sum_{k=1}^{\infty} \sum_{j=1}^{N} \alpha^k \left(A^k\right)_{ji}.$$

The power of adjacency matrix A accounts for the number of paths of length k between every pair of nodes in a network of N nodes. A decay parameter $\alpha < 1$ is introduced to weight the contributions of nodes at increasing path lengths.

Eigenvector centrality is a measure that depends recursively on the centralities of node's neighbours. Katz centrality, hubs and authorities centrality and PageRank are variants of eigenvector centrality. The basic eigenvector centrality measure is used only for undirected networks whereas the three variants are also appropriate for directed networks. Hubs and authorities centrality assigns to each node two different measures both for sending and receiving influence. For undirected networks this definition coincides with the basic eigenvector centrality and the distinction between hubs and authorities disappears. [4]

Community detection methods

Community detection is one of the most important applications of modelling complex networks. One problem with various approaches is the lack of a commonly accepted definition of a community. On the other hand, one definition for all possible purposes may not be possible because of different empirical data available and requirements in applications. A more realistic approach would be first to categorise and define more general concepts of complex networks, such as types of networks, processes, function rules and interactions. In many cases, the concept of a community is not defined exactly, instead the used method and algorithm define the concept implicitly. Mathematical methods and algorithms for detecting communities in complex network topologies have been reviewed and presented in [6–9].

Modularity maximisation, classical graph partitioning, spectral graph partitioning and several information-theoretic methods are examples in the wide context of community detection methods [1, 2]. The classical graph partitioning is the problem of dividing the nodes of a network into a given number of non-overlapping groups of given sizes such that the number of links between groups is minimised. Modularity measures the robustness of division of a network into modules. One definition of a community is a locally dense connected sub-graph in a network [1]. Modularity has been defined as the fraction of links falling within the given groups minus the expected fraction if links were distributed at random. To compute the numerical value of modularity, each link is cut into two halves, called stubs. The expected number of links is computed by rewiring stubs randomly with any other stub in the network, except itself, but allowing self-loops when a stub is rewired to another stub from the same node. Mathematically, modularity can be expressed as

$$M = \frac{1}{2m} \sum_{vw} \left(A_{vw} - \frac{k_v k_w}{2m} \right) \frac{s_v s_w + 1}{2}. \tag{1}$$

In Eq. (1) v and w are nodes in the network, $2m$ is the number of stubs in the network, k_v is the node degree of node v, $A_{vw} = 1$ means that there is a link between nodes v and w, and $A_{vw} = 0$ means that there is no link between the two nodes. Matrix A is called the adjacency matrix. Membership variable s_v indicates if node v belongs to a community: $s_v = 1$ if node v belongs to community 1, and $s_v = -1$ if node v belongs to community 2. [2]

Equation (1) holds for partitioning into two modules but it can be generalised for partitioning into a desired number of modules. In matrix terms Eq. (1) is

$$M = \frac{1}{4m} s^T B s,$$

where $B_{vw} = A_{vw} - \frac{k_v k_w}{2m}$ is called the modularity matrix. The equation for M is similar in form to an expression used in spectral partitioning of graphs for the cut size of a network in terms of the graph Laplacian. This similarity can be used for deriving a spectral algorithm for community detection. The eigenvector corresponding to the largest eigenvalue of the modularity matrix assigns nodes to communities according to the signs of the vector elements. [2]

The Louvain algorithm [10] and Infomap [11] are two fast algorithms for community detection that have been briefly described in [1]. These algorithms have gained popularity because of their suitability for identifying communities in very large networks. Both algorithms optimise a quality function. For the Louvain algorithm the quality function is modularity and for Infomap it is an entropy-based measure. In the Louvain algorithm modularity is optimised by local changes of the modularity measure and communities are obtained by aggregating the modules to build larger communities. Infomap compresses the information about a random walker exploring the graph. [1]

Stochastic blockmodels have been used as a method for detecting community structure in networks and also for generating synthetic benchmark networks [12]. Many community detection methods discover also hierarchical and overlapping sub-communities in complex networks [13]. In a recent study [14] an information-theoretic method has been presented for discovering groups (building blocks) of network nodes that are usually found together in the same community.

Processes on complex networks

In a dynamical system, the state of the system changes over time according to some given rules [15]. Dynamical processes on complex networks have growing interest because of their wide scope of applications. Epidemic spreading in populations [16], influence spreading in social networks [17] and the flow of traffic on roads are important practical applications [15]. In some specific applications, it is not clear which kind of models best describe the system, or whether both static and dynamic behaviour can coexist in the same system [18]. Standard approaches to studying dynamical processes on networks rely on simulations because analytical mathematical expressions are not available or they are very complicated. However, remarkable research has been conducted in percolation theory where analytical solutions exist for some network topologies. These methods are not directly applicable for modelling the detailed topology of empirical networks. [2]

By definition, processes on networks have a time-dependency. Two examples of processes on networks are spreading on network structure and changing network topology [19]. Spreading processes start from a node or a set of nodes in the network and propagate between nodes via links in the network structure. Also, the network topology may change during the spreading process. Changes in network structure and changes in link and node attribute values are common in many applications. For example, virus spreading in computer networks or in human social networks, is slowed down with virus protection software or vaccination programmes in the human population.

Spreading processes on networks cover a variety of situations because processes can depend on states of other nodes or links in the network. For example, virus spreading may not be possible or it is only partial in case nodes are immunised as a result of previous contamination [16]. Another example is when information or rumours are spread more actively when heard for the first time compared to later versions of the same information. In social systems, many overlapping processes are simultaneously influencing our beliefs and opinions [20].

Network connection model

The classical network connection model is designed for describing reliability of communication networks [22]. If the reliability values between all neighbouring pairs of nodes in the network are known, reliability values between any pairs of nodes in the network can be computed. Here, reliability is identified with the probability of a functioning connection. From the general reliability theory [22] the reliability of a network V is

$$r(V) = \sum_{\mathcal{S}\in\mathcal{O}} \prod_{e\notin\mathcal{S}} (1-p_e) \prod_{e\in\mathcal{S}} p_e,$$

where $\mathcal{S}$ is a set of links where the network is connected and $\mathcal{O}$ is the set of all connected states of the network. Links are denoted by e and the probability of a functioning link is denoted by p_e. If all the probabilities p_e are equal,

$$r(V) = 1 - \prod_{h=1}^{N_L} \left(\sum_{s=0}^{h} (-1)^{h-s} \binom{N_L - s}{N_L - h} H_S \right) p_e^h, \tag{2}$$

where H_s is the sum of indicator functions where s is the number of broken links. The above equations are polynomials of the order of the number of links N_L in the network. In this form, the equations describe the reliability of an entire network. In our case, we apply the results for pairs of nodes by only taking the relevant terms in the summations (see section 'Community detection in the context of the network connection model' for more details).

General methodology

Here, we present the general methodology and example network models used in this study. We introduce the idea of using different network models for describing the network structure. The problem of detecting communities and sub-communities is solved with a community detection algorithm based on searching local maxima of an objective function (see Eq. (8)). The objective function is calculated from a probability matrix or an influence spreading matrix that describe interactions between node pairs in the network. Different network models can be used with the community detection method proposed in [21].

Different network models and community detection methods

Community detection methods have been developed in order to understand the formation of groups in social networks or for categorising sub-systems in technological and biological networks. We present two examples of network models: a network connection model [22] and an influence spreading model [21]. In both cases, we use the same community detection algorithm. In these two examples, interactions between nodes are described as spreading probabilities or probabilities of functioning connections. Technically, interactions between pairs of nodes in the network are expressed in a $N \times N$ -dimensional probability matrix where N is the number of nodes in the network. The probability matrix contains computed values for all pairs of nodes in the network, not just for neighbouring nodes.

Figure 1 illustrates the general idea of the methodology. In the following, we use the symbol $'IS'$ for influence spreading, $'NC'$ for network connectivity and $'CD'$ for community detection. Here, we use the symbol $\boldsymbol{C_{IS}}$ for the probability matrix in the context of influence spreading and $\boldsymbol{P_{NC}}$ for the probability matrix in the context of network connectivity. The community detection method (CD) is similar in the context of influence spreading ($IS - \boldsymbol{C_{IS}} - CD$) and in the context of network connectivity ($NC - \boldsymbol{P_{NC}} - CD$). The matrix $\boldsymbol{C_{IS}}$ or the matrix $\boldsymbol{P_{NC}}$ mediates information to the community detection model. In section 'Community detection method in the context of the influence spreading model' we present the community detection method proposed in [21], although variants that use similar input information in the form of a probability matrix are also possible. One alternative method has been presented in [27]. The method of calculating the values of a network connectivity matrix [22] is presented in section 'Community detection in the context of the network connection model'.

Example network models

We use two network models to demonstrate the community detection method: the classical network connection model [22] and the influence spreading model proposed in [21, 26, 27]. The influence spreading model and the network connection model have been designed for different application areas. The primary use of the influence spreading model is for behaviour and opinion spreading and the connection model is commonly used for modelling communication networks.

The two network models have different application areas, definitions and parameterisations. However, for the two example network topologies, the most important

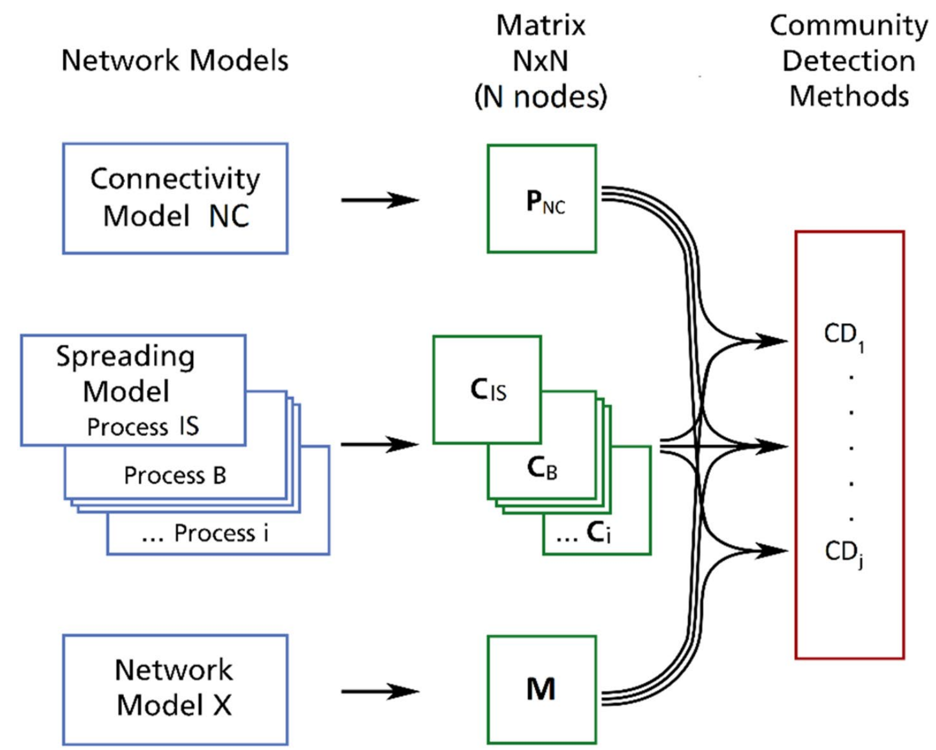


Fig. 1 General procedure for using different network models and community detection models. Information from network models are mediated with a probability matrix describing influence or connectivity between all node pairs in the network

communities and their sub-communities are close to each other. More differences appear in weaker communities and their sub-structure. The fact that the same community detection algorithm (see Eq. (8)) provides reasonable results for different network models suggests that the method is generally useful for community detection in various applications. On the other hand, detailed features of the community detection method are important because sub-structures are also discovered. The community detection method is not limited to particular network characteristics: directed, weighted, time-dependent and layered networks can be analysed.

Temporal spreading is only possible, by definition, in network spreading models. This is the reason why we present the method for modelling temporal distributions for the influence spreading model and not for the network connection model. More detailed descriptions are provided in section 'Influence spreading model' and in [21]. The network connection model is included in this study mostly for comparison purposes and more details of the model can be found in textbooks or research articles [22].

The network connection model describes connectivity between pairs of nodes in the network. This is calculated by considering all possible paths between the source node and the target node. We use the same parameter value for describing a functioning link between all two neighbouring nodes and utilise Eq. (2). Low weighting factors are used in social networks for describing probabilities of social influence [27]. Although we apply the network model designed for physical communication network modelling, we use low values for link weights as in the modelling of influence spreading.

Influence spreading model

Influence spreading models are designed for describing complex social interactions in social network structure [16, 28]. These interactions propagate via connections, or paths, between people. We assume that information content can change and ways of social influence are developing during the spreading process. We allow repeated attempts of influence from a source node to target nodes via all alternative paths.

Other features of the model are: weighted links and nodes, directed links, and the possibility of using different forms of temporal survival distributions as a function of the number of links between a source node and a target node. All the parameters have real-world interpretations. Link and node weights are interpreted as spreading probabilities of forwarding influence between neighbouring nodes and over a node, respectively. Spreading probabilities between all pairs of nodes in a structured network can be calculated from node and link weighting values and the temporal survival distribution function. [21]

The influence spreading model used in this study has been presented previously in [21, 26, 27]. The model has analytical expressions for influence spreading probabilities via different paths from a source node to target nodes in the network topology. In this version of the model, the rate of spreading is assumed to be independent of the state of the network and its elements. To avoid double counting effects, common paths at their beginning from one source node to a target node are taken into account by applying probability theory. In other respects, different paths are assumed to be independent. As a result, in case paths join or cross later, the spreading process is not affected.

An important property of a process is the possibility of loops. If recurrent visits on a node are allowed, one node can be visited several times during the process. In this study, we assume that loops are allowed, but no self-loops. Social influence and spreading of beliefs can be modelled with these kinds of processes. In the process of spreading information and news, loops are less probable. In the algorithm, it is also possible to set the maximum number of visits $V = 1, \ldots, L$, where L is of the maximum path length of computing. Limiting visits may be computationally costly.

The mathematical model considers all paths between all node pairs of a network. From this information, we construct the influence spreading matrix, or the probability matrix, that describes influence spreading from source nodes to all other nodes in the network. Because of the complex structure of networks, the influence spreading probability from node A to node B is not equal to the probability from node B to node A. This means that node A can have more influence on node B than node B on node A or vice versa. As a consequence, in most real cases, the influence spreading matrix is not symmetric. One consequence of this is that peripheral nodes that are locally densely connected can have a considerable effect on other parts of the network. Influence spreading accumulates momentum locally at early phases of the process and later starts a more intensive spreading. [21]

Here, we give an idea of how the model can be programmed with a computing language. A more detailed pseudo-algorithm has been presented in [21]. The algorithm goes through all paths from a source node to a target node. For every path, the probability of spreading is computed by multiplying all the link and node weights along the path. We denote this factor by W_L where L is the path length (number of links along the path). The interpretation of the link weight is the probability of spreading influence via the link and the interpretation of the node weight is the probability of spreading influence over the node. The maximum length of processing paths in computations can be limited by a parameter. Good values of the parameter can be estimated by monitoring the accuracy of calculations and computing times as a function of L in W_L.

In spreading processes, we have to consider also the temporal distribution of spreading as a function of the path length. We denote the probability of temporal spreading at least via L links as a function of time T by $S_L(T)$. Mathematically, the probability is the survival function $S_L(T) = 1 - F_L(T)$, where $F_L(T)$ is the distribution function of the temporal spreading probability. However, in this section, we consider only equilibrium states for time T approaching infinity. We describe how the model is applied with finite time $T < \infty$ in section 'A numerical example of calculating time-dependent values of the influence spreading matrix'. The model predicts that in equilibrium states the spreading process does not reach all nodes with probability one, if link weights are not all equal to one along the paths between all node pairs in the network. Finally, in a network, the spreading has reached all nodes with a probability determined by link and node weights alone. We have for every L the limiting value of the survival function

$$\lim_{T \to \infty} S_L(T) = 1.$$

We write an iterative formula for calculating spreading probabilities between nodes in the network. We consider paths between two nodes in the network: a source node and a target node. Paths from source node s to target node t are combined iteratively in the descending order of common path lengths at the beginning of their paths:

$$P_{i,L_{i,2}} = P_{i,L_{\mathrm{i},1}} + P_{i-1,L_{\mathrm{i}-1,2}} - \frac{P_{i,L_{\mathrm{i},1}} P_{i-1,L_{\mathrm{i}-1,2}}}{W_{L_{\mathrm{i},3}}}, i = 1, \ldots, N_L - 1. \tag{3}$$

The path length $L_{\mathrm{i},2}$ in iteration i is $L_{i,2} = L_{i,max(L_{i,1},L_{i-1,2})}$ and the common path length of $L_{\mathrm{i},1}$ and $L_{\mathrm{i}-1,2}$ is denoted by $L_{\mathrm{i},3}$. The number of different paths from the source node to the target node is denoted by N_{L}. The iteration starts with two paths with $P_{1,L_{1,1}} = W_{L_{1,1}}$ and $P_{0,L_{0,2}} = W_{L_{0,2}}$ having the longest common path length $L_{1,3}$. If there are more than two paths with the same common path length, these paths can be processed in any order. In later steps of the iteration, combined paths are processed in the same way as the original paths of the network. A numerical example in [21] illustrates the algorithm in practice. The probability of influence spreading between the two nodes is the final result of the algorithm after all the paths have been processed. Denoting the source node by s and the target node by t the spreading probability from node s to node t is given by

$$C_{s,t} = P_{\mathrm{i},1,\mathrm{L}_{\mathrm{i},2}}, \; i = N_L - 1. \tag{4}$$

In the last step of the iteration, the length of the last combined path is $L_{N_L-1,2}$. Matrix elements $C_{s,t}$ describe probabilities of influence between each pair of nodes in the network. We coin the name 'influence spreading matrix' for matrix $\boldsymbol{C}$.

Un-normalised out-centrality and in-centrality measures for nodes s and t can be defined as

$$C_s{}^{(out)} = \sum_{t=1}^{N} C_{s,t}, \tag{5}$$

$$C_t{}^{(in)} = \sum_{s=1}^{N} C_{s,t}. \tag{6}$$

Normalised versions of these centrality measures can be obtained by dividing the expressions by N or $N-1$ depending on whether the diagonal elements of influence matrix $\boldsymbol{C}$ are set to one or zero. The corresponding betweenness centrality measure for node n can be defined as

$$b_n = \frac{\mathcal{C} - B_n}{\mathcal{C}}, \tag{7}$$

where

$$\mathcal{C} = \sum_{s,t=1}^{N} C_{s,t},$$

and B_n is calculated similarly to $\mathcal{C}$ with node n removed from the network

$$B_n = \sum_{\substack{s,t=1 \\ n \notin V}}^{N} C_{s,t}.$$

Later in this study Eq. (5) is used as the definition for the node centrality, and Eq. (7) is used as the definition for the betweenness centrality. The quality function of Eq. (8) is based on both aspects of centrality in Eqs. (5) and (6).

Community detection method in the context of the influence spreading model

In this section, we present the community detection method in the context of the influence spreading model (ISMCD) [21]. We take an approach, where instead of performing community detection based on the adjacency matrix of a graph, an influence matrix is first constructed that contains information about a social influence process over the paths of a graph [21]. An element C_{vw} of the influence spreading matrix $\boldsymbol{C}$ accounts for interactions over all the paths from source node v to target node w. In order to study local interactions in the network, a maximum path length can be set in the algorithm. In addition, we take into account that interactions in communities can mean different things depending on the processes that are supposed to operate on a network. This is demonstrated by substituting the social influence matrix with the network connectivity matrix known from the classical communication theory [22].

The method is based on searching local maxima of a community influence measure computed from the probability matrix elements $C_{s,t}, s, t = 1, \ldots, N$. Our basic model has the following form for the community influence measure:

$$P = \sum_{s,t \in V} C_{s,t} + \sum_{s,t \in (G-V)} C_{s,t}. \quad (8)$$

In Eq. (8) the first summation is over the pairs of nodes in a subset V of nodes in network G. The second summation is over the pairs of nodes in the remaining partition of the network denoted by $(G - V)$. Cross terms are ignored in this version of the model because they describe interactions between the two partitions of the division and are not directly involved in the internal cohesion of the partitions (sub-communities). The method detects many kinds of structures in topological complex networks: non-overlapping, overlapping and hierarchical community structure. As a special case, communities consisting of two or more distinct sub-communities that have no direct contact can be discovered.

The method assumes that the network is divided into two communities. However, the model provides many solutions for the local maxima of Eq. (8). Communities with high rankings according to the value of P in Eq. (8) are candidates for the split of the original community in real-life. Note that this may not be the most probable solution to the community formation process. Later, in sections 'Application to the Zachary's karate club social network' and 'Application to the Les Misérables network' we present results for two community measures: the strength P in Eq. (8) of the split into two communities and a statistical measure describing the probability of community formation. In addition, a third measure for the robustness of a composition of sub-communities is proposed.

In the community detection algorithm, a sum of rows and columns of matrix $\boldsymbol{C}$ is used as the quality function. Rows and columns are included in the sum that corresponds to node pairs in the community, and node pairs not in the community. This measure is different from the modularity M of Eq. (1). Typically, a node has a higher influence on neighbouring nodes compared to nodes that are far away in the network topology. Influence is also increasing with the number of alternative paths between nodes. Most community detection methods consider only the local influence amongst nodes in network structure. More accurate results can be obtained when longer path lengths are included in the model and calculations. In order to balance between the increasing number of alternative paths and the distance between a source node and a target node, weighting factors are used to describe probabilities of influence via links between two neighbouring nodes in the network. In static connection models the probability of functioning connection plays the role of interaction. Weighting factors for links or nodes, or both, together with the network topology are the main input data for network models.

Identifying communities of nodes has proven to be challenging due to issues with evaluation and the lack of a reliable gold-standard ground-truth [23]. In [23] a set of 230 social, collaboration and information networks was studied in which the notion of ground-truth communities were defined by nodes explicitly stating their group memberships. Empirical data of spontaneous splits of social networks into partitions are scarce. Real-world social networks can split into two partitions [24, 25] and later more sub-communities may build up. This kind of community formation is a special case of the basic model: if the original network is first divided into A and B, and later

B is divided into B_1 and B_2, usually divisions $(A \cup B_1) \cup B_2$ and $(A \cup B_2) \cup B_1$ are also local maxima of Eq. (8). To be precise, interactions between nodes in one sub-community include also interactions mediated via paths through other communities. In the model, these paths are included whenever both the source node and the target node are inside one community.

Next, we define the robustness measure of communities and sub-communities. Analysing robustness is a method to find nodes who most easily change sides in a division. Robustness for a node belonging to a community is defined as the change of community measure of Eq. (8) when the node is moved from its own partition to the second partition of the division. Because of the change in the sum of Eq. (8) is always non-positive, we define the robustness measure as the negative value of change in Eq. (8). Robustness P_n of node n is given in Eq. (9) using the same notations as in Eq. (8):

$$P_n = \sum_{\substack{s,t \in V \\ n \notin V}} C_{s,t} + \sum_{\substack{s,t \in (G-V) \\ n \notin V}} C_{s,t} - P. \tag{9}$$

Later in this study, we will compare the numerical values of different community measures (see 'Applications of the community detection model'). Strength in Eq. (8) and robustness in Eq. (9) are closely related quantities. However, rankings of the community measures can be different due to the complex structure of networks.

Community detection in the context of the network connection model

In this section, we discuss community detection in the context of the network connection model. It is based on the connectivity of node pairs [22] in the network structure. Equation (2) describes average connectivity in a network with equal probabilities of functioning links. In order to construct a matrix similar to the influence spreading matrix $\left(C_{s,t}\right), s, t = 1, \ldots, N$ of Eq. (4), we calculate the probability of connectivity between all node pairs in the network. Link weights p_e are used as input parameters in Eqs. (1) and (2). Link weights are interpreted as probabilities of functioning connections of directed links between two neighbouring nodes in the network structure.

In the following examples, we assume that links have equal weights and they do not depend on the direction. We denote the probability of a functioning link as p. In order to demonstrate the connectivity matrix, we extract a sub-structure of 12 nodes $\{1, 2, 3, 4, 8, 9, 14, 20, 31, 32, 33, 34\}$ from the Zachary's karate club social network in Fig. 3. Later in this study, we use the same sub-structure as an example when we demonstrate the time-dependence of the influence spreading model in section 'A numerical example of calculating time-dependent values of the influence spreading matrix'. We have developed a computer program for computing polynomials from Eq. (2). For example, the polynomials describing connectivity between nodes 1 and 33 as a function of p is

$$p_{1,33} = 3p^2 + 12p^3 + p^4 - 51p^5 - 132p^6 + \cdots - 3709p^{22} + 846p^{23} - 80p^{24}.$$

We give only the five leading terms and three last terms of the polynomial. If we take only path lengths $L_{max} \leq 2$, the polynomial is $p_{1,33} = 3p^2 - 3p^4 + p^6$. The polynomial is calculated by taking the relevant terms in Eq. (2). The explanation for the first term is that there are three possible paths $\{1-3-33\}$, $\{1-9-33\}$ and $\{1-32-33\}$ with path length $L=2$ and no paths with path length $L=1$. The subsequent terms in the formula of $p_{1,33}$ are calculated from the probabilistic formula of Eq. (2). The polynomial for $p_{1,33}$ can be expressed as a function $f=1-p$ as

$$p_{1,33} = f^5 + 2f^6 + 5f^7 + 20f^8 + 2f^9 - 58f^{10} - \cdots + 6330f^{22} - 1074f^{23} + 80f^{24}.$$

The corresponding formula for path lengths $L_{max} \leq 2$ is $p_{1,33} = 8f^3 - 12f^4 + 6f^5 - f^6$. Either of the two forms of $p_{1,33}$ can be used for calculating the probability matrix of the network. The first term in the formula for $p_{1,2}$ shows that there is one link between nodes 1 and 2:

$$p_{1,2} = p + p^2 + 4p^3 - 3p^4 + \cdots - 3596p^{22} + 779^{23} - 71p^{24}.$$

In this case, the matrix is symmetric. We can set the diagonal terms $p_{s,s}=1$ but the numerical value has no effect on the detected community structures although it changes the numerical values of Eq. (8) and Eq. (9). The community detection method is similar in the contexts of the influence spreading model and the network connection model. The only difference is in using different input information, i.e. the influence spreading matrix is replaced by the probability matrix explained in this section. In the following section, we present the application of the community detection method in the context of the influence spreading model. For comparison, applying the community detection method in the context of the network connection model is presented with the Zachary's karate club social network.

Application of the community detection model

We use two small networks to illustrate the application of the network connection model and the influence spreading model. We compare the results of the community detection method presented in section 'Community detection method'. Figure 2 shows the general arrangement of our investigations. Two network topologies demonstrate the methods: the Zachary's karate club social network [24] and the Les Misérables network [14]. The

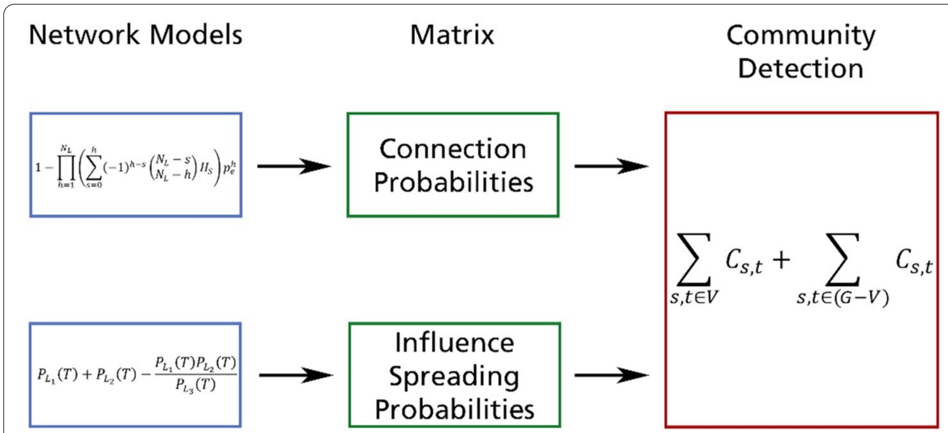


Fig. 2 Two network models and a community detection model based on maximising the value of the objective function in Eq. (8) are used to demonstrate the methodology

Zachary's karate club social network has been used as an example network in several studies in the literature. The Les Misérables network is chosen as an example because a research article investigating building blocks of communities and sub-communities in the Les Misérables network has been published recently in [14]. The set of building blocks discovered in [14] can be compared with the corresponding results of the model proposed in this study. Both example networks have a complex structure from where a set of sub-communities can be detected thus making it possible to compare different models. A third example of an animal social network of 62 bottlenose dolphins [25] has been analysed in [27]. Examples of larger networks, where the model of this study have been used, are in section 'Temporal spreading of networks' and in [21].

The main focus of this study is in the methodology and this is why simple social networks, the Zachary's karate club and the Les Misérables networks are used to demonstrate the method. In addition, we show detailed results provided by the model in order to demonstrate the granularity and different aspects of the model. However, these results are not analysed in detail because such low-level empirical information is not available. Usually, the model predicts the strongest communities accurately, but weaker structures are more sensitive to network models and parameter values.

Applications include discussion about sensitivity to model parameter values, quantitative measures for ranking communities according to their strength, probability of formation and robustness of composition. A new method is proposed for discovering underlying groups of nodes (building blocks) that are usually found together in community structures. The main conclusions are presented in the last section.

Application to the Zachary's karate club social network

W.W. Zachary observed 34 members of a karate club over a period of 2 years [24]. During the study, a disagreement developed between the administrator of the club and the club's instructor. The instructor started a new club, taking 16 members of the original club with him. The karate club social network is pictured in Fig. 3 where line 1 indicates the two partitions after the split of the club with the exception of node 9 who joined the other club. The instructor is node 1 and the administrator is node 34.

Community structure

Zachary's karate club is a social network where we assume that low link weights describe the probability of social influence. In addition, only one community, where all nodes of the network are in one community, is detected with high parameter values. This is not an interesting case in our study. Connectivity and influence spreading probabilities describe different phenomena but they can have some common interpretation in social networks. The two models with the link weight value of 0.05 predict similar rankings for the seven strongest divisions. The link value of 0.05 describes weak social influence and provides a reasonable number of sub-communities. No empirical data or direct observations exist for the link weight values. However, the model predictions calculated with the value of 0.05 [27] for the Zachary's karate club [24] and bottlenose dolphins' [25] social networks agree well with the observations. In a more comprehensive analysis calculations should be performed with a range of parameter values. The compositions of the seven divisions

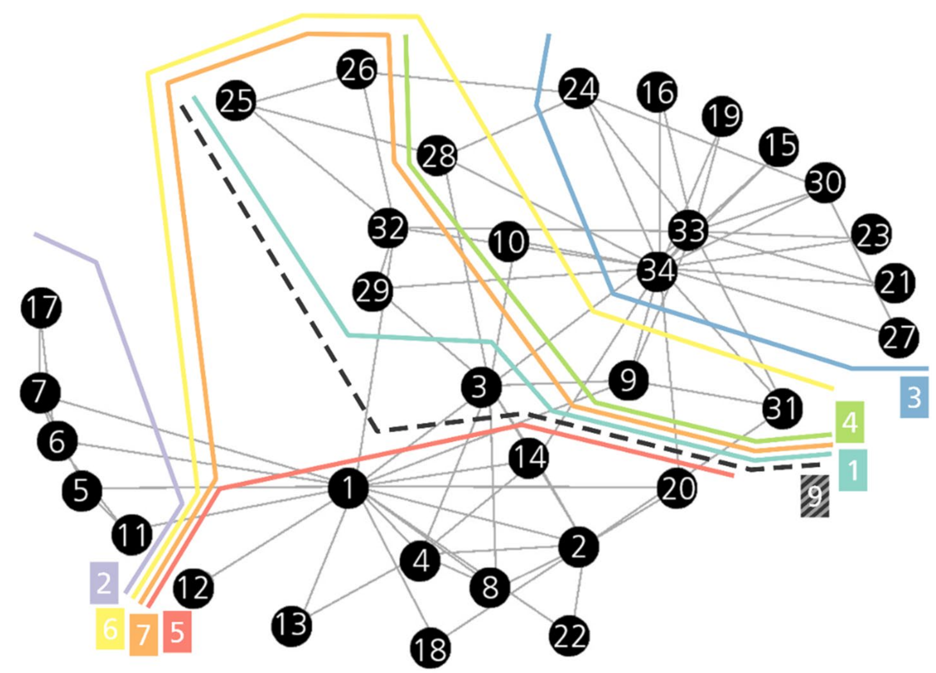


Fig. 3 Zachary's karate club network with divisions indicating detected communities. Divisions 1–7 correspond to lines in Tables 1–3. The additional division 9 is detected in a sensitivity test where link weights have a small random component (see Fig. 8). Sensitivity measures uncertainty in the output of the model in the presence of uncertainty in the link weight values

Table 1 The connectivity and influence spreading models with weights 0.05 predict the same rankings for the seven divisions in Fig. 3. Compositions of the split into two partitions and the number of nodes in each partition are shown

Divisions in Figs. 3 and 4 from connectivity and influence spreading models $(p, w_L = 0.05)$				
id	*#*	*Nodes*	*#*	*Nodes*
id1	16	1-8 11-14 17-18 20 22	18	9-10 15-16 19 21 23-34
id2	5	5-7 11 17	29	1-4 8-10 12-16 18-34
id3	10	15-16 19 21 23-24 27 30 33-34	24	1-14 17-18 20 22 25-26 28-29 31-32
id4	14	9-10 15-16 19 21 23-24 27-28 30-31 33-34	20	1-8 11-14 17-18 20 22 25-26 29 32
id5	10	1-2 4 8 12-14 18 20 22	24	3 5-7 9-11 15-17 19 21 23-34
id6	15	1-4,8,12-14,18,20,22,25-26,29,32	19	5-7 9-11 15-17 19 21 23-24 27-28 30-31 33-34
id7	15	5-7,11,15-17,19,21,23-24,27,30,33-34	19	1-4 8-10 12-14 18 20 22 25-26 28-29 31-32

are listed in Table 1. Table 1 gives both partitions of the divisions, but sometimes we show only the partition with fewer nodes in order to simplify notations.

Figure 4 shows the hierarchical structure of the 14 communities in Fig. 3 and in Table 1. Note that each split has two partitions. Ten different four-level structures are detected in the network under communities id2_29, id3_24 and id5_24. These can be verified as, for example, community id1_18: $\{9-10, 15-16, 19, 21, 23-34\}$ and community id6_19: $\{5-7, 9-11, 15-17, 19, 21, 23-24, 27-28, 30-31, 33-34\}$ are sub-communities of community id2_29: $\{1-4, 8-10, 12-16, 18-34\}$. Two more levels exist below community id1_18 and community id6_19. Colours in Table 1 and in Figs. 3, 4 indicate the two partitions for every seven divisions.

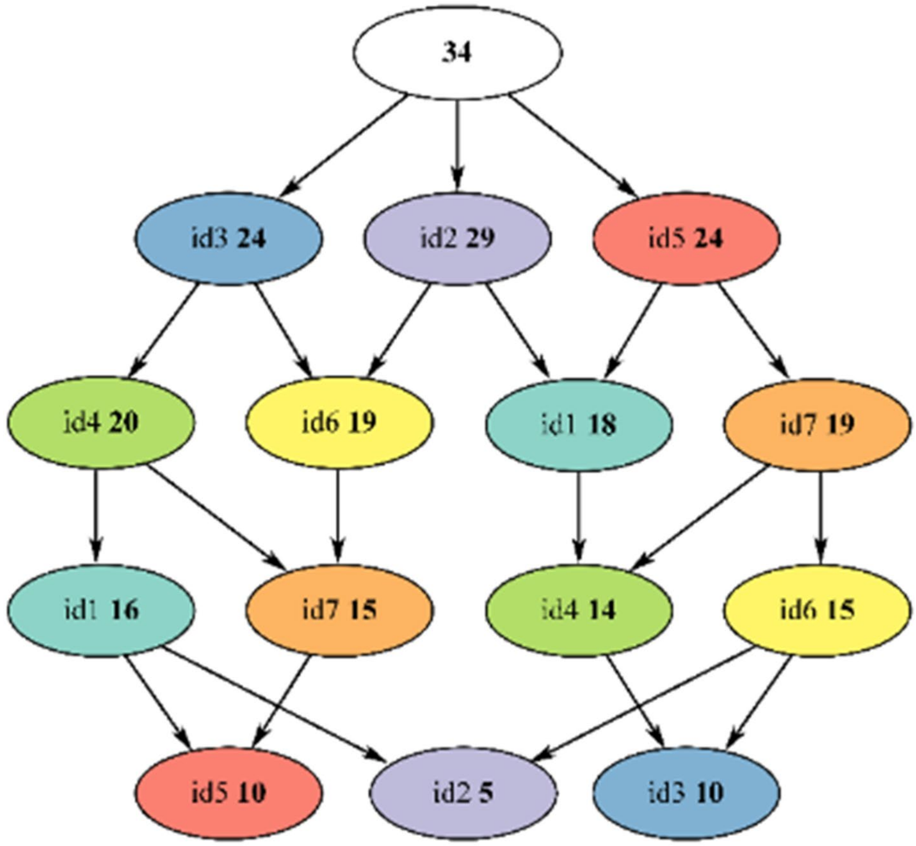


Fig. 4 Hierarchical structure of karate club sub-communities. Seven divisions $id1, \ldots, id7$ detected by the community detection algorithm are indicated by colours (compositions of the 14 communities and sub-communities are listed in Table 1)

Next, we present results of the Zachary's karate club network from the two network models. In Table 2, columns 'A0.05′ and 'A0.1′ are from the network connection model and the other five columns are from the influence spreading model. Nine different solutions for communities are detected. These are lines 1–9 in Tables 2 and 3 (lines 1–7 correspond to id1–id7 in Table 1 and Figs. 3, 4).

The numerical values of the community influence measure of Eq. (8) from the two network models for the nine detected divisions are shown in the left part of Table 2. The corresponding values of statistical community measures are shown in the middle part of the table. The statistical values are probabilities to split into the two communities. These results are simulated by starting from random initial configurations. The second division in line 2 has the highest community measure of Eq. (8) for 'A0.05′ for the network connection model and four influence spreading model calculations with different model parameters.

Table 3 shows the nodes included in the communities. For example, the second line indicates that nodes $\{5, 6, 7, 11 \text{ and } 17\}$ and $\{1, 2, 3, 4, 8, 9, 10, 12, ..., 16, 18, \ldots, 34\}$ are members of the two detected communities. The format of Table 3 is visually useful for small networks, but for larger networks the format of Table 1 is more practical. The last two columns show that the numbers of nodes in the communities are 29 and 5.

Table 2 The values of the community influence measure of Eq. (8) from the two network models for the nine detected divisions are shown in the left part of the table. The corresponding values of the statistical community measures are shown in the middle part of the table. Columns 'A0.05' and 'A0.1' show the results from the network connection model with connectivity probabilities $p = 0.05$ and $p = 0.1$ between neighbouring nodes. Columns 'P0.05' and 'P0.1' show the results from the influence spreading model with influence spreading probabilities $w_L = 0.05$ and $w_L = 0.1$. The next column 'PT0.1′ shows the results during the spreading process at time $T = 0.1$ ($w_L = 1.0$) (all the other columns show results for time approaching infinity $T \to \infty$). Column 'L0.1′ shows the results with the limited path length $L = 1$ ($w_L = 0.1$). Column 'VL0.05′ shows the results with the limited number of visits $V = 1$ on a node during the influence spreading process ($w_L = 0.05, L = 1$). Numerical values in the left, middle and right parts of the table are for strength (Eq. (8)), statistical and robustness measures of the nine divisions on lines 1–9

	Strength (Community Measure)								Statistical Community Measure								Robustness of Community						
id	A0.05	A0.1	P0.05	P0.1	PT0.1	L0.1	VL0.05	id	A0.05	A0.1	P0.05	P0.1	PT0.1	L0.1	VL0.05	id	A0.05	A0.1	P0.05	P0.1	PT0.1	L0.1	VL0.05
1	9.18	24.90	9.42		17.20	21.29	8.76	1	12.96 %	10.03 %	13.11 %		14.20 %	14.93 %	14.74 %	1	14.46	34.31	14.79		27.50	33.50	14.25
2	10.35		10.62		19.24	23.89	9.72	2	10.69 %		10.29 %		10.79 %	10.65 %	11.08 %	2	19.14		19.61		35.63	43.89	18.05
3	8.69		8.91		16.23	20.04	8.25	3	4.65 %		5.13 %		5.22 %	5.27 %	4.72 %	3	12.50		12.76		23.61	28.49	12.19
4	8.72	23.58	8.94		16.33	20.12	8.32	4	2.58 %	0.35 %	2.71 %		3.15 %	2.00 %	3.89 %	4	12.61	29.05	12.88		24.00	28.83	12.48
5	8.58		8.79		16.04	19.70	8.16	5	1.68 %		1.64 %		1.87 %	1.78 %	2.20 %	5	12.06		12.27		22.84	27.14	11.84
6	8.03	21.46	8.22		15.03	18.34	7.67	6	0.38 %	0.09 %	0.37 %		0.41 %	0.31 %	0.43 %	6	9.87	20.56	10.00		18.81	21.69	9.89
7	7.94		8.12		14.84	18.10	7.56	7	0.40 %		0.45 %		0.48 %	0.42 %	0.48 %	7	9.49		9.61		18.03	20.72	9.43
8		24.78						8		0.12 %						8		33.86					
9				29.27				9				8.83 %				9				37.67			

Table 3 Nodes in communities corresponding to lines in Table 2. For example, line 2 means that the five nodes ($N_1 = 5$) 5, 6, 7, 11, and 17 are members of the partition indicated with '0's and the remaining 29 nodes ($N_2 = 29$) are members of the other partition. In the next column '1,010,111' indicates that the division in line 2 is found in computer runs 'A0.05', 'P0.05', 'PT0.1', 'L0.1', and 'VL0.05'. The last two columns show the number of nodes in the two partitions of the network. Three sub-communities (building blocks) are indicated with the orange, purple and blue colours to visualise patterns in this tabular presentation. Note that nodes 14 and 20 (in bold) are not members of their typical sub-community in line 8 (division 8 found only in the network connection model in run 'A0.1′)

id	Nodes 1–34	Found in runs	N_1	N_2
1	1111111100111100110101000000000000	1110111	16	18
2	1111000111011111011111111111111111	1010111	29	5
3	1111111111111100110101001101101100	1010111	24	10
4	1111111100111100110101001100100100	1110111	20	14
5	1101000100011100010101000000000000	1010111	10	24
6	1111000100011100010101001100100100	1110111	15	19
7	1111000111011100010101001101101100	1010111	19	15
8	1101111100111**0**00110**0**01000000000000	0100000	13	21
9	1101111100111100110101000000000000	0001000	15	19

Note, that the computer runs 'A0.1′ and 'P0.1′ have not found the second community, as can be seen also in Table 3 with the second and the fourth zero in '1010111'.

Fewer communities are found with higher link weights and it is even possible that the strongest community calculated with lower link weights is not found or a new combination of nodes emerges. Comparing lines 1 and 9 in Table 3 reveals that the only difference is node 3 moving to the larger partition. This configuration is the only one detected with the higher weight value of $w_L = 0.1$ in the influence spreading model.

Three columns in Table 2 show calculations from the influence spreading model with three different parameters: 'PT0.1′ with the time of spreading $T = 0.1$, 'L0.1′ with the limited path length $L = 1$ ($w_L = 0.1$), and 'VL0.05′ with the limited number of visits $V = 1$ ($w_L = 0.05$) on a node during the influence spreading. These results agree with the basic calculations of 'A0.05′ and 'P0.05′. These results may be different in other more complex network topologies or network configurations.

The measure of robustness, as defined in Eq. (9), provides a method to find nodes who most easily change sides of a division. The strength of divisions and probabilities of division from a random initial state are presented in Table 2. In the right part of Table 2 robustness values have been shown for the seven computer runs and nine divisions of the karate club network. In this case, strength in the left part of Table 2 has the same ranking of divisions 1 – 9 as robustness. As strength and robustness are closely related quantities, this kind of results can be expected. However, rankings of the three different community measures can be different due to the complex structure of networks. Robustness of communities is further discussed in section 'Robustness of community structure'.

Figure 5 shows the closeness centrality values of Eq. (5) from the influence spreading model for the 34 nodes of the karate club network. Bars in the figure are for the link weights 0.05 and 0.1. In general, closeness centrality values are higher for higher

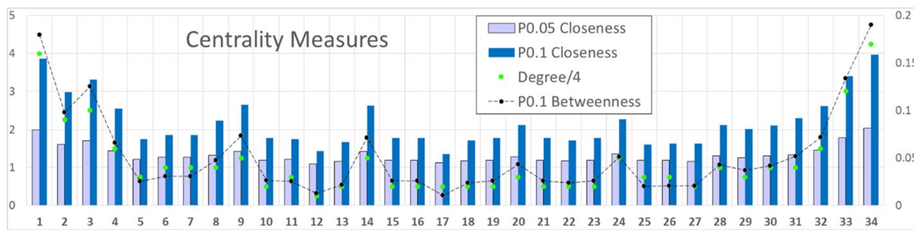


Fig. 5 Closeness centrality values of Eq. (5) for the 34 karate club members for the influence spreading model with link weights 0.05 and 0.1 are shown with bars. Betweenness centrality values of Eq. (7) for the influence spreading model are shown for link weights 0.1. Node degree values divided by 4 are shown for comparison

Table 4 Three divisions found in 100 computer runs of P0.1 with random uniform distribution $U(0.1, 0.01)$ link weights. The last column shows the number of runs where the divisions are found

Nodes (first partition of the division)	Strength	Statistics	Robustness	Count
1–8,11–14,17–18,20,22	28.99	9.00	37.61	52
1–2,4–8,11–14,17–18,20,22	28.81	9.07	37.70	28
9–10,15–16,19,21,23–24,27–28,30–31,33–34	27.27	0.24	31.38	22

link weights but some nodes gain some relative advantage, for example, nodes 1, 2, 3, 4, 8, 9, 14, 20, 24 and 28–34. These nodes are in central positions (not peripheral) in Fig. 3. Consequently, higher link weight values can strengthen communities in central positions of networks. Degree is a measure of interactions between a node's closest neighbours. For example, nodes 9, 14 and 20 have low degrees but relatively high closeness centrality values. In the network of Fig. 3, these nodes are in central positions. In addition, their betweenness values of Eq. (7) with link weights 0.1 are relatively high. Betweenness curve with link weights 0.05 is not included in the figure because it behaves similarly to the degree curve.

Robustness of community structure

Next, we investigate the robustness of communities and sub-communities in the Zachary's karate club network. Table 4 shows one example where strength and robustness are not in the same order for the three divisions found in 100 runs with random uniform link weight distribution.

The case of P0.1 from Table 2 has been selected for more detailed analysis because only one division is found with the link weight value of 0.1 in the influence spreading model. The only difference between divisions 1 and 9 in Table 2 is node 3 (line 9 of Table 3 and line 2 of Table 4). Node 3 in division 1 has moved to the second partition when higher link weights are used. In other words, the second partition has more attraction towards node 3. Sensitivity analysis of 100 computer runs with random uniform link weight distribution shows that division 1 with nodes $\{1-8, 11-14, 17-18, 20, 22\}$ on line 1 in Table 1 also appears in 52 runs. Two weaker divisions in Table 4 are found in 28 and 22 runs.

Figure 6 shows the average robustness values for the link weight value of 0.1 (P0.1) and the results of the sensitivity analysis of 100 computer runs (P0.1 $U(0.1, 0.01)$). Robustness

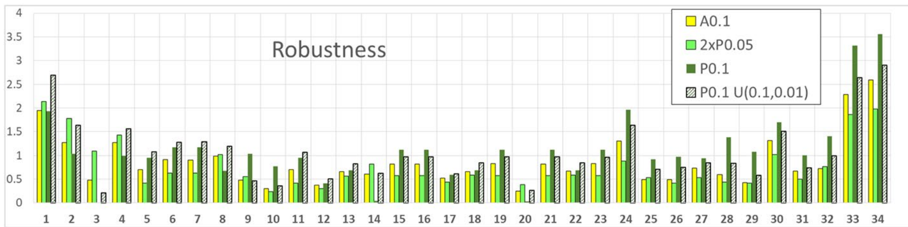


Fig. 6 Robustness values of the 34 karate club members for the influence spreading model with link weights 0.05 (P0.05) and 0.1 (P0.1). Robustness values for the network connection model with probability 0.1 of a functioning link (A0.1). For comparison, P0.1 is computed also for uniformly distributed random link weights $U(0.1, 0.01)$

values for nodes are un-weighted average values over all divisions where the node is a member. In addition, the results for the network connection model with probability values 0.1 and the influence spreading model with link weight values 0.05 are shown. Note that the results are presented for nodes of the network and not for divisions.

Robustness is primarily a measure for studying a node's commitment to its communities, but it can be calculated for an individual node as average over all its subcommunities as in Fig. 6. The results in Fig. 6 can be analysed for every node of the network, but some observations can be made easily. Nodes 1, 33 and 34 are the most robust nodes and also nodes 2, 4, 24 and 30 are robust members of their communities. Differences between the network models and parameter values can be analysed node by node. For example, nodes 33 and 34 are relatively more robust members of their community in the influence spreading model than in the connection model. Both network models give similar results for the robustness of node 1. Comparing the results of the influence spreading model with link weights 0.05 (P0.05) and 0.1 (P0.1) show similar behaviour. However, it is not so clear from the sensitivity analysis of random link weights in Fig. 6. This is one indication that the sensitivity analysis is useful before detailed conclusions can be made.

Nodes 3, 10 and 20 are examples of loosely bound nodes of their communities thus having low robustness values. These nodes appear in several communities and they jump more easily from one community to another when changing model parameter values or network models. Node 3 is one example of this behaviour as discussed earlier. Nodes with low robustness values have high betweenness values if they are in gateway-like positions. This is true for node 3 as can be seen from Fig. 5, but it is not true for node 10, for example. Betweenness and robustness are two different concepts although they are related and correlated in usual network topologies.

Figure 7 goes even deeper in investigating the robustness of communities and their structure. Robustness for selected nodes in the karate club network is shown in Fig. 7 for the seven divisions of Table 2. We study the same case of P0.05 as before. Interesting conclusions can be made about nodes that have high betweenness and low robustness values such as nodes 3 and 9. Node 9 is the one who joined the second partition including node 1. From Fig. 7 we can see that crossing any borders of divisions 1, 4 and 7 are easy for node 9. It would have been even easier for node 3 to change side because crossing the border of division 1 has a low influence on the strength of the division.

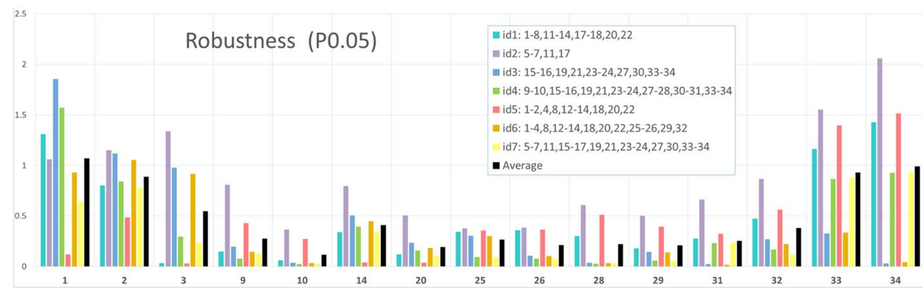
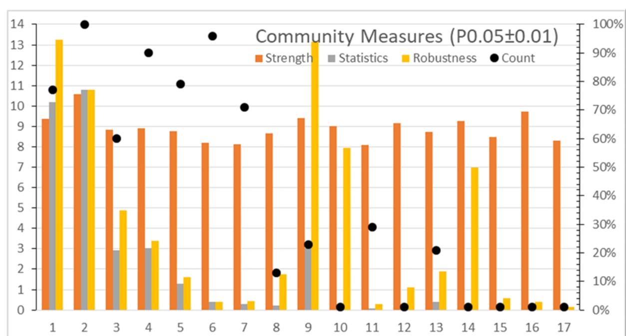


Fig. 7 Robustness values of selected members of the karate club network for P0.05 in Fig. 6. Values are shown for the seven strongest divisions



1	1-8,11-14,17-18,20,22
2	5-7,11,17
3	15-16,19,21,23-24,27,30,33-34
4	9-10,15-16,19,21,23-24,27-28,30-31,33-34
5	1-2,4,8,12-14,18,20,22
6	1-4,8,12-14,18,20,22,25-26,29,32
7	5-7,11,15-17,19,21,23-24,27,30,33-34
8	1-4,8,12-14,18,20,22
9	1-2,4-8,11-14,17-18,20,22
10	10,15-16,19,21,23-30,32-34
11	5-7,9,11,15-17,19,21,23-24,27,30-31,33-34
12	9,15-16,19,21,23-24,27-28,30-31,33-34
13	9,15-16,19,21,23-24,27,30-31,33-34
14	15-16,19,21,23-30,32-34
15	1-4,8-10,12-14,18,20,22,31
16	1-2,4-8,11-13,17-18,22
17	5-7,10-11,15-17,19,21,23-24,27-28,30,33-34

Fig. 8 Results of a sensitivity analysis from 100 computer runs of the karate club network. The influence spreading model is used with random uniformly distributed $U(0.05, 0.01)$ link weights. The figure illustrates the values of strength, statistical measure, robustness measure and the number of runs where the 17 divisions in the table are detected

Results of a sensitivity analysis of 100 computer runs for the influence spreading model with uniformly distributed random link weights between 0.04 and 0.06 are shown in Fig. 8. Altogether 17 divisions are detected and they are listed in the table. Strength, statistics and robustness measures are shown as bars and the number of runs where the divisions are found is shown as dots. Statistical and robustness measures are low for divisions 6 and 7. The interpretation is that the probability of formation is low and also their robustness is weak. These two measures are correlated, but Fig. 8 illustrates that some divisions can have a low probability of formation but a high robustness value. Division 9 is a good example of a robust division. This composition is uncovered by the sensitivity analysis and we have added this division in Fig. 3 with a dotted line. This is the same split as division 1 with one exception of node 3. This finding is in good agreement with our previous discussion about node 3. Note that division 9 is detected with higher link weights 0.1 in Table 2 as the only optimal solution of Eq. (8).

Application to the Les Misérables network

Les Misérables is a French novel by Victor Hugo published in 1862. The social network of fictitious characters in the novel has been studied widely in community detection literature. Recently, the Les Misérables network has been used to study the consistency of optimal community structure and the idea that sub-communities correspond

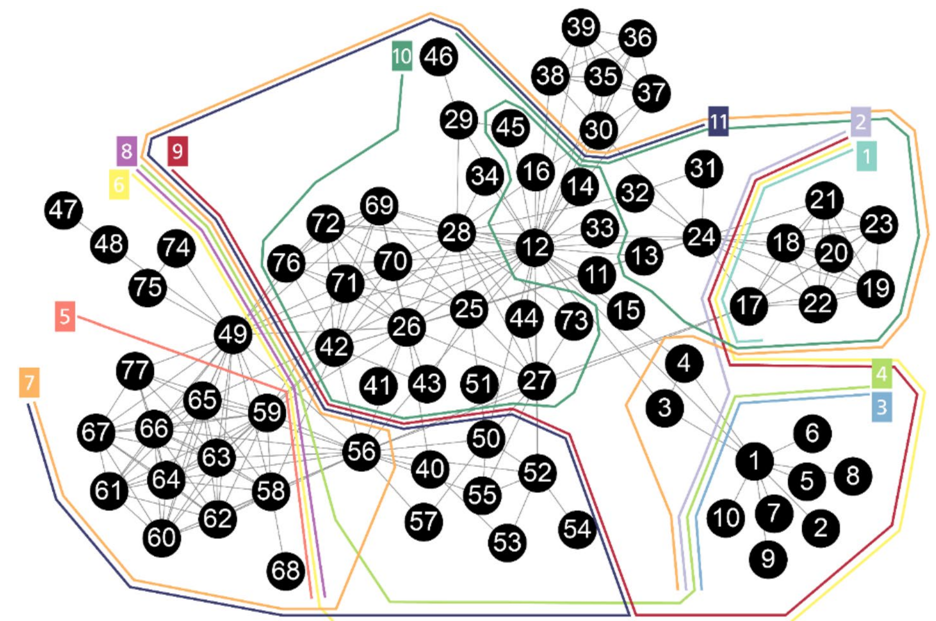


Fig. 9. 11 strong divisions of the Les Misérables network. Borders of the divisions show building blocks of community structure

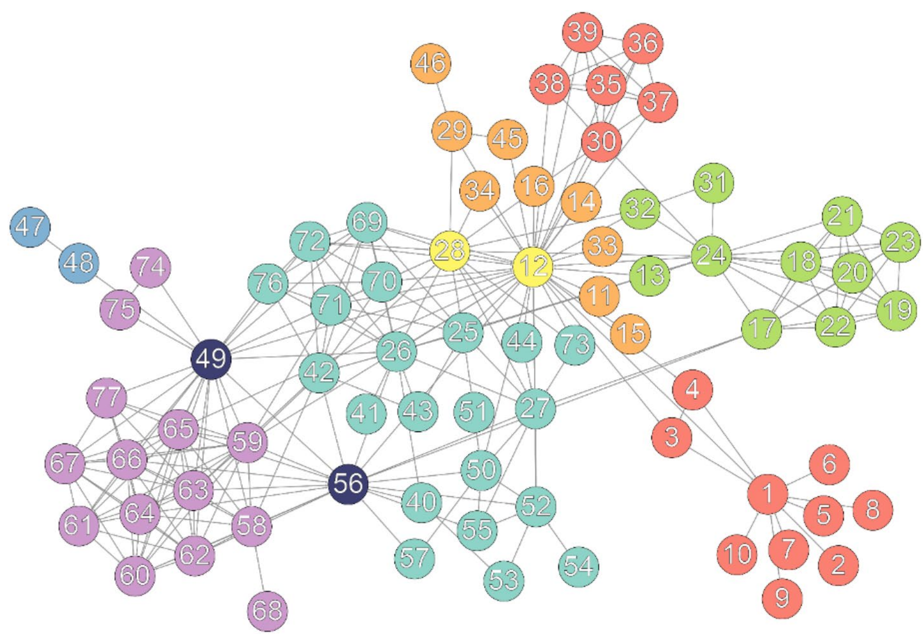


Fig. 10 Building blocks of the Les Misérables network discovered in [14]

to arrangements of a set of underlying building blocks [14]. An information-theoretic method was used to discover building blocks from the social network of Les Misérables. Those results can be compared with the findings of this study.

Community structure

The Les Misérables social network consists of 77 nodes illustrated in Figs. 9 and 10. Figure 9 shows 11 highly ranked divisions from the 201 divisions found in 100 computer runs. The influence spreading model was used with uniformly distributed random link weights between [0.04, 0.06]. The ranking of the 11 divisions is the same for the strength of divisions and for the probability of formation and very close to the ranking of robustness measure. Figure 12 shows the values of the community measures for 50 highly ranked divisions where the 11 divisions of Fig. 9 are the first 11 values.

We propose a new method for discovering building blocks from a network by using borders between divisions provided by a community detection method. Figure 10 shows the corresponding results in [14]. Building blocks discovered in this study are very similar. One difference is that the building block of seven nodes $\{40, 50, 52, \ldots, 56\}$ is not detected in [14]. In Fig. 9, this group of nodes is separated by several boundaries and it can be regarded as a building block. There are other minor differences, for example, nodes $\{29, 34, 46\}$ are not grouped with nodes $\{11, 14, 15, 16, 33\}$ in our model. Central nodes like 12, 28, 49 and 56 are also members of sub-communities in Fig. 9.

Figure 11 illustrates the hierarchical structure of the Les Misérables network. We have included only half of the sub-communities detected in the influence spreading model with link weights 0.05. The complete graph of all hierarchical relationships is similar but larger. In Table 6 (Appendix), smaller partitions (sub-communities) are shown on the left and the graph in Fig. 11 shows partitions having node 1 as a member. Figure 11 can be compared with Fig. 4 with the complete hierarchical structure of the karate club network of 34 nodes. Only 6 of the 11 most highly ranked divisions 1, 2, 3, 4, 5 and 8 are detected

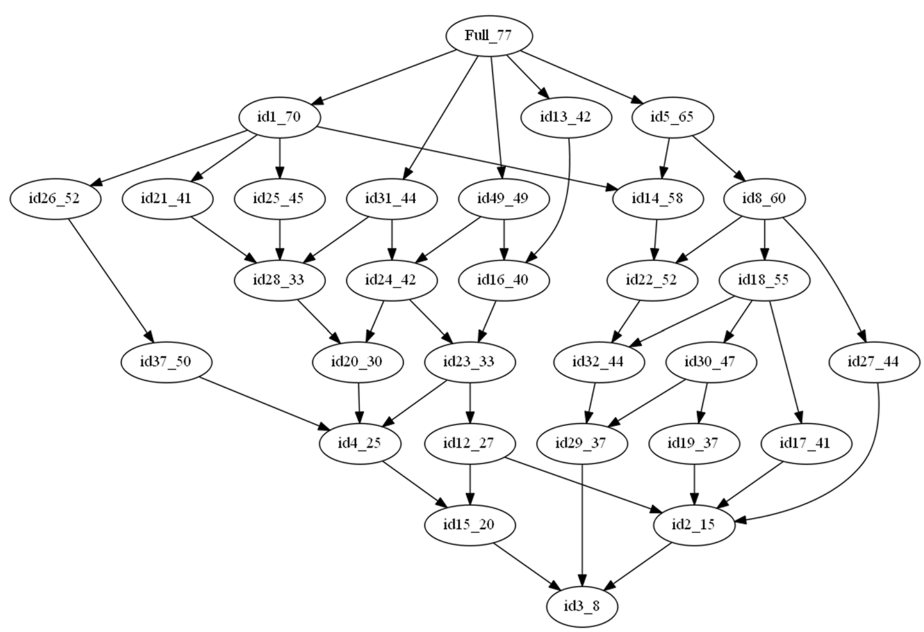


Fig. 11 Hierarchical structure of 29 sub-communities of the Les Misérables network (only half of the hierarchical structure is shown). Divisions indicated by id-numbers are listed in Table 6 in the Appendix with bold ids. Results are for the influence spreading model with link weights 0.05. Sizes of sub-communities are also displayed in the graph

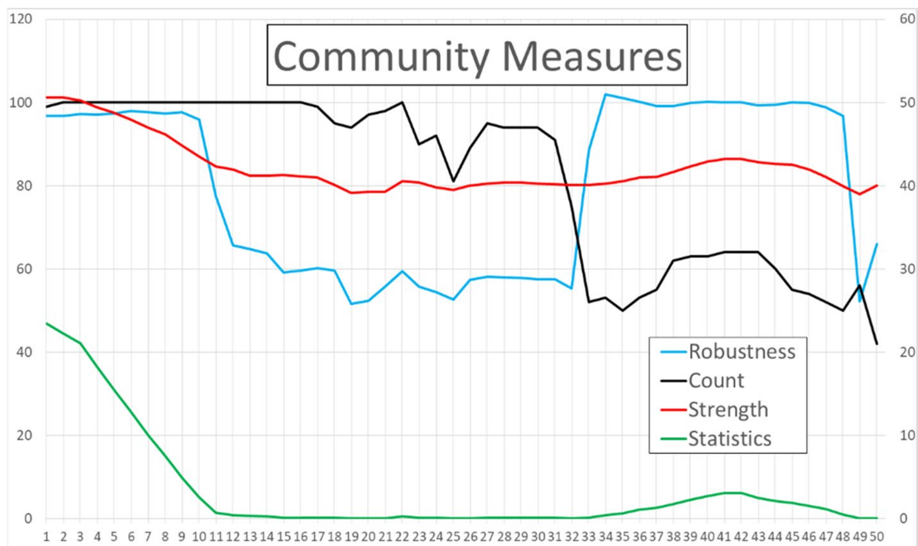


Fig. 12 Values of the community measures for the Les Misérables network. Compositions of the divisions are listed in Table 6 in the Appendix. Fifty highly ranked divisions of the 201 division detected in 100 computer runs with uniformly $U(0.05, 0.01)$ distributed random link weights are shown in the figure

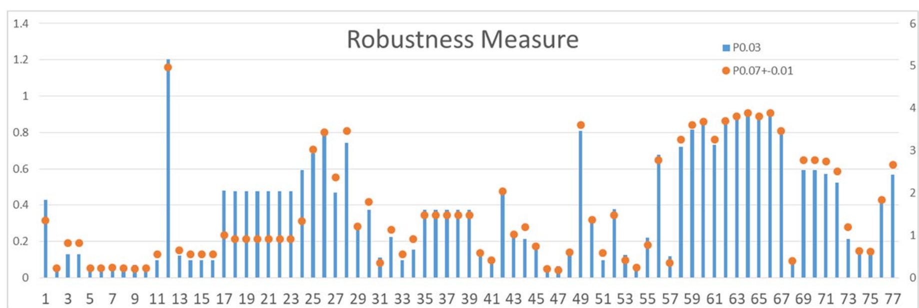


Fig. 13 Robustness values of nodes in the Les Misérables network from the influence spreading model with link weights 0.03 and uniformly distributed random link weights between [0.06, 0.08]

with link weights 0.05. On the other hand, divisions 12–32, 37 and 49 are found. This is an indication that more sub-communities can be detected by varying link weights.

Robustness of community structure

Figure 12 illustrates the community measure values of strength of division (Eq. (8)), probability of formation (Statistics), robustness of composition (Eq. (9)) and the number of computer runs (Count) where the divisions are detected. Compositions of the most highly ranked divisions 1 – 11 are the corresponding lines in Table 6 in the Appendix. Also, some divisions 39 – 46 have good community measure values except that they are only detected in 65% of runs. These divisions are like division 9 in Fig. 8 of the karate club network.

In Fig. 13, the robustness of nodes in the Les Misérables network are presented for link weights 0.03 and uniformly distributed random link weights between [0.06, 0.08].

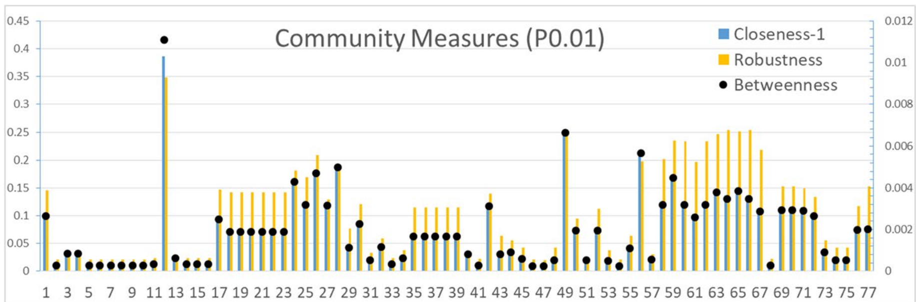


Fig. 14 Closeness and betweenness centrality measures compared with robustness values of nodes in the Les Misérables network. The influence spreading model is used with link weights 0.01. The detailed composition of the robustness values is shown in Fig. 15

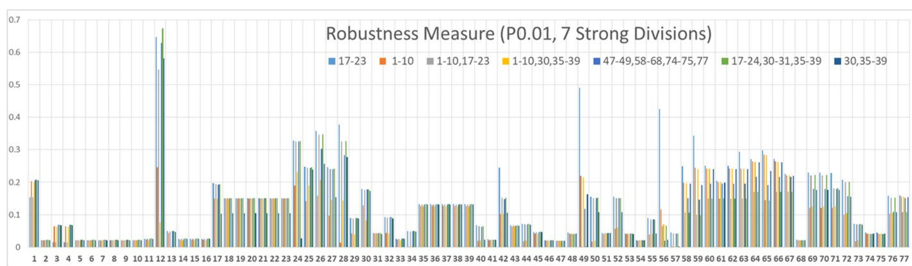


Fig. 15 Robustness values of nodes for seven strong divisions in the Les Misérables network. The corresponding average values of robustness are shown in Fig. 14

Results are similar except that nodes 17–24 have relatively higher robustness values for lower link weights 0.03. Figure 14 compares closeness centrality, betweenness centrality and robustness of nodes in the Les Misérables network for link weights 0.01. Nodes 12, 28, 49 and 56 are examples of nodes with relatively high betweenness compared to centrality. Robustness values of nodes 17–23, 35–39, 58–67 and 77 are relatively high when compared with closeness and betweenness centrality values. These nodes are members of peripheral sub-communities in Fig. 9. Figure 15 shows a detailed view of the average information of robustness in Fig. 14 for seven strong divisions. Figure 7 is the corresponding figure for the karate club network.

Temporal spreading on networks in the influence spreading model

Spreading dynamics and diffusion in complex networks [15] have been studied for example in [19, 20, 28, 29]. Next, we discuss how we can use different temporal spreading distributions in our methodology. Temporal distributions are incorporated in the form of a probabilistic survival distribution function. Instead, empirical numerical values can be used if the spreading process does not obey a known theoretical probability distribution.

We assumed in section 'Influence spreading model' that social networks are in an equilibrium state. In the influence spreading model, this is accomplished by letting time T approach infinity. This was an appropriate assumption for community detection in situations where social influence has been spreading for a long time in the network. In the following sections, we demonstrate how time-dependent spreading phenomena can be studied with the model. In a finite time horizon spreading has not reached all nodes in the network structure with probability one. In that sense, the influence spreading process is still going on in the network structure. All centrality and community measures of previous sections can be calculated as functions of time. For example, the form of Eq. (8) of the spreading probability is valid as such with time-dependent variables. Community structure is also time-dependent. Table 2 shows that similar results for low time values with high link weights (columns 'PT0.1′') and high time values with low link values (columns 'P0.05′') can be expected. However, in complex network topology, there is no simple relation between spreading time and link weights. Also, the ranking of sub-communities can change as a function of time. [21]

In previous sections, the focus has been more structural, whereas in this section we study also temporal development. The structural view takes into account the topology of the network with link and node weights, and the temporal view studies also time-dependent development of spreading processes. Both of the views are built on a network topology with node and link weights but in the previous sections, we have eliminated the time variable by investigating only equilibrium states of networks. In the model, link and node weights along a path of length L are multiplied together and denoted by W_L, and survival functions describing temporal spreading over L links are denoted by $S_L(T)$. Both structural and temporal aspects can be studied by combining these two factors as $W_L S_L(T)$. This holds for a single path between two nodes. The main idea of the spreading model is in the technique of combining multiple paths between node pairs in a network. The method is presented in section 'Influence spreading model' and in more detail in [21].

In what follows, we will first describe the Poisson and e-mail forwarding survival distribution functions that are used in describing the spreading dynamics and diffusion in complex networks [19, 20, 28, 29]. After that, a numerical example of calculating time-dependent values of the influence spreading matrix.is provided. We will then proceed by applying our influence spreading model to three network structures, Zachary karate club, Facebook and Enron networks for each of the mentioned survival distribution functions.

Poisson and e-mail forwarding survival distribution functions

In this study, two different temporal spreading distributions demonstrate the modelling with three real-world social networks. The Poisson distribution describes random response time and the e-mail forwarding distribution describes the process of receiving and forwarding messages. Spreading processes are modelled on constant topological network structure. The examples show that the Poisson temporal distribution is more efficient for spreading at low time values for short path lengths and the situation is reversed for high time values and longer path lengths.

In this context, temporal probability distributions describe spreading probabilities starting from a source node via links to other nodes in the network. Survival distribution function $S_L(T)$ provides the probability of temporal spreading via L links at time T. Function $S_L(T)$ is expressed in terms of probability distribution function $F_L(T)$ as $S_L(T) = 1 - F_L(T)$.

The Poisson distribution is a discrete probability distribution that expresses the probability of giving a number of events occurring in a fixed interval of time if these events occur with a known constant rate and independently of the time since the last event. In this context, the Poisson process describes a process where spreading via successive links occurs randomly as a function of time. For the Poisson distribution $S_L(T)$ is

$$S_L(T) = 1 - \sum_{z=0}^{L-1} e^{-\lambda T} \frac{(\lambda T)^z}{z!}, (S_0 = 1). \tag{10}$$

Figure 16 shows Poisson distribution survival functions for path lengths $L = 1, \ldots, 20$ and time values $T = 1, \ldots, 10$.

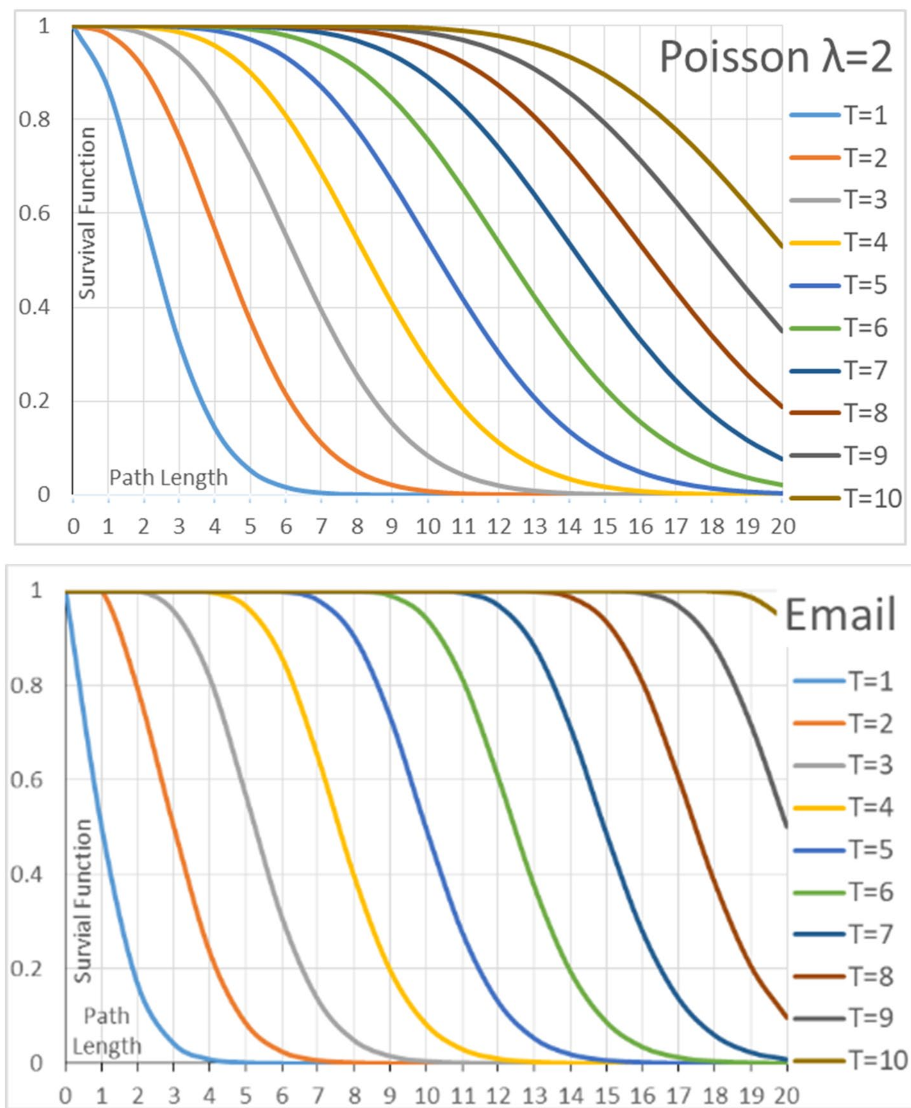


Fig. 16 Survival distribution functions for the Poisson and e-mail forwarding distribution functions as a function of path length for time values $T = 1, \ldots, 10$. The event rate parameter value of $\lambda = 2$ for the Poisson distribution has been used to get comparable results with these path lengths

The e-mail forwarding distribution describes a typical process of receiving and forwarding e-mails or messages. In the model, the source node sends one e-mail in a time unit and all other nodes in the network check their mailboxes once in a time unit and forward the received e-mail if they had received it before the checking time, or the e-mail stays in the mailbox waiting for the next time unit. Nodes send messages and check their mailboxes independently in uniformly distributed random time points. The program code describing the process of receiving an e-mail message once in a time unit and forwarding it once in a time unit is shown in the Appendix.

The two survival distribution functions have different characteristics at short path lengths and at long path lengths. At time $T = 1$ the Poisson probability of reaching nodes at path lengths 1–6 is high compared to the e-mail forwarding probability. Roles are changed with higher path lengths. For example, at time $T = 10$ the spreading probability over path length $L = 20$ for the Poisson distribution is 53% and for e-mail forwarding it is 94%.

Figure 17 shows differences between Poisson and e-mail forwarding survival distribution functions. Curves in Fig. 17 show clearly that Poisson processes dominate e-mail forwarding processes at low time values and short path lengths of $L \leq 10$. Conversely, at high time values and longer path lengths of $L > 10$ e-mail forwarding processes dominate. Note, that the point on the X-axis, where the roles exchange, depends on the value of Poisson distribution's event rate parameter λ.

This kind of process characteristics can have intricate impacts on how spreading processes perform on complex network topology. At low path lengths, the degree of a source node determines how the spreading process starts. The model suggests that Poisson processes are efficient for spreading to neighbouring nodes. On the other hand, spreading processes to distant nodes at high path lengths can get an advantage from using e-mail type of delivery methods.

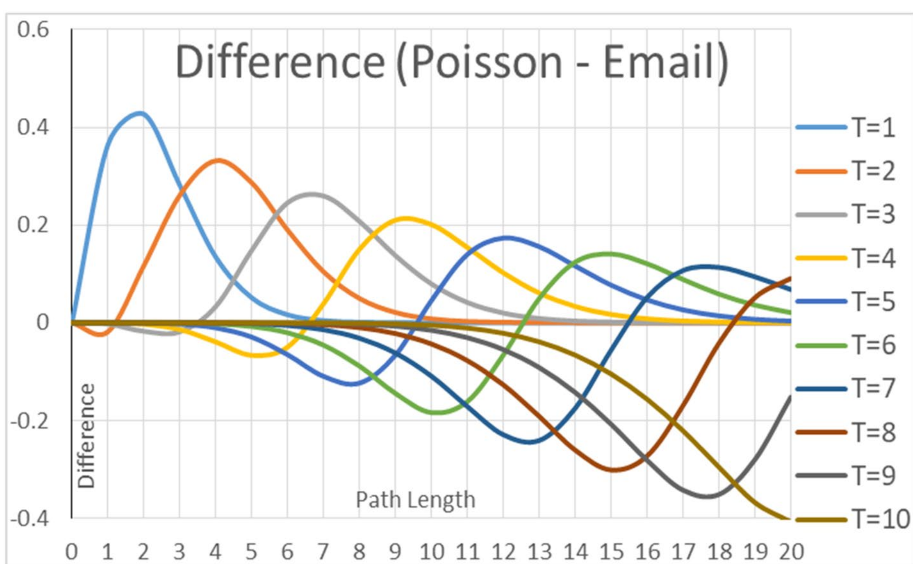


Fig. 17 Difference between Poisson and e-mail forwarding survival distribution functions as a function of path length for time values $T = 1, \ldots, 10$.

A numerical example of calculating time-dependent values of the influence spreading matrix

In this section, we demonstrate how the values of the influence spreading matrix are calculated [21]. We calculate the value of $C_{1,33}(T)$, which describes the probability of influence spreading in time T from node 1 to node 33 in Fig. 3. In order to limit the number of possible paths, we consider path lengths $L \leq 3$ and self-avoiding paths, meaning a node can only appear once in a path. There are eight possible paths which we categorise into four groups. The four groups, denoted by I, I, III and IV, have three, two, two and one different paths correspondingly: $I: 1-4-3-33,\ 1-8-3-33,\ 1-20-34-33,\ II: \{1-2-3-33,,\ 1-2-31-33\}\{1-14-3-33,\ 1-14-34-33\},\ III: \{1-3-33,,1-3-9-33\}$ $\{1-32-33,\ 1-32-34-33\},\quad IV: \{1-9-33,\ 1-9-3-33,\ 1-9-31-33,\ 1-9-34-33\}$. Curly brackets indicate the paths that we are going to combine by using the rules explained in section 'Influence spreading model'.

In Table 6, the numerical values of survival functions $S_L(T)$ for path lengths $L = 1, 2, 3$ are calculated from Eq. (10) for the Poisson distribution. We assume that delays in the influence spreading process are caused by nodes of the network and the role of links is only to pass on information. Different time distributions can be incorporated in the model by using a corresponding survival function instead of Eq. (10). For example, the e-mail forwarding distribution in section 'Temporal spreading distributions' is one alternative for describing time delays in the spreading process.

The three paths in group I all have unique beginnings. We assume that the probability of spreading via a link is w and the probability of spreading via a node is one. Because there are three links, the probability of spreading is $w^3S_3(T)$, where $S_3(T)$ describes the probability of transmitting the influence via three links in time T, for example from node 1 to node 4, from node 4 to node 3 and from node 3 to node 33. The probability of spreading via one of these paths is denoted by $P_I(T)$. Formulas for $P_{II}(T)$, $P_{III}(T)$ and $P_{IV}(T)$ show explicitly how spreading via overlapping paths are calculated as a function of time T. In all these three cases only one link is shared at the beginning of the paths. The fourth case denoted by $P_{IV}(T)$ uses the results of $P_I(T)$, $P_{III}(T)$ combining the four paths in group IV.

The probability of spreading $C_{1,33}(T)$ is calculated by the standard probabilistic formula for mutually non-exclusive events [21]. Note that the degree of node 1 is 16 but only 8 of the paths originating from node 1 reach node 33 within path lengths $L \leq 3$. Numerical values of $P_g(T), g = I, II, III, IV$ and $C_{1,33}(T)$, $T = 1, 2, \infty$ for $w = 0.05, 0.25, 0.5,$ and0.75 are documented in Table 6. Figure 18 shows the time-dependence between $0 \leq T \leq 5$ of spreading probabilities for link weights $w = 0.25, 0.5,$ and 0.75. Black curves are for the Poisson distribution in Eq. (10) with $\lambda = 1$ and blue curves are for the e-mail forwarding distribution calculated from the algorithm that is provided in the Appendix. Formulas for the four independent groups of paths and the final result $C_{1,33}(T)$ in Eq. (11) are the following:

$$P_I(T) = w^3S_3(T),$$

$$P_{II}(T) = w^3S_3(T) + w^3S_3(T) - \frac{w^3S_3(T)w^3S_3(T)}{w^1S_1(T)} = 2w^3S_3(T) - \frac{w^5S_3^2(T)}{S_1(T)},$$

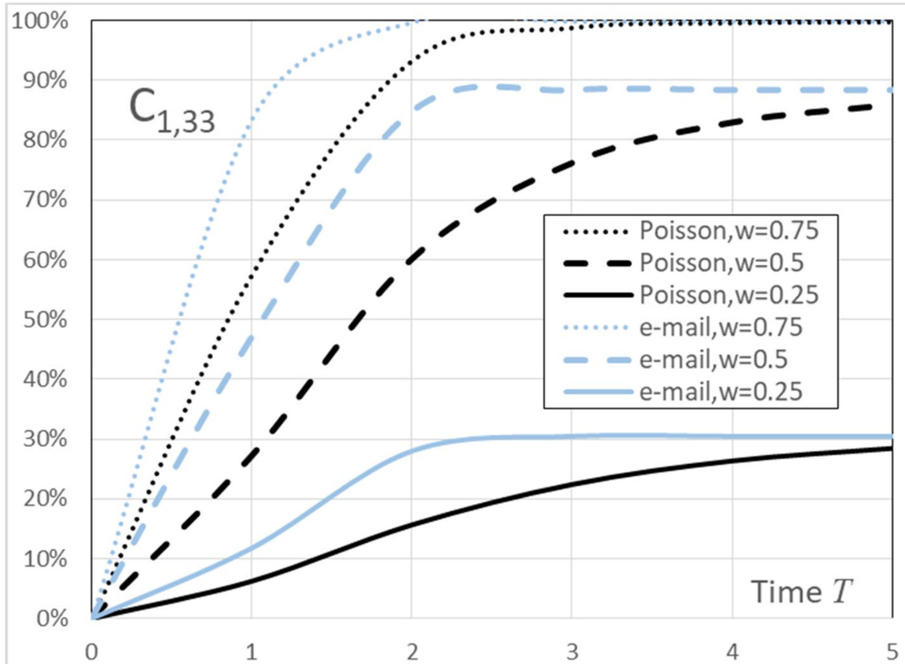


Fig. 18 Influence spreading probabilities $C_{1,33}(T)$ from node 1 to node 33 as a function of time T for link weight values $w = 0.25, 0.5, 0.75$. Black curves are for the Poisson distribution and blue curves for the e-mail forwarding distribution

$$P_{III}(T) = w^3S_3(T)+w^2S_2(T)-\frac{w^3S_3(T)w^2S_2(T)}{w^1S_1(T)} = w^3S_3(T)+w^2S_2(T)-\frac{w^4S_3(T)S_2(T)}{S_1(T)}$$

$$P_{IV}(T) = P_I(T)+P_{III}(T)-\frac{P_I(T)P_{III}(T)}{w^1S_1(T)}+P_I(T)-\frac{\left(P_I(T)+P_{III}(T)-\frac{P_I(T)P_{III}(T)}{w^1S_1(T)}\right)P_I(T)}{w^1S_1(T)},$$

$$C_{1,33}(T) = 1-(1-P_I(T))^3(1-P_{II}(T))^2(1-P_{III}(T))^2(1-P_{IV}(T))^1. \quad (11)$$

The numerical values in Table 5 are for the influence spreading model (see section 'Influence spreading model') and the Poisson survival distribution in Eq. (10). The corresponding results in Table 5 in the case of the e-mail forwarding distribution can be calculated by using the values of $S_L(T)$ for the e-mail forwarding distribution instead of the Poisson distribution. On the other hand, the network connection model could be used as the network model to describe a spreading process in which an uninterrupted operational connection is needed between source and target nodes. In this case, the temporal survival distribution function would be implemented in the network connection model (see section 'Network connection model'). Thus, three aspects can be combined in our methodology: network model, temporal spreading distribution and community detection algorithm. These modelling decisions depend on the process characteristics of the spreading process under investigation [18].

Table 5 Numerical values of $S_L, L = 1, 2, 3$ and $P_i(T), i = I, II, III, IV$ and $C_{1,33}(T), T = 1, 2, \infty$ for $w = 0.05, 0.25, 0.5, 0.75$. Poisson distribution has been used for describing the time delays of the spreading process

	$T=1$	$T=2$	$T=\infty$
$S_1(T)$	0.63212	0.86467	1.00000
$S_2(T)$	0.26424	0.59399	1.00000
$S_3(T)$	0.08030	0.32332	1.00000
$P_I(T)$	0.00001004	0.00004042	0.00012500
$P_{II}(T)$	0.00002007	0.00008079	0.00024969
$P_{III}(T)$	0.00067043	0.00152401	0.00261875
$P_{IV}(T)$	0.00069008	0.00160196	0.00285536
$C_{1,33}(T), w=0.75$	57.512%	93.298%	99.912%
$C_{1,33}(T), w=0.5$	27.188%	60.090%	88.299%
$C_{1,33}(T), w=0.25$	6.239%	15.658%	30.465%
$C_{1,33}(T), w=0.05$	0.20997%	0.49243%	0.89381%

Applications of temporal spreading on three empirical social networks

We demonstrate the model with three empirical social networks: Zachary's karate club [24], a Facebook network [30], and the Enron e-mail network [30]. These networks represent small, intermediate and large social networks. The network topology of the Zachary's karate club social network with 34 nodes is pictured in Fig. 3. The Facebook network has 4039 nodes and the Enron e-mail network has 36,692 nodes.

Low parameter values for link weights are used for describing influence spreading [21]. Link weights are assumed to be low for one event of an influence attempt because, in a normal social context, the probability of convincing a person to change his or her opinion is low. For larger networks, the influence is expected to be even lower because of less cohesion with more different social groups.

Figure 19 shows spreading results for the karate club network with two different link weights $w = 0.05$ and $w = 0.005$. Results are presented for the e-mail forwarding distribution and Poisson distribution with the event rate parameter value of $\lambda = 2$. For both link weight values, the Poisson spreading process is more effective than the e-mail forwarding process for time values $T < 1.5$. For time $T > 1.5$ the situation is reversed. For higher link weights ($w = 0.05$) the spreading process is accelerating for later time values, and for lower values ($w = 0.005$) the process is more linear until the usual saturation effects take over.

The expected number of influenced nodes is shown on the Y-axes of the figures. This is computed by taking one node, at a time, in the network and assuming that the spreading process is initiated with probability 1.0 at time $T = 0$ from that node. Figures 19 and 20 show average results over all source nodes in the network. For example, the influence of the Poisson process has spread, on average, to 7.2 nodes with the link weights of $w = 0.05$ at time $T = 1$.

Figure 20 shows the expected number of influenced nodes when spreading is initiated from one of the nodes 1–34 (indicated in the figure) for the Poisson distribution $w = 0.05$. The results agree with the actual situation in the karate club. The instructor of the club is node 1 and the administrator is node 34. In the same figure, the

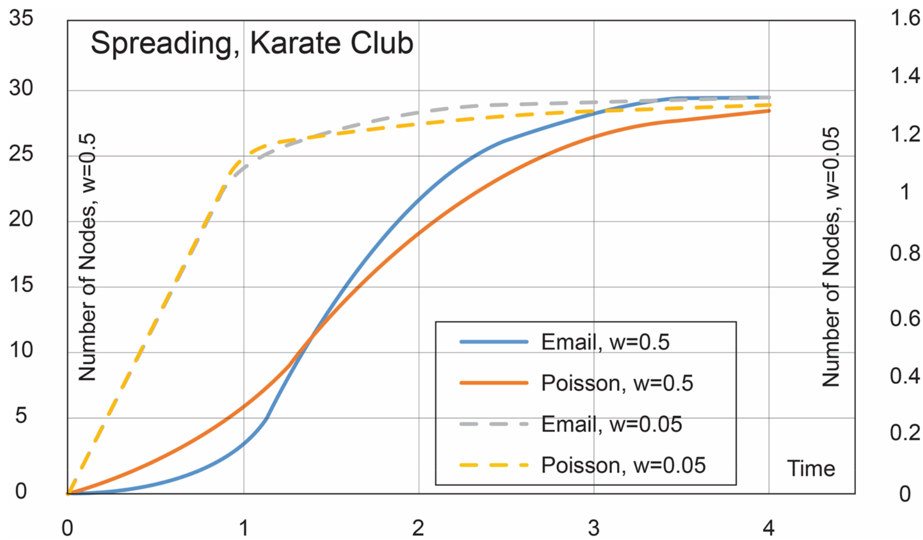


Fig. 19 Spreading in the Zachary's karate club network for the e-mail forwarding process and the Poisson process. Y-axis on the left shows the expected number (normalised) of influenced nodes for $w = 0.5$. Y-axis on the right shows the corresponding values for $w = 0.05$.

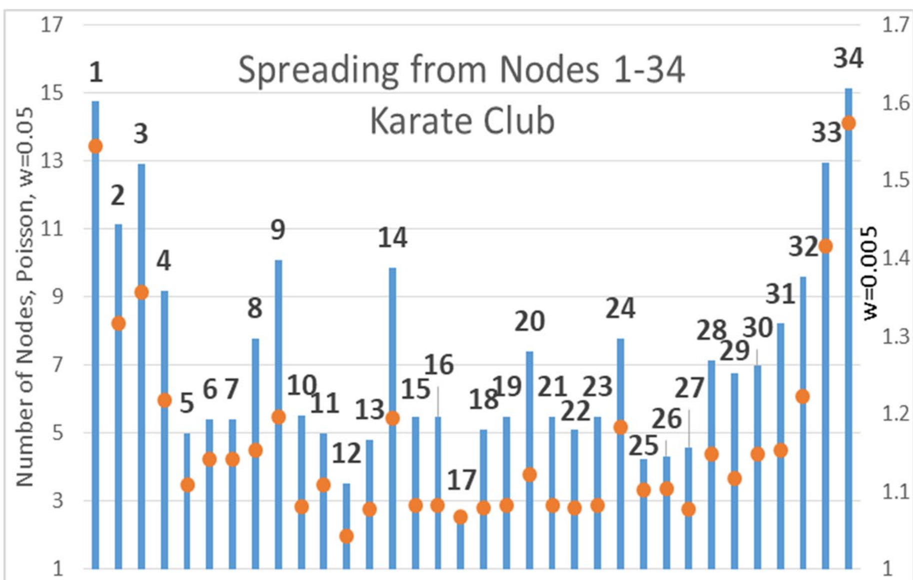


Fig. 20 Spreading from individual nodes 1–34 to other nodes (including the node itself) for the equilibrium state when time approaches infinity. Expected number (normalised) of influenced nodes for $w = 0.05$ on the left Y-axis and for $w = 0.005$ on the right. Numerical values for $w = 0.05$ are shown as bars and for $w = 0.005$ as circles

numerical values for the Poisson distribution $w = 0.005$ are shown as circles. The expected values of the number of influenced nodes are very low for $w = 0.005$.

Rankings for $w = 0.005$ are very close to the corresponding rankings for $w = 0.05$, but they are not the same. For example, node 27 has a higher ranking for $w = 0.05$ than for $w = 0.005$. Node 17 has a relatively higher number of influenced nodes for $w = 0.005$.

These are both peripheral nodes, but they have different accessibility to central nodes. Node 27 has a link to central node 34, but node 17 has only connections to other peripheral nodes. Node 17 has not as much advantage of more active nodes ($w = 0.05$) than node 27. We can see from Fig. 20 that nodes 8, 9, 14, 20 have particularly favourable locations in the network. They have short distances to the most central nodes 1, 33 or 34, and have gateway roles between other nodes in the network. This is a transient phenomenon because, at later time points, influence has propagated more evenly to all nodes in the network.

The Facebook [30] and Enron networks [30] are examples of larger empirical networks. Figure 21 shows spreading results for these networks with two different link weights $w = 0.005$ and $w = 0.0005$. The results are for total sums from all source nodes because normalised values would be very small. Normalised values can be calculated by dividing the results by the number of nodes in the network. The same phenomenon of crossing curves at time $T = 1.5$ can be seen for both link weights. Again, the Poisson distribution with the event rate parameter value of $\lambda = 2$ is efficient at the beginning of the spreading process and vice versa at later time points. At time $T = 3$ ($w = 0.005$) the e-mail forwarding distribution has reached about 400 nodes more in the Facebook network. With lower link weights ($w = 0.0005$), the difference is less because of weaker spreading power.

Conclusions

In this article, we propose a set of methods for modelling community structure and temporal spreading on complex networks. Models of this study can be used to study community or module structure and temporal spreading in social, biological and technological networks. The community detection method is based on separating the network model from the community detection model. Different network models can

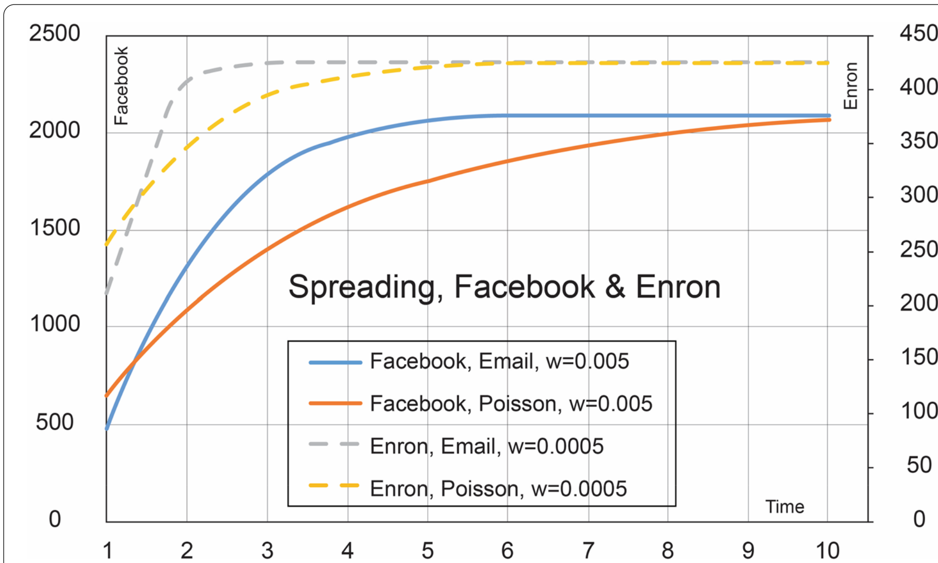


Fig. 21 Spreading in the Facebook and Enron networks for the Poisson process and the e-mail forwarding process. Y-axis on the left shows the expected number (un-normalised) of influenced nodes for the Facebook network. Y-axis on the right shows the corresponding values for the Enron network

be used to provide input for the community detection algorithm. We demonstrate this by the standard network connection model and an influence spreading model. Different temporal distributions can be incorporated into the influence spreading model or the network connection model by using a desired probabilistic survival distribution function.

Communities and sub-communities are identified by local maxima of a quality function that measures the internal strength of two partitions in the network. The quality function used in this study is the sum of probability matrix element values describing interactions between pairs of nodes in both of the two partitions. Cross terms are not included, but the model allows influence via paths that go through the other partition. Weak interactions between nodes produce more solutions than strongly connected networks. By varying link weights, it is possible to get an understanding of the landscape of local maxima and to identify structure in networks. As a new application, we propose that the community detection method can be used as a tool for discovering underlying groups of nodes that are usually found together in community structures. Communities and sub-communities correspond to different combinations of these building blocks.

We present different approaches for measuring and ranking communities. This is an important topic because the method usually gives several solutions and highly ranked solutions are candidates for real-life communities. Three different quality measures are presented for the strength of a community, the probability of formation of a community and the robustness of formation of a community. These measures are correlated, but they represent different views of evaluating communities.

We demonstrate the use of temporal spreading distributions with the Poisson distribution and an e-mail forwarding distribution. The e-mail forwarding distribution is defined with a program code listed in the Appendix. The Poisson distribution is more efficient for spreading at low time values with short path lengths and the situation is reversed for higher time values and longer path lengths. The exact time value where the crossing occurs depends on the event rate parameter value of the Poisson distribution.

The main contribution of this study is proposing a common methodology for analysing network structure and dynamics on complex networks. A quality function is defined based on elements of the probability matrix or the influence spreading matrix. Elements of the matrix describe interactions in the network structure. The quality function is used for detecting communities and studying community structures. Properties and characteristics of a network can be analysed with several network metrics. The generalised centrality measure proposed in this study is a form of node influence metrics that rank or quantify the influence of nodes.

Appendix

See Table 6.

Table 6 Fifty highly ranked divisions out of the 201 detected divisions in 100 computer runs of the Les Misérables network in the influence spreading model with uniformly distributed $U(0.05, 0.1)$ random link weights

id	#	*Nodes*	#	*Nodes*
1	7	17–23	70	1–16,24–77
2	15	1–2,5–10,17–23	62	3–4,11–16,24–77
3	8	1–2,5–10	69	3–4,11–77
4	25	1–2,5–10,47–49,58–68,74–75,77	52	3–4,11–46,50–57,69–73,76
5	12	58–68,77	65	1–57,69–76
6	26	17–23,47–49,56–68,74–75,77	51	1–16,24–46,50–55,69–73,76
7	35	1–10,30,35–39,47–49,56–68,74–75,77	42	11–29,31–34,40–46,50–55,69–73,76
8	17	47–49,58–68,74–75,77	60	1–46,50–57,69–73,76
9	32	17–23,40,47–50,52–68,74–75,77	45	1–16,24–39,41–46,51,69–73,76
10	29	13,17–29,31–32,34,41–44,46,51,69–73,76	48	1–12,14–16,30,33,35–40,45,47–50,52–68,74–75,77
11	31	30,35–40,47–50,52–68,74–75,77	46	1–29,31–34,41–46,51,69–73,76
12	27	1–2,5–10,17–23,58–68,77	50	3–4,11–16,24–57,69–76
13	35	11–16,24–39,41–46,51,69–73,76	42	1–10,17–23,40,47–50,52–68,74–75,77
14	19	17–23,58–68,77	58	1–16,24–57,69–76
15	20	1–2,5–10,58–68,77	57	3–4,11–57,69–76
16	37	3–4,11–16,24–39,41–46,51,69–73,76	40	1–2,5–10,17–23,40,47–50,52–68,74–75,77
17	36	3–4,11–12,14–16,29–30,33,35–39,45–49,56–68,74–75,77	41	1–2,5–10,13,17–28,31–32,34,40–44,50–55,69–73,76
18	22	35–39,47–49,58–68,74–75,77	55	1–34,40–46,50–57,69–73,76
19	37	1–10,13,17–28,31–32,34,41–44,51,69–73,76	40	11–12,14–16,29–30,33,35–40,45–50,52–68,74–75,77
20	30	1–2,5–10,35–39,47–49,58–68,74–75,77	47	3–4,11–34,40–46,50–57,69–73,76
21	36	11–29,31–34,41–46,51,69–73,76	41	1–10,30,35–40,47–50,52–68,74–75,77
22	25	17–23,47–49,56,58–68,74–75,77	52	1–16,24–46,50–55,57,69–73,76
23	33	1–2,5–10,17–23,47–49,56,58–68,74–75,77	44	3–4,11–16,24–46,50–55,57,69–73,76
24	35	3–4,11–16,25–29,32–34,40–46,50–55,69–73,76	42	1–2,5–10,17–24,30–31,35–39,47–49,56–68,74–75,77
25	32	13,17–28,31–32,40–44,50–55,69–73,76	45	1–12,14–16,29–30,33–39,45–49,56–68,74–75,77
26	25	17–26,30–31,35–39,41–43,69–72,76	52	1–16,27–29,32–34,40,44–68,73–75,77
27	33	25–26,28,40–43,47–49,51,53,56–72,74–77	44	1–24,27,29–39,44–46,50,52,54–55,73
28	33	1–10,35–39,47–49,56,58–68,74–75,77	44	11–34,40–46,50–55,57,69–73,76
29	37	1–16,25–29,32–34,41–46,51,69–73,76	40	17–24,30–31,35–40,47–50,52–68,74–75,77
30	30	35–40,47–50,52–68,74–75,77	47	1–34,41–46,51,69–73,76
31	33	11–16,25–29,32–34,40–46,50–55,69–73,76	44	1–10,17–24,30–31,35–39,47–49,56–68,74–75,77
32	33	17–24,30,35–39,47–49,56–68,74–75,77	44	1–16,25–29,31–34,40–46,50–55,69–73,76
33	38	11–12,14–16,30,33,35–40,45,47–50,52–68,74–75,77	39	1–10,13,17–29,31–32,34,41–44,46,51,69–73,76
34	35	1–10,42,47–49,56–72,74–77	42	11–41,43–46,50–55,73
35	37	3–4,11–12,14–16,29–30,33–39,45–49,56–68,74–75,77	40	1–2,5–10,13,17–28,31–32,40–44,50–55,69–73,76
36	33	1–2,5–10,42,47–49,56–72,74–77	44	3–4,11–41,43–46,50–55,73
37	27	13,17–26,30–32,35–39,41–43,69–72,76	50	1–12,14–16,27–29,33–34,40,44–68,73–75,77
38	33	1–2,5–10,17–26,30–31,35–39,41–43,69–72,76	44	3–4,11–16,27–29,32–34,40,44–68,73–75,77
39	36	1–10,17–23,47–49,56–68,74–75,77	41	11–16,24–46,50–55,69–73,76
40	14	1–2,5–10,18–23	63	3–4,11–17,24–77
41	6	18–23	71	1–17,24–77
42	25	42,47–49,56–72,74–77	52	1–41,43–46,50–55,73
43	37	11–34,41–46,51,69–73,76	40	1–10,35–40,47–50,52–68,74–75,77

Table 6 (continued)

id	#	*Nodes*	#	*Nodes*
44	36	3–4,11–24,27,29–39,44–46,50,52,54–55,73	41	1–2,5–10,25–26,28,40–43,47–49,51,53,56–72,74–77
45	38	17–23,30,35–40,47–50,52–68,74–75,77	39	1–16,24–29,31–34,41–46,51,69–73,76
46	36	11–16,24–29,31–34,40–46,50–55,57,69–73,76	41	1–10,17–23,30,35–39,47–49,56,58–68,74–75,77
47	28	13,17–26,28,30–32,35–39,41–43,69–72,76	49	1–12,14–16,27,29,33–34,40,44–68,73–75,77
48	28	11–16,25–29,31–34,41–46,51,69–73,76	49	1–10,17–24,30,35–40,47–50,52–68,74–75,77
49	28	3–4,11–12,14–16,25–29,32–34,41–46,51,69–73,76	49	1–2,5–10,13,17–24,30–31,35–40,47–50,52–68,74–75,77
50	30	25–26,40–43,47–49,56–72,74–77	47	1–24,27–39,44–46,50–55,73

The following program code describes a process of receiving an e-mail message once in a time unit and forwarding it once in a time unit. In the program code, S is the survival distribution function used to demonstrate the e-mail forwarding model in section 'Temporal spreading on networks'.

```
! Survival function S is a function of time t and path length l
S=0
S(1,1)=1.0
do t=1,iT      ! time t
  if (t>1) then
    do l=1,iL ! path length l
      S(t,l)=S(t-1,l)
    end do
  end if
  do jl=1,iL
    l=iL+1-jl
    do i2=1,l-1
      S(t,l)=S(t,l)+S(t,l-i2)/factorial(i2)-S(t,l)*S(t,l-i2)/factorial(i2)
    end do
  end do
  if (t==1) then
    do l=1,iL-1
      S(t,l)=S(t,l+1)
    end do
  end if
  S(1,iL)=1/factorial(iL)
end do
```

Acknowledgements
Not applicable.

Authors' contributions
The author read and approved the final manuscript.

Funding
No funding.

Availability of data and materials
Not applicable.

Declarations

Competing interests
The author declare that he has no competing interests.

Received: 2 March 2020 Accepted: 3 March 2021
Published online: 18 March 2021

References

1. Barabási A-L: Network Science. Cambridge University Press (2016).
2. Newman MEJ. Networks. Oxford: An Introduction. Oxford University Press; 2010.
3. Luciano da Fontoura Costa, What is a complex network? (CDT-2), (2020). doi: https://doi.org/10.13140/RG.2.2.10450.04804/1.
4. Gómez S: Centrality in networks: finding the most important nodes. In: Moscato, P, de Vries, NJ (eds.) Business and Consumer Analytics: New ideas. Part III, Chapter 8, pp. 401–434 (2019).
5. Katz L. A new status index derived from sociometric analysis. Psychometrika. 1953;18(2):39–43.
6. Coscia M, Giannotti F, Pedreschi D. A classification for community discovery methods in complex networks. Stat Anal Data Min. 2011;4(4):512–46.
7. Fortunato S, Hric D. Community detection in networks: a user guide. Phys Rep. 2016;659(11):1–44.
8. Lancichinetti A, Fortunato S. Community detection algorithms: a comparative analysis. Phys Rev E. 2009;80:056117.
9. Yang Z, Algesheimer R, Tessone CJ. A comparative analysis of community detection algorithms on artificial networks. Sci Rep. 2016;6:30750. https://doi.org/10.1038/srep30750.
10. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E: Fast unfolding of communities in large networks. Journal of Statistical Mechanics. P10008 (2008)
11. Rosvall J, Bergstrom CT: Maps of random walks on complex networks reveal community structure. PNAS105: 1118 (2008).
12. Karrer B, Newman MEJ. Stochastic blockmodels and community structure in networks. Phys Rev E. 2011;83(2):016107.
13. Lancichinetti, A, Fortunato, S, Kertész, J: Detecting the overlapping and hierarchical community structure in complex networks. New Journal of Physics. 11 (2009).
14. Riolo MA, Newman MEJ: Consistency of community structure in complex networks. Phys. Rev. E 101, 052306 (2020), https://arxiv.org/abs/1908.09867.
15. Barrat A, Barthélemy M, Vespignani A: Dynamical Processes on Complex Networks. Cambridge University Press (2008).
16. Smilkov D, Kocarev L. Influence of the network topology on epidemic spreading. Phys Rev E. 2012;85:016114.
17. Centola D: How behavior spreads, the science of complex contagions. Princeton University Press (2018).
18. Kuikka V: Subsystem Cooperation in complex networks – case brain network. In: Barbosa H., Gomez-Gardenes J., Gonçalves B., Mangioni G., Menezes R., Oliveira M. (eds) Complex Networks XI. Springer Proceedings in Complexity. Springer, Cham (2020).
19. Holme P, Saramäki J. Temporal networks. Phys Rep. 2012;519:97–125.
20. Wang W, Liu Q-H, Liang J, Hu Y, Zhou T. Coevolution spreading in complex networks. Phys Rep. 2019;820:1–51.
21. Kuikka V. Influence spreading model used to analyse social networks and detect sub-communities. Computational Social Networks. 2018;5:12. https://doi.org/10.1186/s40649-018-0060-z.
22. Ball MO, Colbourn CJ, Provan JS: Network reliability. In: Handbooks in Operations Research and Management Science. Chapter 11. vol 7, pp. 673–762 (1995).
23. Yang, J, Leskovec, J: Defining and evaluating network communities based on ground-truth, MDS '12: Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics, Article No. 3 (2012), https://doi.org/https://doi.org/10.1145/2350190.2350193.
24. Zachary WW. An information flow model for conflict and fission in small groups. J Anthropol Res. 1977;33:452–73.
25. Lusseau D, Schneider K, Boisseau OJ, Haase P, Slooten E, Dawson SM. The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations: Can geographic isolation explain this unique trait? Behav Ecol Sociobiol. 2003;54:396–405.
26. Kuikka V: Influence spreading model used to community detection in social networks. In: Cherifi C, Cherifi H, Karsai M, Musolesi M (eds.) Complex Networks & their applications VI. COMPLEX NETWORKS 2017. Studies in computational intelligence, vol. 689. Cham: Springer, pp. 202–215 (2018).
27. Kuikka V: A General method for detecting community structures in complex networks. In: Cherifi H. et. al. (eds.) Complex Networks & Their Applications VIII. COMPLEX NETWORKS 2019. Studies in Computational Intelligence, vol. 881. Springer (2019). https://doi.org/https://doi.org/10.1007/978-3-030-36687-2_19.
28. Zhang Z-K, Liu C, Zhan X-X, Lu X, Zhang C-X. Dynamics of information diffusion and its applications on complex networks. Phys Rep. 2017;651:1–34.
29. Horváth DX, Kertész J. Spreading dynamics on networks: the role of burstiness, topology and non-stationarity. New J Phys. 2014;16:073037.
30. Leskovec J, Krevl A: SNAP Datasets, Stanford Large Network Dataset Collection (2014)

Publisher's Note