Applied Network Science

RESEARCH Open Access

Ego-zones: non-symmetric dependencies reveal network groups with large and dense overlaps

Milos Kudelka, Eliska Ochodkova*, Sarka Zehnalova and Jakub Plesnik

*Correspondence:
eliska.ochodkova@vsb.cz
Dept. of Computer Science, Faculty of Electrical Engineering and Computer Science, VSB-Technical University of Ostrava, 17. listopadu 2172/15, 708 00 Ostrava-Poruba, Czech Republic

Abstract

The existence of groups of nodes with common characteristics and the relationships between these groups are important factors influencing the structures of social, technological, biological, and other networks. Uncovering such groups and the relationships between them is, therefore, necessary for understanding these structures. Groups can either be found by detection algorithms based solely on structural analysis or identified on the basis of more in-depth knowledge of the processes taking place in networks. In the first case, these are mainly algorithms detecting non-overlapping communities or communities with small overlaps. The latter case is about identifying ground-truth communities, also on the basis of characteristics other than only network structure. Recent research into ground-truth communities shows that in real-world networks, there are nested communities or communities with large and dense overlaps which we are not yet able to detect satisfactorily only on the basis of structural network properties.

In our approach, we present a new perspective on the problem of group detection using only the structural properties of networks. Its main contribution is pointing out the existence of large and dense overlaps of detected groups. We use the non-symmetric structural similarity between pairs of nodes, which we refer to as dependency, to detect groups that we call zones. Unlike other approaches, we are able, thanks to non-symmetry, accurately to describe the prominent nodes in the zones which are responsible for large zone overlaps and the reasons why overlaps occur. The individual zones that are detected provide new information associated in particular with the non-symmetric relationships within the group and the roles that individual nodes play in the zone. From the perspective of global network structure, because of the non-symmetric node-to-node relationships, we explore new properties of real-world networks that describe the differences between various types of networks.

Keywords: Network, Community, Overlap, Dependency, Prominency, Ego-zone

Introduction

Frequently solved problems in complex network analysis include the study of network structures. One of the challenges in this area is to design methods capable of detecting groups of nodes that have empirically determined properties that are common in real-world networks.

The procedure associated with this task is community detection, and it is a well-known fact that some real-world networks, e.g., social networks, have a community structure.

However, the concept of a network community is not precisely defined. Informally, a network community is often described as a group of nodes that are strongly connected inside the community but weakly connected with other communities. Unfortunately, this definition cannot be applied in real-world situations, where one node may belong to multiple communities. In this case, communities either partially overlap or one community is entirely nested into another community.

There are many different methods used to detect communities in networks. These methods are based on various approaches, the first comprehensive overview of which can be found in Fortunato (2010). This survey is also focused on the methods used for detecting overlapping communities. In this survey, overlaps are understood mostly either as a group of nodes connecting several communities (hubs) or as a connection within a hierarchy described in a way similar to hierarchical clustering by a dendrogram. A detailed overview of overlapping community detection methods can be found in Xie et al. (2013). As the authors point out, a common feature of the methods being investigated is the small fraction of the nodes in the overlaps.

Recent results have shown three essential properties describing network community structure: (a) there are large overlaps that have a higher density compared to the density of overlapping communities (Yang and Leskovec 2012); (b) the similarity of links has a significant influence on the size of communities and their overlaps (Ahn et al. 2010), and (c) on the basis of a close relationship between the high density of triangles and the existence of a community structure, triadic closure as a natural mechanism leads to the emergence of a community structure (Bianconi et al. 2014).

In our approach, we combine ego-network analysis and seed-based community detection methods (Bagrow and Bollt 2005; Clauset 2005) in that we choose a node as a seed for the detection of a group. It differs from them in that, as in ego-network analysis, each seed (ego) is the basis of the group which we call an ego-zone (a zone in short). Ego-zone detection is, similarly to Ahn et al. (2010), based on the analysis of similarities in a networks. However, we analyze the similarity of adjacent nodes, and moreover, we understand the similarity as non-symmetric, which corresponds better to the reality (Tversky 1977). Therefore, the approach presented in this article combines the properties mentioned above with one additional principle – non-symmetric similarity.

To measure similarity, we use dependency (Kudelka et al. 2015). The calculation of the dependency is based on the ratio of the weights of triangles shared by the adjacent nodes and the weights of all the edges of each adjacent node. Using the non-symmetric relation of dependency, in this article we present several key findings based on observations of real-world networks. First, we show that there are three types of nodes in the networks in terms of dependency. They are (1) nodes that are not dependent on any other node, (2) nodes on which no other node depends, and (3) nodes that have around them both dependent nodes and nodes they depend on. In particular, the first type of nodes (independent) describes “key players”, especially in networks with social interaction. The nodes of both the first and the third types significantly affect overlapping groups of nodes, which we call ego-zones. For zones, we define the roles that nodes of each type play in them. We will show that our definition of a zone as a group with two types of internal dependencies and specific roles of nodes, not only in the neighborhood but also in the wider surroundings of the chosen node (ego), leads to overlapping zones. We will explain why overlaps are created and also that they can be large and other zones may be nested inside them. In experiments with both generated and real-world networks, we will show what properties they have in terms of dependency and zones, and how the real-world networks differ from the ones that are generated and among themselves. In experiments with real-world networks, we will explore the relationship between zones and communities in the traditional sense and ground-truth communities. An interesting conclusion is that in some types of networks, it is possible to find zones that correspond to the traditional view of communities, while in others, they correspond to the ground-truth communities. In our experiments, we investigate in particular those properties that are related to dependency and detected zones. However, for some comparisons, we also utilize known structural properties of networks, especially average degree, modularity, and average clustering coefficient.

Related work

Groups of nodes that are likely to share common features and/or play similar roles in the network are called clusters or, more often, communities. It is a well-known fact that some real-world networks, e.g., social networks, have a community structure. Community detection is not a well-defined problem because there is no universal definition of community and the nature of communities is not known in advance. The problem is also complicated by the variability of community forms: disjoint, overlapping or, for example, hierarchical communities may appear. As a result, there is no manual on how to use the algorithm, how to evaluate the performance of different algorithms or how to compare them. The authors (Fortunato and Hric 2016) offer a guided tour of the main aspects of this issue, discuss the strengths and weaknesses of popular methods, and provide guidance on how to use them.

One of the first publications about communities that is most often mentioned is Girvan and Newman (2002). The authors proposed a community detection algorithm based on edge betweenness, which is a generalization of Freeman's node betweenness centrality (Freeman 1977). This method is an example of detection methods based on a division of a network (or underlying graph). Simultaneously, it is an example of a global, or a top-down, method. The method is not capable of finding overlapping communities, because each node is assigned to only one community. Other representatives of global methods, the results of which are non-overlapping communities, include, among others, one of the oldest algorithms, the Kernighan-Lin algorithm (Kernighan and Lin 1970), the spectral bisection method (Barnes 1982) and hierarchical clustering. The last example uses the symmetrical similarity rate because it assumes that communities are made of mutually similar nodes and this similarity is symmetrical. An example of a different approach to hierarchical clustering is, e.g., the Walkatrap algorithm, which is based on a random walk (Pons and Latapy 2005). The novel CAN algorithm (Zhang et al. 2018) is proposed to reveal community structure using the correlation analysis of nodes. A wide scale of methods is further represented by methods based on modularity (Newman and Girvan 2004) and its optimization (Blondel et al. 2008; Guimera et al. 2004). A large number of metrics have been proposed, a detailed survey of the metrics proposed for community detection and evaluation can be found in Chakraborty et al. (2017).

Local (or seed-based) methods begin searching from a random node and then gradually add neighboring nodes one by one on the basis of the optimization of measured metrics or heuristics. This process is named local expansion. From among the many methods, the following can be named: the well-known method of Bagrow and Bollt (2005) or the agglomerative algorithm of Clauset (2005), which uses greedy maximization of local modularity to find local communities. The starting nodes need not only be chosen at random. For instance, in Khorasgani et al. (2010), the community is created as a group of followers assembled around a potential leader.

It is a natural property of many real-world networks, especially social networks, that a node may be a member of multiple communities and not only of one community, which leads to the emergence of overlapping communities. The Clique Percolation Method (Palla et al. 2005), in which the community that is obtained, named the k-clique community, is the union of all k-cliques that can be reached from each other through a series of adjacent k-cliques, is a very popular method. This method, however, assumes the existence of cliques, which looks, even for social networks, like an unreal assumption. The idea of partitioning edges instead of nodes was also explored. The node in the original graph is called overlapping if the edges associated with it belong to more than one community (Ahn et al. 2010; Evans and Lambiotte 2010). Local expansion is also used to detect overlapping communities (Lancichinetti et al. 2009; Baumes et al. 2005). Another, dynamic, approach is the algorithm to detect overlapping communities in networks by label propagation called COPRA (Gregory 2010).

There is a question whether the structural view on communities corresponds to real-world communities, about the existence of which information is available from non-topological properties of networks (or from the attributes of nodes). A negative answer can be found in Hric et al. (2014). The authors (Yang and Leskovec 2015) introduced the concept of ground-truth communities and proposed a methodology, which compares and evaluates how do various structural definitions of network communities correspond to ground-truth communities. They allow ground-truth communities to be nested and to overlap. The existence of these nested communities and their detection was also published by, e.g., Tatti and Gionis (2013).

The community view on groups of nodes is one of the possible ones. A different approach to the analysis of groups of nodes is the egocentric approach. It is focused on the node referred to as the "ego" and its neighbors, known as "alters". This approach naturally applies mainly to the analysis of social networks. For example, in Abbasi et al. (2012) the authors dealt with the analysis of co-authorship networks and the question of whether the collaboration skills and research performance of researchers were correlated. McAuley and Leskovec (2014) designed an algorithm to automatically detect circles in ego-networks, so that alters may belong to any number of circles, including none. They found circles that were disjoint, overlapping and hierarchically nested.

Our approach to the detection of groups of nodes (ego-zones) is related to Danisch et al. (2013). The authors suspect that a well-chosen set of few nodes could define a single community. The key idea is that, although one node generally belongs to numerous communities, a small set of appropriate nodes can fully characterize a single community. They work with similarity measure called Carryover opinion metric.

The term "dependency" can be found in Parshani et al. (2011); Bashan et al. (2011). The authors work with what are termed "dependency links" and "dependency networks" and analyze the cascade dissemination of errors in a system and state that if a node has a lot of neighbors that are dependent on it, then its vulnerability will affect the vulnerability of all of the dependent nodes. This fits in with our concept of ego-zones (see

"Ego-zones" section), where we can watch ego-zones through the lens of "dependency links", so that the removal of the ego from a network means, for example, the removal of the entire zone (if it is small and has no sub-zones). Alternatively, it can mean only the break-up of a large zone into sub-zones, in which the removed ego does not play an important role (most of the nodes in such a sub-zone are not dependent on this ego).

A similar term, "influence", is used by Jacob et al. (2016), who propose a graph theory approach that focuses on the correlation influence between selected brain regions, named Dependency Network Analysis. Partial correlations are used to quantify the level of influence of each node during the performance of this task.

Dependency

If we consider a group of nodes fulfilling a particular purpose or function in a network, then we can expect that the nodes in a group will be similar in terms of this purpose. On the other hand, we can assume that the similarity between two objects in a group does not generally have to be symmetrical. This is based on the assumption that, in assessing the similarity of two objects, it is necessary to take into account not only their common properties but also the properties in which both objects differ (Tversky 1977).

Let us now project this assumption into the structure of a network in order to use this structure to measure the similarity between a pair of adjacent nodes, x and y. Consider all the nodes adjacent to node x or node y. These nodes can be divided into three groups. The first group is the shared neighbors of nodes x and y. These neighbors represent triangles shared by both nodes and can be understood as the basis of the similarity. Therefore, a higher number of triangles increases the similarity of nodes x and y. The remaining two groups of nodes include those nodes that are adjacent either to node x or node y. Here, a higher number of non-shared neighbors of nodes x or y reduces their similarity.

When formalizing these considerations, let us further assume that we are working with a weighted undirected network. The non-symmetrical similarity of node x to node y will be called a structural dependency, from now on referred to as dependency (Kudelka et al. 2015).

Definition 1 *Structural dependency. Let x, y be nodes, then dependency $D(x,y)$ of node x on node y is defined as follows:*

$$D(x,y) = \frac{w(x,y) + \sum_{v_i \in CN(x,y)} w(x,v_i) \cdot r(x,v_i,y)}{\sum_{v_j \in N(x)} w(x,v_j)} \tag{1}$$

$$r(x,v_i,y) = \frac{w(v_i,y)}{w(x,v_i) + w(v_i,y)}, \tag{2}$$

where $CN(x,y)$ is set of all common neighbors of x, y, $N(x)$ is set of all neighbors of x, $w(x,y)$ is weight of edge between x, y, and $r(x,v_i,y)$ is the coefficient of the dependency of node x on node y via the common neighbor v_i.

Equation 1 shows that the numerator contains the dependency of node x on y with the edge weight between nodes x and y counted in, as well as reduced edge weights between node x and particularly shared neighbors. The reduction is a value dependent on the weight of the edges between nodes x or y and their shared neighbors. The reduction value increases or decreases with an increase or decrease in the weight of the edge between a shared neighbor and node y. The denominator contains the sum of the weights of the edges between node x and all its neighbors. When we consider a reverse dependency of node y on node x, then the denominator will be the sum of the weights of the edges between node y and all its neighbors, and the numerator will also differ because of different weights and reduction values. Therefore, the dependency of node x on node y can differ from the dependency of node y on node x.

If we work with an unweighted network, then the weights of all the edges will be equal to 1, and all the reduced values will be equal to 0.5. The value of an expression in the numerator will be the same for both dependencies, but the values of denominators can vary. Therefore, even for an unweighted network, the dependency of the nodes is non-symmetric. Thus, our method is designed with weighted networks in mind, but can also be applied to unweighted ones. Moreover, the formulas from Definition 1 can also be used for directed networks; however, this case lies beyond the scope of this article. Therefore, below we will work only with weighted or unweighted undirected networks.

Figure 1 shows an undirected unweighted network with nine nodes to illustrate different dependencies of neighboring nodes and two zones with their overlap (which will be explained in detail in "Ego-zones" section and the experimental "Zones in generated networks" and "Zones in real-world networks" sections).

IsDependent relationship

To simplify the view on the dependency between two adjacent nodes x and y, let us define the relationship IsDependent as follows:

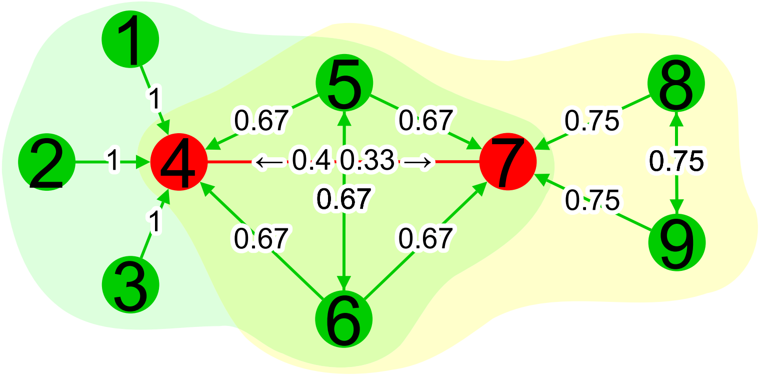

Fig. 1 Dependency and zones in unweighted network. The network presented here is undirected and unweighted. However, the arrows in this network express the direction of the dependency in a case when $D(x, y) \geq 0.5$ (informally, a strong dependency). The values of these dependencies are displayed on the edges. It can be seen that all the green nodes are strongly dependent either on node 4 or node 7 or both of them (for nodes 5 and 6). Moreover, the pairs of nodes 5, 6, and 8, 9 are mutually strongly dependent. Nodes 4 and 7 are specific (independent); they are not strongly dependent on any other node and have only weaker mutual dependencies (0.33 and 0.4). On the basis of the explanation in the "Ego-zones" section, the two largest detected zones in the network are displayed, the green zone with ego 4 and the yellow with ego 7. Their overlap with four nodes is a multi-ego zone with egos 5 and 6. Nodes 4 and 7, thanks to their independency on other nodes, induce this overlap. Moreover, nodes 5 and 6 belong to the overlap because of their strong dependency on both of these independent nodes. It can be noticed that the overlap is a clique and is denser than both overlapping zones. A total of seven zones were detected in this network. In addition to the three already mentioned, these are zones [1 4], [2 4], [3 4] with egos 1, 2, and 3, and one multi-ego zone [8 9 7] with egos 8 and 9

Definition 2 *IsDependent. Let x, y be neighboring nodes, then IsDependent relationship is defined as follows:*

IsDependent(x, y) = True if $D(x, y) \geq 0.5$; otherwise IsDependent(x, y) = False. The dependency threshold is set to 0.5 to take into account and reasonably balance a mutual dependency between two neighboring network nodes.

This relationship can be used to transform the original network into an unweighted directed network. In Fig. 2a is a well-known Karate Club network after the transformation. Edges exist only between nodes where at least one is dependent on the other, and their direction corresponds to the relationship IsDependent. The node size corresponds to the in-degree centrality of the given node. The transformed network in Fig. 2a emphasizes information about the structure of the original network, which is in Fig. 2b.

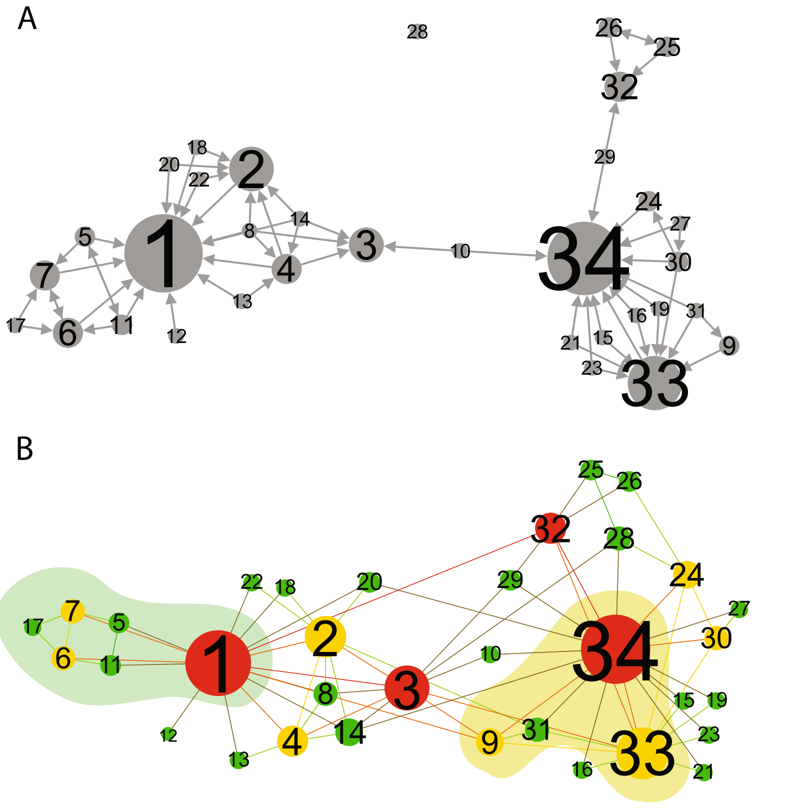

Fig. 2 Karate Club. Transformed unweighted directed network (**a**), and network with strongly-prominent nodes marked in red and weakly-prominent nodes in yellow (**b**). For the isolated node 28 in sub-figure **a** it is true that it is not dependent on any other node and at the same time, no other node in the network is dependent on it. Node 10 is dependent on nodes 3 and 34. Most nodes are dependent on nodes 1 or 34. There are nodes 32, 6, 7, 14, 28 with different prominency (see Definition 5), but a comparable degree. Node 32 is strongly-prominent (*OwIndep* = 3, *TwDep* = 0, *OwDep* = 0, *Prominency* = 1, *Degree* = 6). Nodes 6 and 7 are weakly-prominent (*OwIndep* = 2, *TwDep* = 1, *OwDep* = 1, *Prominency* = 0.667, *Degree* = 4) because of their dependency on each other and on node 1. Nodes 14 or 28 are non-prominent (*OwIndep* = 0, *TwDep* = 0, *OwDep* = 4 or 0, resp., *Prominency* = 0, *Degree* = 5 or 4, resp.) because no other nodes are dependent on them. As explained in Fig. 7 in "Ego-zones" section, sub-figure **b** displays one multi-ego zone (green) with egos 6 and 7, and one group of nodes (yellow) representing two alternative zones with egos 9 or 31, respectively

After the network has been transformed into its unweighted directed version, all the neighbors of each node of the network can be, by using Definition 3, divided into four groups described by different types of dependencies (for examples, see Fig. 2).

Definition 3 *Four types of dependencies. Let x be a node, then:*

$OwDep_x$ is the number of neighbors on which x is dependent, but which are not dependent on x (one-way dependency);

$OwIndep_x$ is the number of neighbors which are dependent on x, but x is not dependent on them (one-way independency);

$TwDep_x$ is the number of neighbors which are dependent on x, and x is dependent on them (two-way dependency);

$TwIndep_x$ is the number of neighbors which are not dependent on x, and x is not dependent on them (two-way independency).

The nodes that have a non-zero value for OwIndep deserve special attention. These nodes can be divided into two groups (see Fig. 2b). The first group includes (red) nodes that are not dependent on other nodes. The second group contains (yellow) nodes that are dependent on at least one other node.

Definition 4 *Prominent nodes. Let x be a node, then:*

- *a node x is called prominent if $OwIndep_x > 0$;*
- *a prominent node x is called strongly-prominent if $OwDep_x = 0$ and $TwDep_x = 0$.*
- *a prominent node x which is not strongly-prominent is called weakly-prominent.*

Strongly-prominent or weakly-prominent nodes play roles of global or local authorities for those network nodes that are unilaterally dependent on them. Below, we will call the nodes in the roles of authorities "centers". In "Cause of overlaps" section, we show that the existence of prominent nodes is an important aspect causing overlaps between groups.

To determine whether and to what extent a node plays the center role, we define the value of node prominency (see Definition 5). When calculating this value, we measure the degree of independency of the node as the F1 score, based on a confusion matrix in which $true\ positives = OwIndep$, $false\ negatives = TwDep$, and $false\ positives = OwDep$. The point is to assess the network node x from the perspective of dependency of its neighbors on it and, also conversely, its independency on its neighbors; it means that *positives* are neighboring nodes dependent on the x node, and *negatives* are other neighbors.

Definition 5 *Prominency. Let x be a node, then its prominency is*

$$Prominency(x) = \frac{2 \cdot OwIndep_x}{2 \cdot OwIndep_x + TwDep_x + OwDep_x}. \tag{3}$$

Prominency is not defined for nodes having zero values for all types of dependencies in the formula. In this case, we set $Prominency = 0$.

In fact, using prominency, we can divide all network nodes into three prominency types. For strongly-prominent nodes, the $Prominency = 1$, and for weakly-prominent nodes, the $Prominency > 0$. The remaining network nodes are non-prominent and have

Prominency = 0. However, prominency should not be seen as a new centrality. For example, there may be nodes that have a comparable degree, but with different types of prominency. Nodes 6, 7 (weakly-prominent), 14, 28 (non-prominent), and 32 (strongly-prominent) in Fig. 2 are examples. Basically, prominency expresses the importance of a node for its neighbors, regardless of the number of these neighbors.

While strongly-prominent nodes are entirely independent, weakly-prominent nodes share their prominency with weakly-prominent or strongly-prominent nodes in their surroundings. In "Zones in real-world networks" section, we analyze 16 real-world networks. One of the key findings is the different proportion between the number of nodes of the three types of prominency for different types of networks (see Fig. 3).

In Fig. 2b, strongly-prominent or weakly-prominent nodes are marked in red or yellow. In Figs. 18 and 19 in Appendix C, the Les Misérables network is presented as well as the largest connected component of the Net Science network.

The node properties of all three small networks are summarized in Table 1, and the properties associated with the IsDependent relationship are shown in Table 2 (the NetDep property will be explained in "Zones in generated networks" section).

Ego-zones

Using the dependency, we are able to describe a node group within a network with specific characteristics exactly and unambiguously. This description is based on one "central" node and dependencies in its surroundings.

Definition 6 *Ego-zone. The ego-zone is a group of network nodes meeting three following criteria:*

1. *the default member of the ego-zone is any network node called ego;*
2. *a member of the ego-zone is any node that is dependent on ego or another node of the ego-zone; the set of all such nodes including the ego is called the inner-zone;*

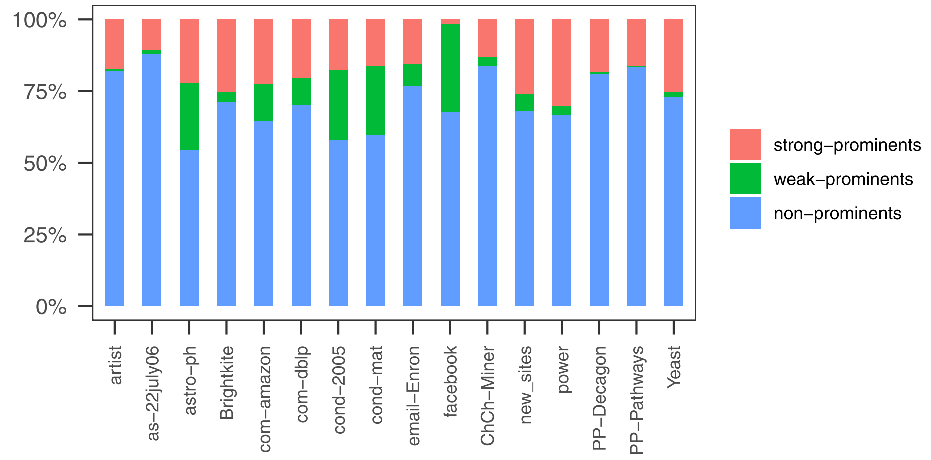

Fig. 3 Prominency types in real-world networks. The bar chart shows, in particular, the difference in the proportion of weakly-prominent nodes. This proportion is especially high in the collaboration networks (astro-ph, cond-mat, cond-mat-2005), and low in the biological (ChCh-Miner, PP-Decagon, PP-Pathways, Yeast) and technological (as-22july06, power) networks

Table 1 Properties of small networks

Network	n	m	Degree		CC	Strong-prominents		Weak-prominents		Modularity
			Max	Avg		Total	Percent	Total	Percent	
karate	34	78	17	4.58	0.485	4	11.7	8	23.5	0.416
lesmis	77	254	36	6.59	0.580	5	6.5	26	33.7	0.565
netscience	379	914	34	4.82	0.741	40	10.5	96	23.5	0.850

A high percentage of weakly-prominent nodes shows, despite the small network size, that these networks are close to real-world collaboration networks (see Fig. 3)

3. *a member of an ego-zone is each node outside the inner-zone on which at least one node of the inner zone is dependent; the set of all such nodes is called the outer-zone.*

Outer-zone nodes can be divided into two groups based on whether they are dependent on other nodes in the outer-zone.

Definition 7 *Outer-zone nodes. The outer-zone consist of two types of nodes:*

Liaison *is the outer-zone node which is not dependent on any other nodes of the outer-zone;*
Co-liaison *is the outer-zone node which is dependent on at least one another node of the outer-zone.*

For ego-zones, an alternative name – dependency zone – can be used in networks other than social ones; below, we will only use zone. The algorithm to detect zones based on an iterative procedure derived from Definition 6, including its scalability, is given in Appendix A.

For illustration, Fig. 4 shows a zone with the red ego and four regular nodes in the yellow inner-zone, one liaison and two co-liaisons in the blue outer-zone and, four green nodes outside the zone. The edge directions represent the dependency between nodes.

Thus, for each node of the network (ego), there exists its inner-zone, the size of which depends on the degree of direct or indirect dependency of the surrounding nodes on this ego. The natural characteristic is that there may be, especially in clique-close structures, more egos that have the same inner-zone. In this case, as is apparent from point 3 in Definition 6, their outer-zone must also be the same, and thus the zones as a whole. We consider those zones with the same inner-zone a single zone and refer to them as a *multi-ego zone*. Individual pairs of egos in these multi-ego zones must be dependent on one another (TwDep); otherwise, the individual egos would generate different inner-zones. On the other hand, there may be zones with corresponding nodes but with different inner-zones, and therefore their outer-zones also differ. Such zones are considered different and

Table 2 Node dependencies and network dependency in small networks

Network	OwDep		OwIndep		TwDep		TwIndep		NetDep
	Max	Avg	Max	Avg	Max	Avg	Max	Avg	
karate	4	1.647	13	1.647	1	0.176	6	1.118	0.756
lesmis	9	1.909	21	1.909	6	1.065	15	1.714	0.740
netscience	8	1.599	27	1.599	4	0.755	10	0.871	0.819

The average values of OwDep and OwIndep are the same because they are only opposite views of the IsDependent property. As can be seen, the Net Science network has stronger internal dependencies because its nodes have a lower average number of neighbors with which they are mutually independent

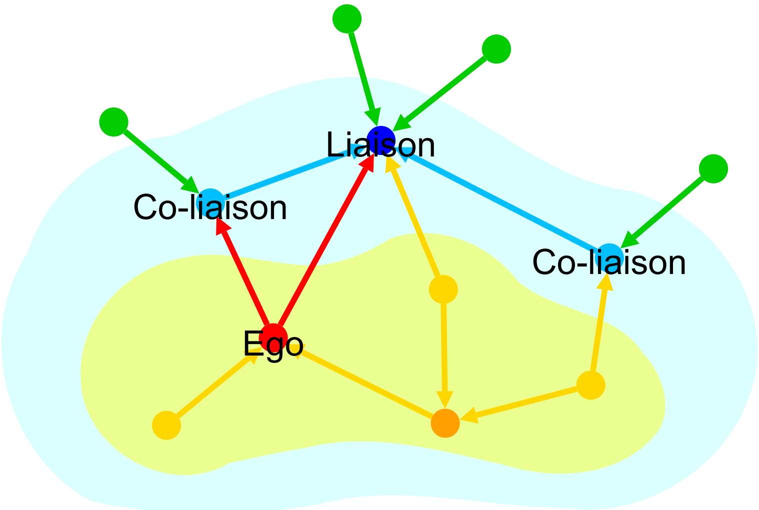

Fig. 4 Zone with eight nodes in four different roles. The zone has yellow inner-zone and blue outer-zone. Ego is dependent on Liaison and Co-liaison on the left, both Co-liaisons are dependent on Liaison, and other yellow nodes from inner-zone are dependent either on Ego or on Liaison or Co-liaison on the right. From a global perspective, the Liaison is a strongly-prominent node because it is not dependent on any other nodes and all its neighbors are one one-way dependent on it. The Ego, two Co-liaisons, and the dark yellow neighbor of Ego are weakly-prominent nodes; all of them have at least one neighbor dependent on them. Other nodes are non-prominent

are referred to as zones with alternative role configurations or an *alternative zones*. For examples of multi-ego and alternative zones, see Fig. 2.

As we will demonstrate in "Zones in real-world networks" section, depending on the network structure, zones of various sizes may exist, including zones with hundreds of nodes. Nevertheless, there may be trivial zones with a single node (e.g., node 28 at the Karate Club). In a more detailed view of the zone as a whole, there may be zones that have, for example, more nodes in the outer-zone than the inner-zone, or vice versa - zones that have no outer-zone.

From Definition 6 it follows that the zone in the network is unambiguously detected and that, in addition to the group of nodes belonging to the zone, we also receive further information:

1 the first one is about non-symmetric dependencies within a group. This information results in knowledge of the prominency of the individual nodes in the group;
2 the second one is the roles that individual nodes play in the zone. In particular, the questions are whether and to what extent there are prominent nodes in the group playing the role of egos in the inner-zone or liaisons and co-liaisons in the outer-zone. Both these pieces of information are crucial for a detailed assessment of intra and inter-group relationships;
3 the third one is the size of the zone and its density. Small densely connected zones are expected to be homogeneous, and their heterogeneity increases as the number of nodes (especially liaisons) in the zone increases and their density decreases.

We can suppose that the more interconnected the group is, the higher the consensus on the purpose or function of the group will be. Nodes that correspond entirely to this purpose have no edges that face outward. Conversely, if the nodes have edges that face outward, they represent this purpose only in part of their egos. Moreover, the more prominent nodes the zone contains, the higher the potential of the zone overlapping with other zones will be, as will be explained in "Cause of overlaps" section.

Although this is an intuition-based estimation, in an experiment in "Zones and ground-truth communities" section with four networks with identified ground-truth communities, we will show that the zones detected in some of these networks are a relatively good match for a non-trivial number of real communities.

We divide the nodes of the zone into four groups. In the first group, there are egos, and, the other nodes of the inner-zone are the second group. The third and fourth groups contain liaisons and co-liaisons, both belonging to the outer-zone. An example of the inner and outer-zones in the Les Misérables network is shown in Fig. 5. The properties of the nodes associated with zones are summarized in Table 3, the properties of zones are listed in Table 4, and the properties related to zone overlaps are listed in Table 5.

The values in Table 4 show that zones may overlap, so that one node can be a member of multiple zones. It is also clear that the maximum value of zone membership corresponds to the maximum value of membership in the liaison or co-liaison role. Here the key parameter is prominency; nodes with a non-zero value of prominency, i.e., strongly-prominent or weakly-prominent nodes, have a non-zero value of OwIndep; for this reason, they have neighbors that are dependent on them, but they are not dependent on these neighbors. In the zones to which these neighbors belong, there can be a strongly or weakly-prominent node in the liaison role, and weakly-prominent node in the co-liaison role. Thus, it is evident that the higher the value of OwIndep, the higher the potential

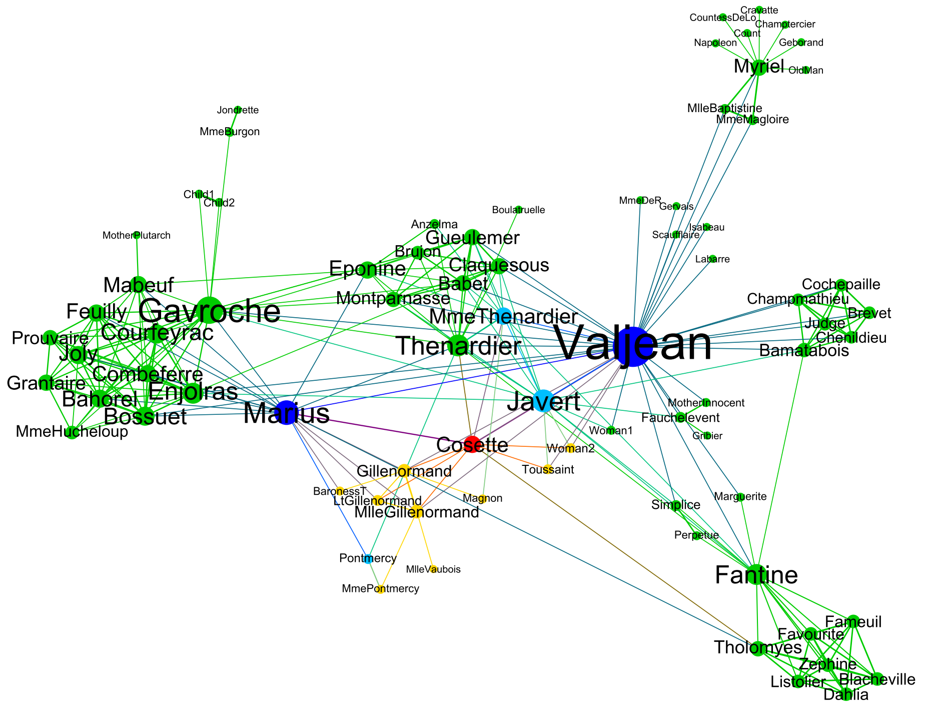

Fig. 5 Cosette's zone in Les Misérables network. The zone in which the red node Cosette is ego; the yellow nodes belong to inner-zone and the blue nodes to outer-zone; dark blue nodes play liaison role and light blue nodes co-liaison role. Green nodes are outside the zone

Table 3 Zone properties in small networks

Network	Zones	Zone size		Inner-zone size		Outer-zone size		Trivial	Dyad	Triad	Multi ego	Embed.
		Max	Avg	Max	Avg	Max	Avg					
karate	31	16	5.129	15	3.065	4	2.065	1	1	14	3	0.384
lesmis	58	37	6.483	32	3.776	12	2.707	0	17	10	8	0.366
netscience	276	49	5.859	40	3.547	9	2.312	0	26	66	71	0.500

In all networks, relatively large zones were detected considering the total number of network nodes and the average zone size. The Net Science network has a higher quality of zones measured by the embeddedness value; this means that nodes in zones are more strongly connected inside than outside the zone

of the prominent node for membership in different overlapping zones in the liaison role is. In all the three networks mentioned above, the node with the maximum membership value is in the liaison role in all the zones it belongs to (except for its own, where it is the ego).

Cause of overlaps

Nodes in the role of liaison or co-liaison are the cause of large overlaps. This follows from the fact that these nodes can be in both the outer and inner-zones of different zones. If we have such a node in two different zones, in the first of which it is in the outer-zone and in the second zone it is in the inner-zone, then the nodes from the first zone that are dependent on this node must be in the overlap of these zones as well.

For a better idea, let us have zone Z in which node v is a liaison (or co-liaison). Then node v is in its outer-zone. Thus, according to point 3 of Definition 6, one or more nodes u_i from the zone Z belonging to its inner-zone must be dependent on node v. Next, let us have zone Z_v in which v is any node of its inner-zone (e.g., ego). Then from point 2 of Definition 6 it follows that zone Z_v also contains nodes u_i which are dependent on v. As a result, nodes v and u_i must belong to the overlap of zones Z and Z_v.

The emergence of overlaps is illustrated in Fig. 1, where node 4 is the ego of the green zone and node 7 is its liaison; at the same time it applies that node 7 is the ego of the yellow zone. Thus, the overlap of these zones includes both nodes 4 and 7 and nodes 5 and 6, which are dependent on node 7. The same is true vice versa, where node 7 is the ego of the yellow zone, for which the liaison is node 4. Moreover, the overlap is a multi-ego zone in which the egos are nodes 5 and 6; nodes 4 and 7 are the liaisons of this zone.

To assess network community structure and quality of detected zones, we utilize two parameters; the first one is modularity, and the second one is embeddedness. With a higher value of modularity, the network community structure becomes more clear. Table 1 shows the weighted modularity Q for each network which comes from the interval $[-1, 1]$ and is defined as follows (Blondel et al. 2008):

Table 4 Memberships in small networks

Network	Membership		Liaisonship		Co-liaisonship	
	Max	Avg	Max	Avg	Max	Avg
karate	16	4.676	15	1.088	9	0.794
lesmis	21	4.883	20	0.883	7	1.156
netscience	26	4.266	25	0.995	16	0.689

The maximum value of zone membership corresponds to the maximum value of membership in the liaison role. The difference of one is because the node is an ego in its own zone

Table 5 Properties of zone overlaps in small networks

Network	Overlaps	Overlap size		Zones in overlaps		Zones in O size	
		Max	Avg	Total	Percent	Max	Avg
karate	145	7	1.490	4	2.6	7	5.000
lesmis	206	16	1.845	17	8.3	15	5.824
netscience	1275	13	1.562	93	7.3	11	4.763

Zones nested in overlaps exist even in small networks; the size of a nested zone may be close to or equal to the size of the overlap

$$Q = \frac{1}{2m} \sum_{ij} \left[w_{ij} - \frac{k_i \cdot k_j}{2m} \right] \delta(c_i, c_j), \tag{4}$$

where w_{ij} is the edge weight between nodes i and j, c_i is the community to which node i is assigned, $k_i = \sum_j A_{ij}$ is the sum of weights of the edges attached to node i, Kronecker delta δ_{c_i,c_j} is 1 if $c_i = c_j$ and 0 otherwise, and $m = \frac{1}{2} \sum_{ij} A_{ij}$.

To evaluate the quality of zones, we use the value of embeddedness from the interval $[0, 1]$; see Table 3. Group embeddedness is defined as the ratio between the internal degree of the group and its total degree (Hric et al. 2014). The higher the zone embeddedness value, the stronger the belonging of groups of nodes to the group as a whole. For a group of nodes, a sum of internal degrees k_{in} and a sum of total degrees of k_{tot}, the group embeddedness ξ is defined as follows:

$$\xi = \frac{k_{in}}{k_{tot}}. \tag{5}$$

The modularity in Table 1 shows that all three small networks have a community structure. However, the zones that are detected have a lower average value of zone embeddedness and therefore a lower quality than can be expected from communities. This fact is the first factor to indicate that zones, despite their exact and deterministic definition, cannot generally be considered communities. This is because the dependencies in both the inner and outer-zones implicitly do not provide a higher degree of interconnection within the zone than outside of it.

For each network, all pairs of overlapping zones were found. Table 5 lists the total number of overlaps, and their maximum and average sizes. The last four columns of the table show the total number of zones that are nested in some overlap of two other zones and their maximum and average sizes. The maximum overlap sizes are 7, 16, and 13, and the maximum zone sizes in the overlaps are 7, 15, and 11. In Fig. 6, for illustration, the Les Misérables network is shown with zone overlaps marked. There exist zones with other nested zones. We will use the terms *super-zone* and *sub-zone* to describe such situations. Simply put, in this example, the two listed zones are super-zones for all the zones in the overlap, and, vice versa, the zones in the overlap are sub-zones of both zones.

Figure 6 also shows two essential characteristics that correspond to the observation of real-world networks (Yang and Leskovec 2015). The overlaps of groups can be large and in the overlaps other groups can exist. To gain comprehensible visualization of these relations, we visualize the zone structures as weighted directed networks of sub-zones and zones that are nested in the overlaps of other zones. The visualized networks also include multi-ego zones and zones with the same group of nodes, but with alternative role configurations. Figure 7 shows the structures of zones for the Karate Club (A) and Les Misérables (B) networks. Figure 20 in Appendix C shows the same structure for the largest connected component of the Net Science network.

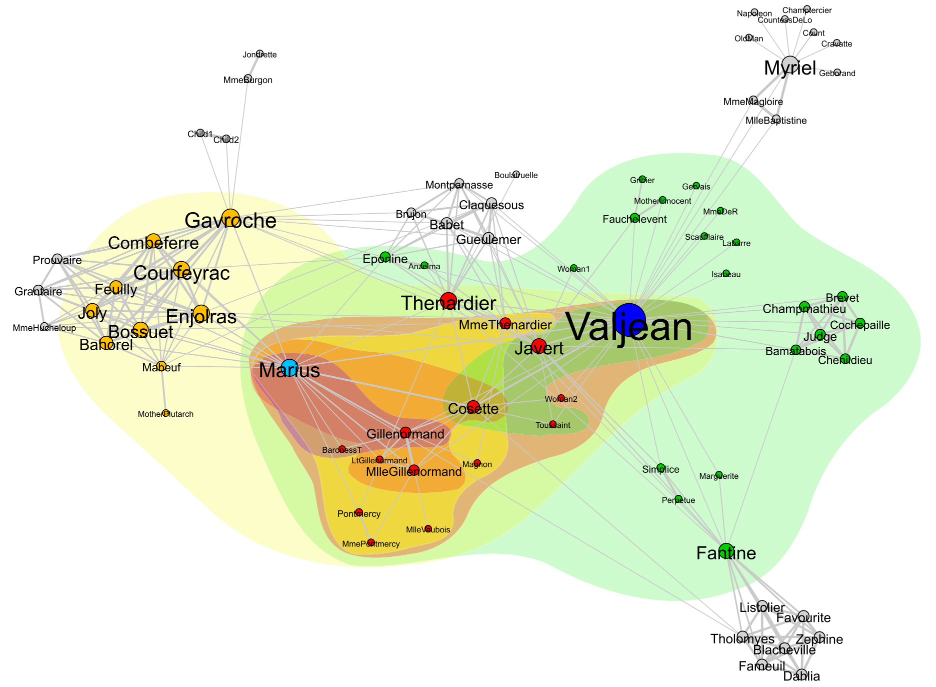

Fig. 6 Overlapping and nested zones in Les Misérables network. Two zones with Valjean (green) and Marius (yellow) egos overlap, and the other nodes of this overlap are red. The overlap nests Cosette's zone (peach color), for which the two selected egos play the liaison role. The Cosette's zone thus covers the entire overlap except for the Thenardier node. Further, other sub-zones nested in the overlap and in Cosette's super-zone are marked in dark yellow, dark green, orange and violet colors

In the experiments described in "Zones in generated networks" and "Zones in real–world networks" sections, we analyze the zone properties of generated and real-world networks. To analyze the overlaps, we detected only those overlapping pairs of zones in each network for which each pair of overlapping zones had at least ten nodes. For each overlap with at least four nodes, we found the largest zone that the overlap contained (if such exists).

Zones in generated networks

How the network structure affects the properties of zones is a natural question. Typical properties of real-world networks include, for example, scale-freeness related to the power-law distribution of node degree or community structure. Generative models can be used to explore various properties of networks. We use three models to assess how the properties of zones with network properties are related. The first is Erdös-Rényi (ER) model which generates random networks (Erdös and Rényi 1959), the second is Barabási-Albert (BA) model based on preferential attachment generating scale-free networks (Albert and Barabási 2002), and the third is Triadic Closure based (TC) model generating scale-free networks with a community structure (Bianconi et al. 2014). We used various settings for the experiments. The result was unweighted undirected networks with 10000 nodes. If an unconnected network was generated, we used the largest connected component for the analysis.

For networks generated using the ER model, we set the probability p of having an edge between a pair of nodes to the values of 0.0005 and 0.001. These are the values

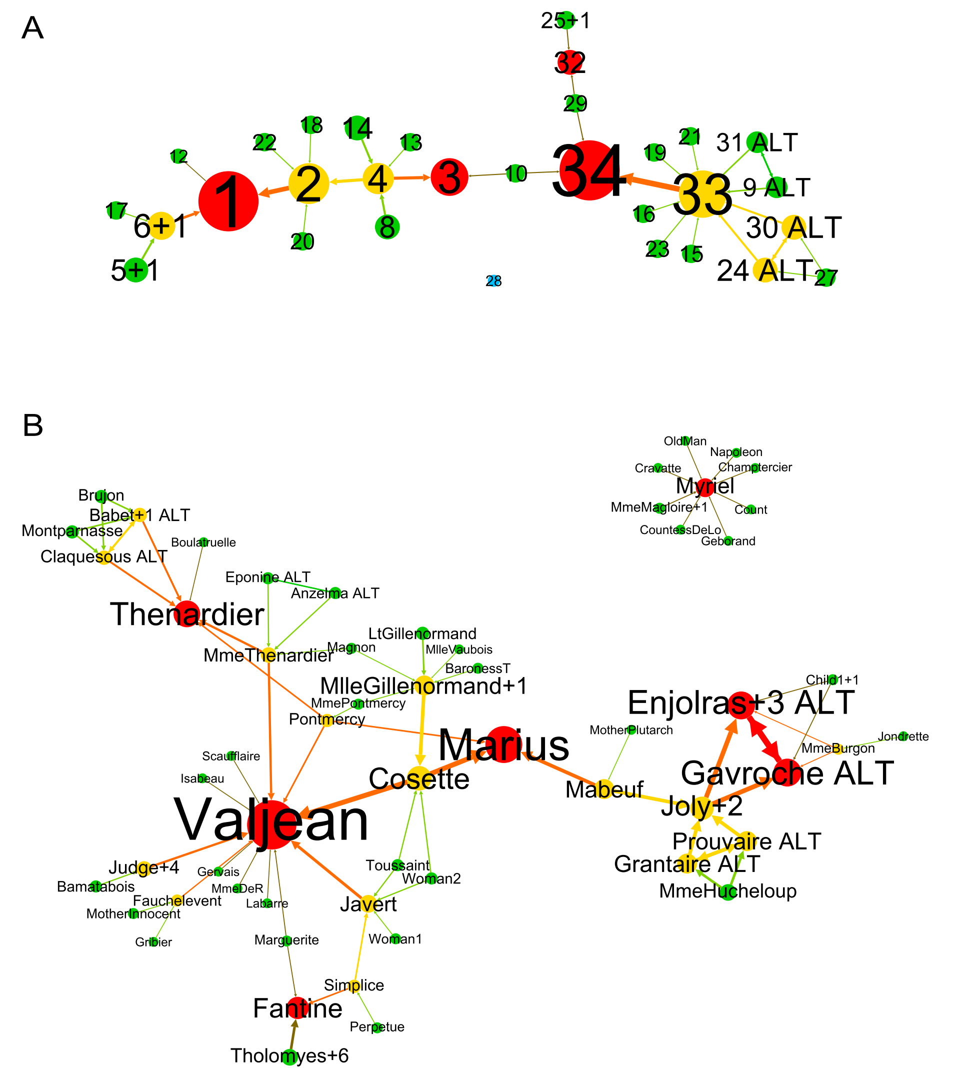

Fig. 7 Karate Club (**a**) and Les Misérables (**b**) zone networks. The networks were transformed into the weighted directed networks of sub-zones and zones nested in overlaps. Each pair of nodes in this transformed network connected by a directed edge represents zones; the smaller node represents the sub-zone that is nested into the super-zone represented by the larger node. The size of each zone corresponds to the number of nodes in this zone, and the edge strength corresponds to the number of nodes of the sub-zone. Red nodes represent zones that have no super-zone, green ones zones without sub-zones. The yellow zones have both sub-zones and super-zones, and blue zones have neither sub-zones nor super-zones. In sub-figure **a** with Karate Club, there is a multi-ego zone (node 6+1) with two egos 6 and 7, one liaison 1 and three other nodes 5, 11, 17 (see also Fig. 2). Moreover, there are, e.g., two alternative zones with egos 9 or 31 (9 ALT or 31 ALT). The first one has one liaison, 34, one co-liaison, 33, and one other node, 31. The second one has the same liaison, 34; however, it has two co-liaisons, 9, 33, and no other node. Similarly, in sub-figure **b**, Enjolras's zone has three other egos and one alternative Gavroche's zone. If the zone has at least two edges directed to different zones, then it is contained in their overlap. While in Karate Club, there are no zones nested in overlaps, the Les Misérables network has, for instance, such zones for Cosette, Pontmercy, and Mabeuf

that levitate around the threshold of when the network becomes connected (0.0009). For the BA model, we chose the values 3 and 4 of m representing a number of existing nodes to which a new node is connected. In its basic version, the TC model works with two parameters that are related to the connection of a new node to the network. This new node is connected to the randomly selected network node in the first step.

In the second step, the first parameter is the probability p, with which a neighbor of the selected node is preferred for connecting before a randomly selected node. The second parameter m defines the total number of connections for the new node, and, as a result, defines $m - 1$ repeated connections in the second step. A higher probability value of p increases the local connectedness among nodes and thus emphasizes the community structure of the network. In contrast, a higher value of m increases the network density. For the experiments, we chose 0.7 and 0.97 for p and $2, 3$, and 4 for m. The results of the analysis of the networks that were generated are summarized in Tables 6, 7, 8, 9, and 10.

The values of some parameters point to differences from the results of the analysis for the three small networks in "Ego-zones" section. These differences are most significant in the ER model, as can be expected (see Table 6).

Table 7 shows that the low maximum values of OwDep and TwDep are a common feature of all the networks that were generated, not only when compared with the three small networks from "Ego-zones" section but also compared to the real-world networks that will be analyzed in "Zones in real-world networks" section.

Table 7 also shows that ER and BA models generate networks that have low average values of OwDep, TwDep, and OwIndep. On the contrary, the average value of TwIndep is higher than in TC networks. These characteristics can be interpreted as the cause of the fact that in these generated networks there is a very high proportion of small zones, such as trivial zones (only with ego node), dyads, and triads (zones with two or three nodes), as shown in Table 8. Therefore, the average size of the zone is small, as is the average number of nodes in the zones. Networks also have very few (or no) multi-ego zones. This is because network nodes predominantly have neighbors with which they are not two-way dependent (the two-way dependency of a pair of egos is a condition for both being in the same zone). We will informally refer to networks that predominantly have pairs of two-way independent neighboring nodes as weakly dependent. For this characteristic, we define NetDep – the total network dependency.

Table 6 Properties of generated networks

Network	n	m	Degree		CC	Strong-prominents		Weak-prominents		Modularity
			Max	Avg		Total	Percent	Total	Percent	
BA $_{3}$	10000	29991	1319	5.998	0.059	516	5.1	16	0.3	0.380
BA $_{4}$	10000	39984	1877	7.996	0.106	133	1.3	7	0.7	0.301
TC $_{0.7\ 2}$	10000	19996	40	3.999	0.463	3044	30.4	1314	13.1	0.809
TC $_{0.7\ 3}$	10000	29991	83	5.998	0.375	3679	36.7	1019	10.1	0.675
TC $_{0.7\ 4}$	10000	39984	117	7.996	0.314	3124	31.2	499	4.9	0.572
TC $_{0.97\ 2}$	10000	19996	88	3.999	0.669	2093	20.9	1619	16.9	0.945
TC $_{0.97\ 3}$	10000	29991	117	5.998	0.618	2472	24.7	2404	24.0	0.899
TC $_{0.97\ 4}$	10000	39984	193	7.996	0.581	2730	27.3	2355	23.5	0.793
ER $_{0.0005}$	9936	25109	15	5.054	0.001	1702	17.1	0	0	0.439
ER $_{0.001}$	10000	49957	23	9.991	0.001	45	4.5	0	0	0.259
football	115	613	12	10.660	0.404	4	3.4	0	0	0.604

When compared to the three small networks (see Table 1), a higher proportion of prominent nodes was found for the TC model, where the networks have a community structure represented by a high modularity value

Table 7 Node dependencies and network dependency in generated networks

Network	OwDep		OwIndep		TwDep		TwIndep		NetDep
	Max	Avg	Max	Avg	Max	Avg	Max	Avg	
BA_{3}	3	0.261	386	0.261	1	0.000	933	5.476	0.087
BA_{4}	4	0.184	392	0.184	0	0.000	1532	7.629	0.046
$\text{TC}_{0.7\,2}$	3	1.088	16	1.088	2	0.284	29	1.540	0.615
$\text{TC}_{0.7\,3}$	4	1.022	21	1.022	2	0.104	66	3.850	0.358
$\text{TC}_{0.7\,4}$	4	0.591	9	0.591	2	0.048	110	6.767	0.154
$\text{TC}_{0.97\,2}$	3	1.224	28	1.224	2	0.469	62	1.082	0.729
$\text{TC}_{0.97\,3}$	4	1.691	39	1.691	2	0.309	88	2.307	0.615
$\text{TC}_{0.97\,4}$	5	1.740	45	1.740	2	0.241	153	4.277	0.465
$\text{ER}_{0.0005}$	2	0.191	4	0.191	1	0.008	15	4.665	0.077
$\text{ER}_{0.001}$	2	0.004	1	0.004	0	0.000	23	9.982	0.001
football	2	0.035	1	0.035	0	0.000	12	10.591	0.007

In the networks that were generated, there are no nodes that have a high number of neighbors on which they are dependent. Moreover, the ER and BA networks have low average values of OwDep, TwDep, and OwIndep, and also a very low NetDep value

Definition 8 *Network dependency. Let m^* be the number of edges connecting mutually (two-way) independent nodes and m is the number of all the edges of the network. Then network dependency NetDep is defined as follows:*

$$NetDep = 1 - \frac{m^*}{m}. \tag{6}$$

The NetDep value is from the interval $[0, 1]$, and the lower it is, the higher the proportion of two-way independent neighbors, while the overall dependency of the network becomes weaker. NetDep is low for connected networks with a high proportion of small zones. On the other hand, low NetDep value does not automatically imply a higher fraction of trivial zones. Even though individual nodes may not be dependent on most of their neighbors, they may have neighbors on which they are dependent or vice versa. In this case, as shown in Observation 2, the low NetDep value influences, in particular, the quality (embeddedness) of the zones. Our experiments show that ER and BA networks are weakly dependent because the NetDep value is very low (see Table 7). The same does not apply to TC networks.

Table 8 Zone properties in generated networks

Network	Zones	Zone size		Inner-zone size		Outer-zone size		Trivial	Dyad	Triad	Multi ego	Embed.
		Max	Avg	Max	Avg	Max	Avg					
BA_{3}	9999	555	1.680	390	1.262	165	0.418	8228	65	1291	1	0.002
BA_{4}	10000	450	1.421	393	1.184	57	0.237	8545	1011	169	0	0.000
$\text{TC}_{0.7\,2}$	8581	26	4.623	18	2.664	12	1.960	219	37	4103	1172	0.438
$\text{TC}_{0.7\,3}$	9479	39	4.246	24	2.213	16	2.033	1325	404	2169	483	0.315
$\text{TC}_{0.7\,4}$	9762	19	2.844	10	1.666	11	1.178	3403	2323	372	231	0.173
$\text{TC}_{0.97\,2}$	7656	55	5.160	38	3.225	17	1.935	16	13	2828	1768	0.483
$\text{TC}_{0.97\,3}$	8453	83	6.564	48	3.424	36	3.139	64	81	319	1237	0.411
$\text{TC}_{0.97\,4}$	8797	99	6.723	50	3.278	50	3.446	247	775	255	991	0.331
$\text{ER}_{0.0005}$	9897	8	1.553	5	1.202	3	0.351	7087	570	1998	39	0.085
$\text{ER}_{0.001}$	10000	3	1.013	2	1.004	2	0.008	9930	10	60	0	0.001
football	115	3	1.087	2	1.035	2	0.052	108	4	3	0	0.008

In the ER and BA networks, there is a very high proportion of small zones, such as trivial zones, dyads, and triads and, conversely, a low proportion of multi-ego zones. Moreover, the embeddedness of these networks is extremely low

Table 9 Memberships in generated networks

Network	Membership		Liaisonship		Co-liaisonship	
	Max	Avg	Max	Avg	Max	Avg
BA $_{3}$	552	1.680	551	0.416	2	0.002
BA $_{4}$	450	1.421	449	0.235	2	0.001
TC $_{0.7\ 2}$	20	3.967	19	1.568	4	0.113
TC $_{0.7\ 3}$	34	4.025	33	1.757	6	0.170
TC $_{0.7\ 4}$	19	2.776	18	1.064	5	0.086
TC $_{0.97\ 2}$	43	3.950	42	1.251	5	0.230
TC $_{0.97\ 3}$	67	5.548	66	2.105	9	0.549
TC $_{0.97\ 4}$	81	5.915	80	2.409	12	0.623
ER $_{0.0005}$	8	1.547	7	0.349	0	0.000
ER $_{0.001}$	3	1.013	2	0.008	0	0.000
football	3	1.087	2	0.052	0	0.000

The ER networks have very low maximum membership of nodes in zones, and all the networks that are generated have a low maximum co-liaisonship. Moreover, the ER and BA networks have extremely low (or equal to zero) co-liaisonship

Table 10 summarizes the results of the detection of zone overlaps and the detection of zones within the overlaps in the networks that are generated. The ER networks have a low maximum zone size and node memberships in zones (see Table 9); as can be seen in Table 10, no overlaps exist in these networks. Conversely, there are large zones in the BA networks, and it can be noticed that despite the relatively large maximum size of the zone overlap (75 or 73 nodes), there are no large zones in the overlaps (the maximum zone size in the overlap is 5 or 7 nodes respectively). The networks generated by the TC model differ from ER and BA networks. There are relatively small overlaps (the maximum overlap size is 24 nodes), but they may contain zones of comparable size (up to 11 nodes). For most of the properties, TC networks resemble the three small networks analyzed above and, as will be seen later, large-scale real-world networks.

The small maximum sizes of overlaps in TC networks are related to the low OwDep and TwDep values (see Table 7), which represent the dependency of one node on multiple nodes. The first property increases the chance of there being more nodes in liaison or co-liaison roles in the surroundings of the node. These are the primary cause for the

Table 10 Properties of zone overlaps in generated networks

Network	Overlaps	Overlap size		Zones in overlaps		Zones in O size	
		Max	Avg	Total	Percent	Max	Avg
BA $_{3}$	707	75	3.833	132	18.7	5	3.788
BA $_{4}$	169	73	7.385	64	37.9	7	4.609
TC $_{0.7\ 2}$	774	10	2.535	372	48.1	8	3.839
TC $_{0.7\ 3}$	807	10	2.615	331	41.0	7	4.299
TC $_{0.7\ 4}$	57	7	3.456	36	63.2	7	4.667
TC $_{0.97\ 2}$	1111	13	2.809	527	47.4	8	3.947
TC $_{0.97\ 3}$	7419	17	2.549	2077	28.0	11	4.847
TC $_{0.97\ 4}$	12037	24	2.524	2843	23.6	11	5.534
ER $_{0.0005}$	0	0	0.000	0	0	0	0.000
ER $_{0.001}$	0	0	0.000	0	0	0	0.000
football	0	0	0.000	0	0	0	0.000

The ER networks have no overlaps and, conversely, the BA networks have large overlaps; however, no large zones are contained in these overlaps. The networks generated by the TC model have relatively small overlaps, but they may contain zones of comparable size overlapping of multiple zones. The TwDep property then ensures that there are pairs of nodes that are part of the same inner or outer-zone.

Random networks generated by ER model are not scale-free and do not have a community structure; BA networks are scale-free but are known to have no community structure (they have low modularity). It is, therefore, a question what properties would be possessed by networks which do not have a high proportion of small zones. A natural expectation may be that, for non-trivial zones to exist, it should suffice for the network to have a community structure. On a well-known football network that has a community structure (its modularity is 0.604; see Fig. 8) it can be seen that it does not.

The results of the analysis of the football network are summarized in the last row of the tables with generated networks. It can be seen that in the network with 115 nodes, there are only seven non-trivial zones with a maximum size of three. It can be concluded that apart from community structure, more varied occurrences of differently-sized zones are determined by other properties. From the point of view of NetDep and four parameters based on different types of dependencies between pairs of nodes (see Table 7), the football network is closest to the ER networks, less to the BA networks, and the least to the TC networks. The low NetDep value and the low averages of OwDep, OwIndep and TwDep, and, conversely, the high average of TwIndep for the football, ER and BA networks, affect together the small or zero percentage of strongly-prominent and weakly-prominent nodes, and therefore, the low share of the centers in these networks (see Table 6). But as can be seen, the football network does not contain any centers. Consequently, we can assume that the size of the zones (as well as the structure associated with their overlaps) is influenced by two factors.

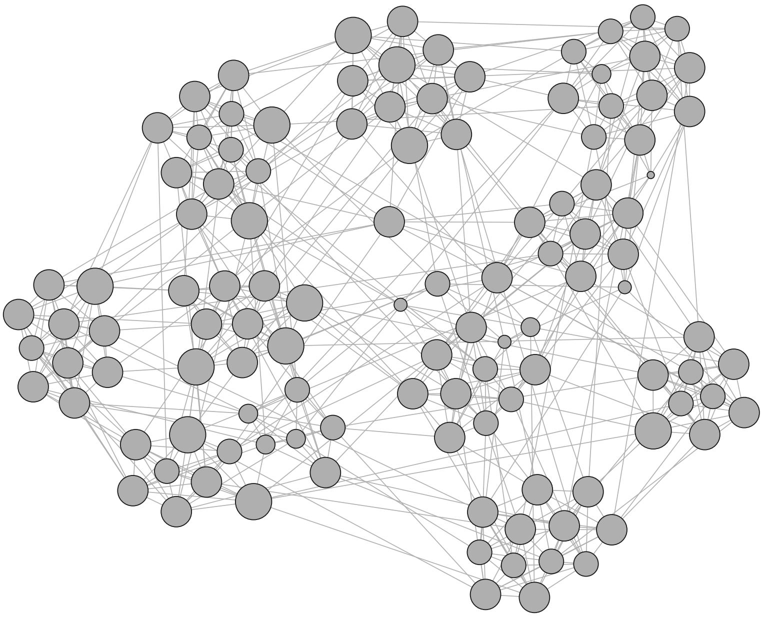

Fig. 8 The football network. Zones are not communities; no larger zones can be detected in the football network. It is evident that the network has a community structure. However, the network almost does not include central nodes in communities or centers linking communities; the node size is proportional to its degree and, as can be seen, most of the nodes have a similar degree

Observation 1 *Zones are not communities: Zones cannot, in general, be considered communities. Large zones exist in networks that have a community structure and, moreover, centers representing both global and local authorities, i.e., strongly and weakly-prominent nodes with a higher degree.*

To further assess the properties of the networks concerning the parameters associated with the detected zones, Figs. 9 and 10 show the distributions (CCDF – complementary cumulative distribution function) and plots of selected properties of four of the generated networks described above.

The upper row in Fig. 9 shows the degree distributions and distributions of four zone-related properties (zone size, liaisonship, co-liaisonship and membership). For the ER networks analyzed here, there are no nodes in the co-liaison role, for the BA networks only a small number of them and only exceptionally in more than two zones. This indicates that in these two networks, there are virtually no dependencies of weakly-prominent nodes on both types of prominent nodes. But as we show in "Zones in real-world networks" section, the dependency between strongly-prominent and weakly-prominent nodes is a common feature in real-world networks (see Observation 3). An example may be the dependencies of node 4 on nodes 2 and 3, node 33 on node 34 and node 2 on node 1 in the Karate Club network (see Fig. 2).

The lower row shows the distributions of properties related to types of node dependencies on their neighbors. As shown in Table 7, for the BA and TC networks there is a non-trivial number of nodes that have a larger number of neighbors with a higher maximum and average value of the OwIndep property than for the ER networks. This can be

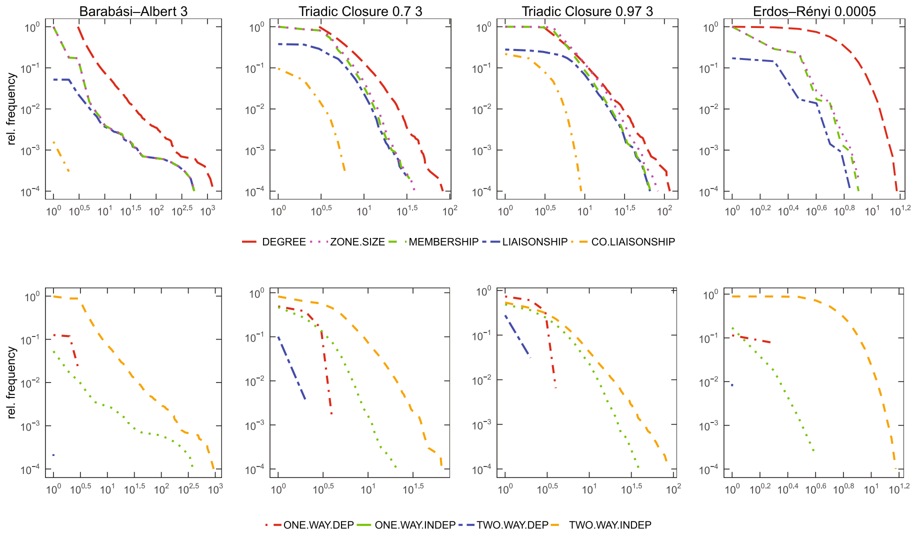

Fig. 9 Node and zone properties in generated networks. The upper row shows distributions of degree and four zone-related properties of four selected generated networks. It can be seen that the shapes of these four distributions copy the shape of the degree distribution. Moreover, it is clear that three of them (zone size, liaisonship, membership) have almost identical distributions, while for liaisonship it is true approximately from the value of 10. The lower row shows the distributions of properties related to types of node dependencies. A common characteristic of all the networks is a very low (or zero) number of neighbors on which the nodes are dependent (OwDep and TwDep properties). Moreover, the shape of the distribution of the TwIndep property copies the shape of the node degree distribution, with nodes not being dependent on most of their neighbors (higher TwIndep values prevail)

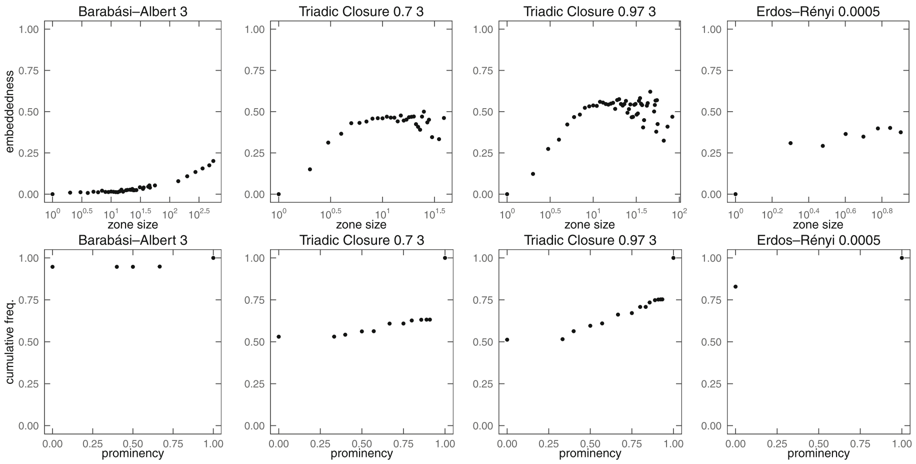

Fig. 10 Embeddedness and prominency in generated networks. The average embeddedness in relation to the size of the zone is shown in the upper row of plots. In the ER networks, there are only very small zones (up to ten nodes), with the average zone embeddedness increasing with the size of the zone and the maximum being around 0.4. A similar behavior is seen in the BA network with the modularity of 0.380, where the maximum average zone embeddedness is only around 0.2. The TC networks show a growing trend of the average embeddedness value up to the zone size being around 10 nodes; then the trend is rather reversed. However, the zone embeddedness value is higher than for the ER and BA networks, which is also associated with high modularity (see Table 6). At its peak, the average value of the zone embeddedness exceeds 0.5 or 0.6 for networks with the parameters 0.7 3 or 0.97 3 and with the modularity of 0.675 or 0.899. The occurrence of nodes with a given prominency in generated networks is shown in the lower row of cumulative distributions. In ER and BA networks, which are weakly dependent, nodes with a prominency value of 0 prevail, and most or all of the other nodes have a value of 1. It is different in TC networks, which have a larger number of weakly-prominent nodes with a prominency value between 0 and 1

interpreted as the existence of network centers; in the BA network, they are nodes with up to hundreds of neighbors, while in the TC networks there are dozens of neighbors. Besides, the TC networks show that the distribution of OwIndep and OwDep properties is almost identical for networks with a stronger community structure (higher modularity). This means that one-way dependencies increase at the expense of independency, which can also be confirmed by the NetDep value in Table 7, which, for TC $_{0.7\ 3}$, is equal to 0.358 and, for TC $_{0.97\ 3}$, is equal to 0.615 (for BA $_3$ it is 0.087).

The upper row in Fig. 10 displays plots showing the relationship between zone size and zone quality measured by the average zone embeddedness. The lower row contains cumulative distributions (CDF) expressing the frequency of occurrences of nodes with a given prominency value. We further show that higher occurrence and diversity in prominency values are a significant characteristic of networks resulting from human interaction (see Observation 5).

Zones in real-world networks

There are three key findings related to the analysis of the three small and ten generated networks. The first is the effect of the weakly dependent network (low NetDep value) on zone quality measured by embeddedness and on the ratio of trivial and generally small zones. The second finding is the existence of zone overlaps. For the small and TC networks, there are larger zones inside overlaps. The third is the finding that zones cannot be considered communities. The prerequisite for the existence of the zones is not only the community structure but also the more complex dependencies in the network. We continue to study these findings when analyzing real-world networks, mainly focusing on differences between them. Besides, we focus more on overlaps and their sizes and densities compared to zone densities, and moreover, on the relationship between zones and non-overlapping or ground-truth communities, respectively.

For the experiments, we used a total of 16 known networks serving different analytical purposes. They are three collaboration networks (astro-ph, cond-mat, cond-mat-2005), five communication (Brightkite, Email-Enron) and social (artist, facebook, new_sites) networks, two technological networks (as-22july06, power), four biological networks (ChCh-Miner, PP-Decagon, PP-Pathways, Yeast) and two networks constructed from ground-truth communities (com-amazon, com-dblp). For details of the individual datasets, see Appendix B. It is natural that the networks that are analyzed differ in their structure, which is also affected by the way the networks were constructed. The results of the analysis of real-world networks are summarized in Tables 11, 12, 13, 14, and 15.

Table 11 shows the percentage of nodes with a non-zero prominency, i.e., strongly-prominent and weakly-prominent nodes. In these values, the collaboration networks and networks with ground-truth communities differ from others. For each network of these two types, there are higher proportions of weakly-prominent nodes (9 percent or more), which is not the case for other networks. For networks with ground-truth communities, the share of prominent nodes is slightly lower, which is also reflected in the lower average value of co-liaisonship (see Table 14). All of these networks represent the result of human activities. In collaboration networks, it is the direct activity of the authors associated with the publication of research articles. For both networks with ground-truth communities, specific activities are involved. In the com-amazon dataset, the result of this activity is a network of products that people buy together (co-purchasing). At com-dblp, the network

Table 11 Properties of real-world networks

Network	n	m	Degree		CC	Strong-prominents		Weak-prominents		Modularity
			Max	Avg		Total	Percent	Total	Percent	
artist	50515	819090	1469	32.438	0.137	8755	17.3	325	0.6	0.604
as-22july06	22963	48436	2390	4.218	0.230	2416	10.5	363	1.5	0.663
astro-ph	14845	119652	360	16.120	0.669	3291	22.1	3473	23.3	0.755
Brightkite	56739	212945	1134	15.012	0.173	14297	25.1	1961	3.4	0.660
com-amazon	334863	925872	549	5.529	0.396	75603	22.5	43161	12.8	0.926
com-dblp	317080	1049866	343	6.622	0.632	65090	20.5	29139	9.1	0.810
cond-mat	13861	44619	107	6.438	0.651	2241	16.1	3315	23.9	0.862
cond-2005	36458	171735	278	9.420	0.656	6372	17.4	8904	24.4	0.786
email-Enron	33696	180811	1383	21.463	0.509	5180	15.3	2591	7.6	0.584
facebook	4039	88234	1045	43.691	0.605	61	1.5	1242	30.7	0.835
ChCh-Miner	1510	48512	443	64.254	0.304	197	13.0	48	3.1	0.392
new_sites	27917	205964	678	14.776	0.294	7276	26.0	1588	5.6	0.611
power	4941	6594	19	2.669	0.080	1494	30.2	144	2.9	0.935
PP-Decagon	19065	715602	251	75.069	0.233	3500	18.3	147	0.7	0.445
PP-Pathways	21521	338625	213	31.812	0.124	3492	16.2	46	0.2	0.386
Yeast	2224	6609	64	6.339	0.125	563	25.3	38	1.7	0.587
LFR $_{20\ 500\ 2000}$	10000	102054	200	20.411	0.399	1499	15.0	225	2.3	0.659
LFR $_{7\ 60\ 4000}$	10000	32262	99	6.452	0.349	2280	22.8	472	4.7	0.578

The collaboration networks and networks with ground-truth communities differ from others (except facebook) in weakly-prominent nodes. For networks of these two types, there is a higher percentage of weakly-prominent nodes than for other networks. Biological networks, social networks (except facebook), and communication networks have a lower clustering coefficient

Table 12 Node dependencies and network dependency in real-world networks

Network	OwDep		OwIndep		TwDep		TwIndep		NetDep
	Max	Avg	Max	Avg	Max	Avg	Max	Avg	
artist	16	0.308	41	0.308	8	0.019	1463	31.795	0.020
as-22july06	12	1.399	2024	1.399	3	0.038	503	1.383	0.672
astro-ph	55	3.256	136	3.256	30	1.414	239	8.193	0.492
Brightkite	31	0.857	245	0.857	7	0.130	1027	5.661	0.246
com-amazon	6	1.221	348	1.221	6	0.281	201	2.806	0.493
com-dblp	75	1.602	112	1.602	39	0.882	237	2.536	0.617
cond-mat	15	1.648	53	1.648	9	0.686	79	2.456	0.619
cond-2005	24	2.024	83	2.024	15	0.595	228	4.777	0.493
email-Enron	10	1.575	1319	1.575	7	0.426	870	7.155	0.333
facebook	26	1.959	1001	1.959	12	0.236	290	39.536	0.095
ChCh-Miner	16	1.053	62	1.053	3	0.021	432	62.127	0.033
new_sites	27	0.793	107	0.793	13	0.128	672	13.042	0.116
power	4	0.711	12	0.711	3	0.340	10	0.906	0.661
PP-Decagon	180	0.572	267	0.572	16	0.040	2426	73.886	0.016
PP-Pathways	80	0.485	584	0.485	7	0.007	1950	30.493	0.031
Yeast	6	0.819	24	0.819	2	0.040	56	4.265	0.282
LFR $_{20\ 500\ 2000}$	16	1.215	63	1.215	6	0.054	194	17.927	0.122
LFR $_{7\ 60\ 4000}$	10	1.207	42	1.207	3	0.111	74	3.927	0.391

Biological networks, social networks, and communication networks have a low NetDep value. The technological networks have a low average TwIndep compared to the other networks; in both of technological networks, there is the highest NetDep value. Thus, the highest dependencies were found in the technological networks

Table 13 Zone properties in real-world networks

Network	Zones	Zone size		Inner-zone size		Outer-zone size		Trivial	Dyad	Triad	Multi ego	Embed.
		Max	Avg	Max	Avg	Max	Avg					
artist	50126	73	1.885	43	1.325	33	0.560	32169	5701	7455	307	0.037
as-22july06	22522	2328	4.423	2083	2.499	245	1.924	784	8111	10300	430	0.070
astro-ph	11521	530	14.738	199	5.542	331	9.196	247	1316	1566	1792	0.323
Brightkite	53308	375	3.494	265	2.063	110	1.431	6364	21930	12344	2891	0.181
com-amazon	297805	451	4.551	360	2.556	91	1.995	45566	40210	64374	24143	0.316
com-dblp	236688	255	6.557	116	3.530	146	3.027	3585	43712	50846	50311	0.315
cond-mat	10712	91	6.444	59	3.442	48	3.002	166	1635	2122	2200	0.412
cond-2005	29389	271	8.490	117	3.857	154	4.633	483	3664	4755	4870	0.309
email-Enron	28704	1371	6.251	1326	3.056	247	3.195	1309	9837	3467	3299	0.089
facebook	3788	1017	6.587	1002	3.078	42	3.510	30	1635	346	118	0.047
ChCh-Miner	1498	78	3.766	63	2.004	17	1.762	769	209	173	9	0.027
new_sites	26729	177	3.537	110	1.894	67	1.642	9777	3783	4612	758	0.141
power	4120	24	3.674	20	2.382	8	1.292	259	1239	1162	562	0.457
PP-Decagon	18933	614	6.627	269	1.593	378	5.034	12652	1692	1787	76	0.037
PP-Pathways	21471	652	2.939	585	1.491	120	1.448	11135	5060	2992	37	0.025
Yeast	2181	32	3.169	26	1.887	10	1.282	414	796	476	38	0.177
LFR $_{20\ 500\ 2000}$	9848	88	4.591	64	2.235	31	2.356	4819	909	693	57	0.587
LFR $_{7\ 60\ 4000}$	9476	72	4.638	46	2.418	27	2.219	2270	129	3619	430	0.425

All the biological networks (ChCh-Miner, PP-Decagon, PP-Pathways, Yeast), social networks (artist, facebook, new_sites), and communication networks (Brightkite, Email-Enron) have a low average value of zone embeddedness

Table 14 Memberships in real-world networks

Network	Membership		Liaisonship		Co-liaisonship	
	Max	Avg	Max	Avg	Max	Avg
artist	71	1.870	70	0.543	18	0.013
as-22july06	2262	4.338	2261	1.875	12	0.012
astro-ph	272	11.438	271	4.827	207	2.309
Brightkite	354	3.283	353	1.274	76	0.070
com-amazon	429	4.047	428	1.439	29	0.335
com-dblp	177	4.894	176	1.938	96	0.322
cond-mat	72	4.980	71	1.524	62	0.796
cond-2005	170	6.844	169	2.492	99	1.243
email-Enron	1314	5.325	1313	2.457	36	0.265
facebook	993	6.178	992	1.057	69	2.234
ChCh-Miner	77	3.736	76	1.351	48	0.397
new_sites	165	3.386	164	1.402	27	0.171
power	18	3.063	17	1.062	6	0.015
PP-Decagon	589	6.581	588	4.623	351	0.376
PP-Pathways	652	2.932	651	1.434	42	0.010
Yeast	32	3.107	31	1.210	10	0.048
LFR $_{20\ 500\ 2000}$	88	4.521	87	2.554	11	0.066
LFR $_{7\ 60\ 4000}$	68	4.395	67	2.008	11	0.095

Some biological (PP-Pathways, Yeast), technological (as-22july06, power), communication (Brightkite, email-Enron), and social (artist, new_sites) networks have a very low average value of co-liaisonship. The same applies to both the LFR networks

Table 15 Properties of zone overlaps in real-world networks

Network	Overlaps	Overlap size		Zones in overlaps		Zones in O size	
		Max	Avg	Total	Percent	Max	Avg
artist	750	26	3.667	292	38.9	22	6.007
as-22july06	58095	583	2.763	3628	6.2	13	4.089
astro-ph	500677	311	3.705	33119	6.6	203	23.743
Brightkite	27925	114	4.779	4942	17.7	55	10.169
com-amazon	65057	128	4.601	29576	45.5	24	5.498
com-dblp	925712	177	2.859	116482	12.6	163	12.946
cond-mat	27516	47	2.771	4519	16.4	47	8.483
cond-mat-2005	375065	184	2.345	30731	8.2	78	10.124
email-Enron	302590	169	1.817	15609	5.2	41	6.379
facebook	26322	42	1.759	973	3.7	34	12.807
ChCh-Miner	670	73	10.213	238	35.5	43	16.008
new_sites	3819	41	5.933	1637	42.9	28	9.296
power	125	12	3.272	66	52.8	12	4.712
PP-Decagon	86548	382	19.909	4453	5.1	382	128.530
PP-Pathways	17880	122	11.439	2510	14.0	122	43.175
Yeast	227	17	3.802	112	49.3	8	4.688
LFR $_{20\ 500\ 2000}$	5420	81	9.782	2290	42.3	35	8.932
LFR $_{7\ 60\ 4000}$	14577	27	2.060	2331	16.0	17	5.961

Except for the two biological networks, the average size of the overlaps for most networks is small. However, it can be seen in all the networks that a non-trivial number of larger zones exists in the overlaps; some of the networks have zones (nested in overlaps) with a high number of nodes. In the last column, there are average zone sizes detected in overlaps of two other zones. These sizes are greater than the average overlap sizes in 'avg overlap size' column. Thus, the zones exist more likely in larger overlaps. The size of a nested zone can correspond with the size of the entire overlap; e.g., in two biological networks (PP-Decagon and PP-Pathways), the maximum overlap of two zones is also a zone is constructed on the basis of co-authorship activity related to the participation of people in the same conference, or publishing articles in the same journal. All of these networks have a distinct community structure represented by high modularity and a high clustering coefficient (except com-amazon). This is also related to the higher average values of zone embeddedness and NetDep values (see Tables 12 and 13); the lower clustering coefficient of the com-amazon network is projected to a smaller average zone size (see Table 13). Here, let us note that this is due to the low average number of nodes in the outer-zone (1.995). This, in turn, means that most of the zones are connected to their surroundings more weakly than in other networks; this is confirmed by the very high modularity of this network (0.926).

While social, communication, technological and biological networks share a higher proportion of strongly-prominent nodes, they have a very low fraction of weakly-prominent nodes. The only exception is the facebook network. This network, uniquely among the social networks, has a very low fraction of strongly-prominent nodes and a larger fraction of weakly-prominent nodes. This network is an exception in the whole group of real-world networks, which is probably due to the specific construction of the network by merging more ego-networks. It is possible that the ego-networks were chosen in such a way that almost every ego was dependent on some other ego. As a result, there are only a few strongly-prominent nodes in the network. Note also that the facebook network has high modularity but a very low NetDep value and a very low average of zone embeddedness.

What is noteworthy is the relationship between low NetDep values and zone embeddedness. Tables 12 and 13 show that a low NetDep value and a low average value of zone embeddedness have biological networks, social networks, and communication networks. Table 11 also shows that all of these networks, except facebook, have a low clustering coefficient.

Observation 2 *Relationship of NetDep and embeddedness. If the network has a low NetDep value, then it also has a low average of zone embeddedness.*

This observation can naturally be interpreted in such a way that if the network does not have a high degree of dependency between pairs of nodes, it is not possible to assume the frequent occurrence of zones strongly interconnected inwardly and weakly outwardly. However, the opposite relationship between NetDep and zone embeddedness does not apply; e.g., a technological network such as-22july06 has a very low value of zone embeddedness and yet a high NetDep value.

A higher NetDep value is a characteristic feature of technological and collaboration networks, and networks with ground-truth communities. Both technological networks have a low clustering coefficient; the other mentioned networks have a higher clustering coefficient and zone embeddedness. For technological networks, Table 12 also shows that they have (in addition to a lower average degree 4.218 or 2.669, respectively) a lower average TwIndep compared to other networks (1.383 or 0.906, respectively). The two technological networks that were analyzed, therefore, have the highest dependency (the highest NetDep value).

Figures 11 and 12 display the distributions and plots of properties of four selected real-world networks described above (as-22july06, com-dblp, Email-Enron, PP-Decagon). This selection provides both common and distinct characteristics for

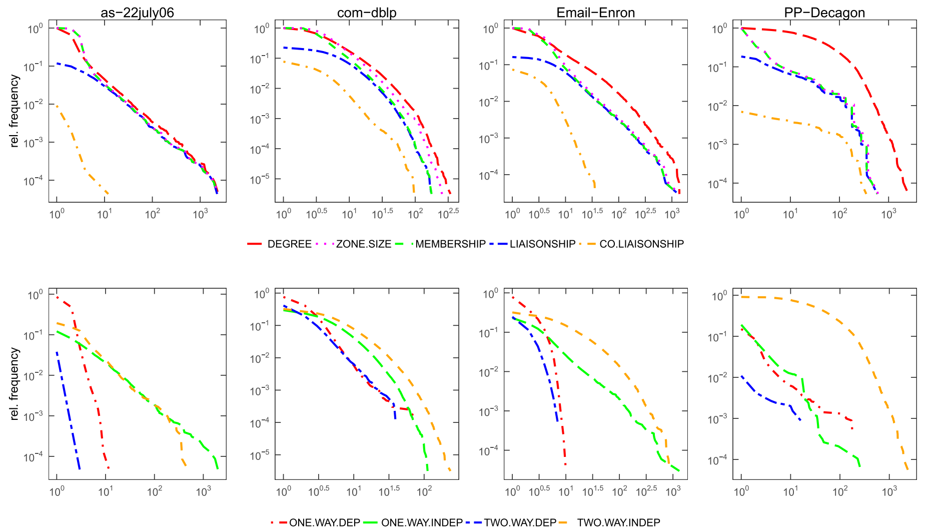

Fig. 11 Node and zone properties in real-world networks. The upper row shows that the distributions of zone-size, membership, and liaisonship roughly copy the shape of the node degree distribution. In as-22july06 and PP-Decagon networks, there is a smaller fraction of nodes in the co-liaison role than in com-dblp and Email-Enron networks. From dependency properties at the lower row plots, it is evident, that PP-Decagon network has a higher proportion of nodes with two-way independent neighbors (TwIndep property)

different types of networks from the perspective of dependencies and detected zones. The distributions and plots of the remaining twelve real-world networks that were analyzed are shown in Appendix C in Figs. 21, 22, 23 and 24.

The upper row of distributions in Fig. 11 shows five properties related to nodes and zones. The four networks that were analyzed (and also the distributions for the other

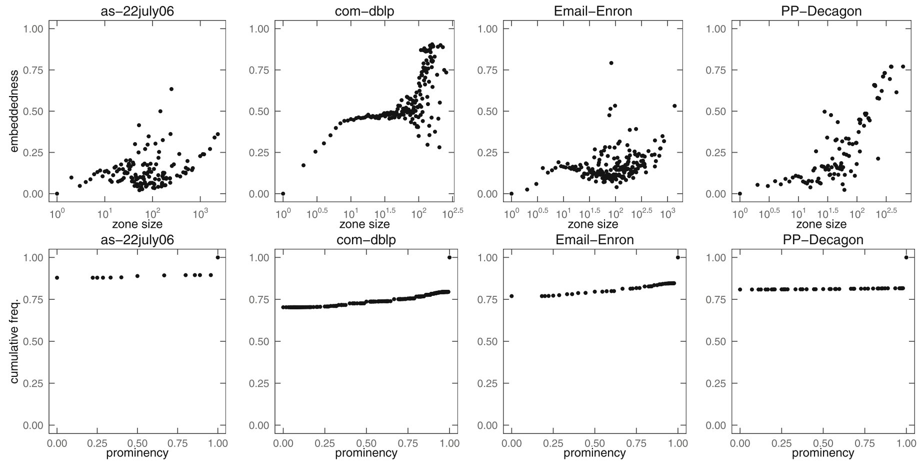

Fig. 12 Embeddedness and prominency in real-world networks. The average embeddedness in relation to the size of the zone is shown in the upper row of plots. Obviously, for all the networks, small zones have a low embeddedness. This is because small zones are often sub-zones of larger zones, and their connection to the surroundings is stronger than in large zones. The occurrence of nodes with a given prominency is shown in the lower row of cumulative distributions. In as-22july06 and PP-Decagon networks, there are not many weakly-prominent nodes (the cumulative distributions are almost not growing). In particular, the technological network as-22july06 has only a few different prominency values. On the other hand, in com-dblp collaboration network, the occurrence of different prominency values is most varied, and frequencies of individual values are higher. For other real-world networks see Figs. 23 and 24 in Appendix C

networks in Appendix C) show that from a relatively low value the distributions of three properties (zone-size, membership and liaisonship) roughly copy the shape of node degree distribution. Conversely, there are two characteristics in which the networks differ as described in Observations 3 and 4.

Figure 11 shows that the technological and biological networks (as-22july06, PP-Decagon and other networks of these types in Appendix C) have a smaller fraction of nodes in the co-liaison role than is the case in collaboration networks (including com-dblp) and Email-Enron communication network (and also other networks with social interaction in Fig. 20 in Appendix C).

Observation 3 *Dependencies in networks with social interaction. In real-world networks, dependencies exist between the liaison nodes, i.e., the mutual dependencies between pairs of co-liaisons or the dependency of co-liaisons on liaisons. In technological and biological networks, however, there is a low proportion of co-liaison nodes; therefore, dependencies between them (e.g., in outer-zones) do not exist so often as in networks resulting from human interaction (e.g., collaboration and communication networks, see Fig. 3).*

When looking at the lower row of distributions of dependency properties, obviously, the low number of co-liaisons is related to a small fraction of nodes, which have at least one two-way dependent neighbor (TwDep property is shown in blue in the lower row of plots).

From distributions in Fig. 11 and Figs.21 and 22 in Appendix C, it can be seen for biological networks that there prevail nodes with two-way independent neighbors. This is projected also to a greater distance of degree distribution from zone-size, membership and liaisonship distributions, especially for those biological networks that also have lower modularity (see Table 11 and degree distributions in plots shown in red).

Observation 4 *Independent neighbors in biological networks. Nodes in biological networks have a higher proportion of neighbors with which they are mutually independent in comparison with technological, collaboration and communication networks.*

However, a similar characteristic to that found in a biological network is found in the social networks artist, new_sites and facebook (see Figs. 21 and 22 in Appendix C). The distributions also show that the Email-Enron network and Brightkite communications network have characteristics at the boundary between collaboration and biological networks and the com-amazon network at the boundary between technological and collaboration networks.

Differences in the individual types of networks are also shown in plots in Fig. 12. The upper row of plots represents zone quality, measured by the average embeddedness in relation to zone size. When the zone size increases, the average embeddedness is more varied, and for most networks, it is slightly higher. For the com-dblp and collaboration networks, however, it is evident that the quality of the large zones is very heterogeneous. This is probably due to differing relationships in large collaborating teams.

The lower row of plots shows cumulative distributions (CDF) expressing the frequency of occurrence of nodes with a given prominency value. Here, the collaboration, communication, and social networks are distinctly different from the technological and biological networks. Plots with the embeddednes and prominency values of the remaining twelve real-world networks are shown in Figs. 23 and 24 in Appendix C).

Observation 5 *Prominency in networks with social interaction. The networks resulting from human interaction have a more varied occurrence of prominency values. This suggests that in these networks, there are more complex non-symmetric dependencies between nodes than in the case of technological and biological networks.*

This observation is due to the fact that technological and biological networks, unlike collaboration, communication and social networks, have much less weakly-prominent nodes with prominency between 0 and 1, i.e., the nodes with a potential to act in the zones as a co-liaison. Additionally, especially in collaboration networks, the occurrence of different prominency values is most varied; it implies high flexibility of connections within zones and between zones.

Zones in LFR networks

The last two rows in Tables 11, 12, 13, 14, 15 with the properties of real-world networks show the results of the analysis of two of LFR benchmark networks that were generated (Lancichinetti and Fortunato 2009). The generator of the LFR networks provides settings ensuring the existence of overlapping communities. Both the networks we analyzed were generated with 10000 nodes. The first network LFR $_{20\,500\,2000}$ was generated to have an average degree of 20, a maximum community size of 500, and a total of 2000 nodes in the overlaps; a total of 186 communities with a minimum size of 7 and an average size of 98 was generated in this network. In the second network LFR $_{7\,60\,4000}$, for an average degree of 7, a maximum community size of 60, and a total of 2000 nodes in the overlaps, a total of 1000 communities with a minimum size of 3 and an average size of 19 was generated. The tables show that the LFR networks, unlike the other generated networks from "Zones in generated networks" section, do not significantly differ from real-world networks in any of the properties that were investigated. The distributions and plots in Fig. 13 confirm the same, describing other properties of the LFR networks. However, two interesting results can be seen.

The first is the maximum size of the zones found in the LFR networks. For the first network (see Table 13), the largest zone is considerably smaller than the largest community (88 vs. 500), while in the second network it is the opposite (72 vs. 60). The second result is a low average value of co-liaisonship (0.066 and 0.095, respectively), which is particularly characteristic of some biological, technological, and social networks (see Table 14). Overall, though, the LFR networks that were generated are closest to the biological networks, as shown in the comparison of the LFR network properties in Fig. 13 with the properties of real-world networks in Figs. 21, 22, 23 and 24 in Appendix C. Unlike the biological networks, however, the LFR networks have a higher NetDep and embeddedness value. From this more detailed view, the LFR networks differ from the other real-world networks that were analyzed.

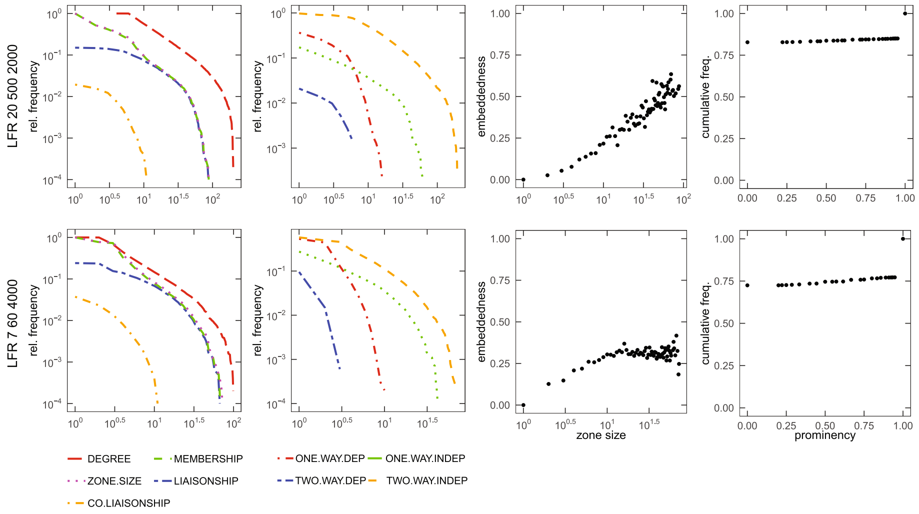

Fig. 13 Properties of LFR networks. The two LFR networks have similar characteristics to the biological networks. However, they have higher average embeddedness, which is mainly due to the higher embeddedness of larger zones with more than ten nodes

Zone overlaps and their density

Recent research on the properties of node groups in networks with ground-truth communities has shown that (1) groups of nodes may overlap, (2) other overlapping groups may exist in overlaps, and (3) overlaps may be denser than overlapping groups. We have dealt with the first two properties in previous experiments with small and generated networks. Table 15 summarizes the results of detecting pairs of overlapping zones with at least ten nodes and zone detection within overlaps with at least four nodes. As shown through the average size of the overlaps, most of them are small (except for biological networks). However, in all the networks, there is a non-trivial number of zones in larger overlaps, and there are also zones with a high number of nodes. This can be read from both the maximum and the average size of zones within overlaps in the last two columns of the table.

In the next experiment, we investigated the density of zones and overlapping zones of the same size. The goal was to verify that the overlaps are more densely connected compared to zones. In the upper row of plots in Fig. 14, the average density of the zones in relation to their size can be seen. The plots in the lower row provide the same information for zone overlaps. The same plots for the remaining twelve real-world networks are shown in Figs. 24 and 26 in Appendix C. When comparing the average densities for zones and overlaps with the same number of nodes, it can be seen that, especially for a smaller number (dozens) of nodes, the average density of the overlaps for most networks is higher than the density of zones of the same size. For large overlaps, although not so clear, the situation is similar. The exceptions are, in particular, the collaboration networks (astro-ph, cond-mat-2005, cond-mat) networks, and also Email-Enron communication network, where the average density of small zones is higher than overlaps of the same size. This is because, in these networks, small zones are often formed by cliques.

Zones in community structure

As mentioned in "Zones in generated networks" section, zones cannot be considered communities. That is why we prepared an experiment in which we applied the Louvain

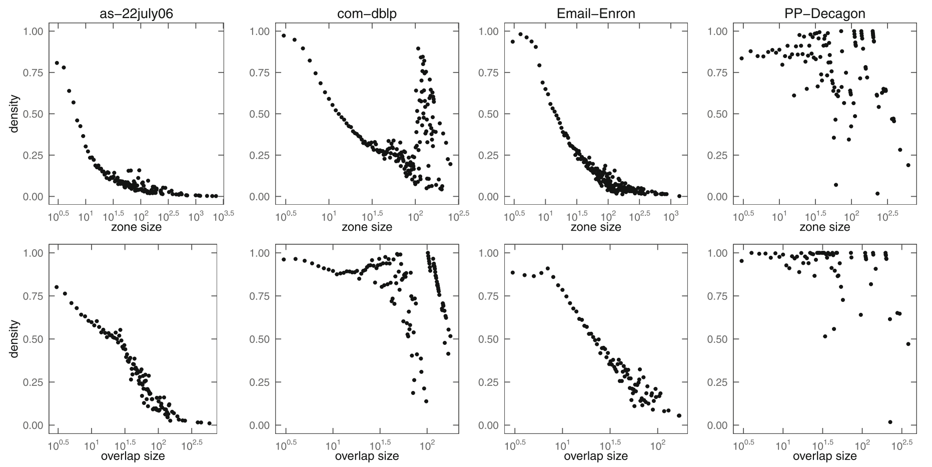

Fig. 14 Density of zones and overlaps in real-world networks. When comparing densities for zones in the upper row and overlaps with the same size in the lower row, it can be noticed that the density of the overlaps tends to be higher than the density of zones of the same size. For other real-world networks see Figs. 25 and 26 in Appendix C

algorithm to detect non-overlapping communities across all of the real-world networks. We then found the best matching zone to every detected community. To compare the community and the zone, we utilized the Matthews Correlation Coefficient (MCC) based on a confusion matrix in which true positives are the number of nodes in the intersection of community and zone ZC, false negatives are the number of zone nodes outside the community Z, false positives are the number of community nodes outside the zone C, and true negatives are the remaining number of nodes outside the zone and the community O (see Equation 7). MCC returns a value between -1 and $+1$; a coefficient of $+1$ represents the perfect fit and -1 indicates total disagreement between community and zone (in our case of overlapping zones and communities, $MCC > 0$). The results are summarized in Table 16.

$$MCC = \frac{ZC \cdot O + Z \cdot C}{\sqrt{(ZC + Z) \cdot (ZC + C) \cdot (O + Z) \cdot (O + C))}} \tag{7}$$

For all the networks that were analyzed (except facebook), the maximum and average community sizes are higher than those for zones. For some networks, even the maximum community size is at tens of thousands of nodes, compared to zones with a maximum of dozens to hundreds of nodes. For facebook, this is the opposite, which is probably due to the construction of this network by merging individual ego-networks.

It can also be seen that the Louvain algorithm provides a relatively small number of predominantly large communities; this is a consequence of the resolution limit issue of modularity. To overcome this problem, it is possible to use the ECG approach (Ensemble Clustering for Graphs, Poulin and Théberge (2018)). Moreover, when many small communities exist in the network, there are approaches working better than the Louvain algorithm (e.g., InfoMap algorithm, Rosvall and Bergstrom (2008)). In our experiments, however, the Louvain algorithm is adequate to assess the match between such detected communities and large zones; we will focus on more communities, including the small ones, in "Zones and ground-truth communities" section.

Table 16 Match of zones and communities detected by Louvain algorithm

Network	Comm.	Comm. size		Zone size		MCC		Modularity	Comm. embed.	Zone embed.
		Max	Avg	Max	Avg	Max	Avg			
anobii	882	37944	177.971	1084	9.672	1.000	0.879	0.573	0.797	0.651
artist	30	10139	1683.833	73	22.033	1.000	0.479	0.604	0.776	0.328
as-22july06	42	5099	546.738	1991	187.690	1.000	0.512	0.663	0.787	0.222
astro-ph	55	648	269.909	382	65.291	0.974	0.423	0.755	0.695	0.640
Brightkite	457	12119	124.155	375	9.626	1.000	0.856	0.660	0.913	0.828
com-amazon	252	11727	1328.821	451	39.222	1.000	0.354	0.926	0.965	0.755
com-dblp	463	33896	684.838	255	27.633	1.000	0.716	0.810	0.910	0.772
com-youtube	9264	197482	122.476	10538	12.265	1.000	0.904	0.686	0.843	0.641
cond-mat	82	759	169.037	75	27.098	1.000	0.413	0.862	0.835	0.704
cond-mat-2005	79	3266	461.494	271	56.190	0.931	0.366	0.786	0.980	0.966
Email-Enron	308	7550	109.403	1371	27.461	1.000	0.891	0.584	0.968	0.902
facebook	16	548	252.438	1017	257.312	1.000	0.727	0.835	0.946	0.644
ChCh-Miner	8	397	188.750	78	31.875	0.636	0.376	0.392	0.616	0.359
new_sites	65	4964	429.492	177	23.738	1.000	0.550	0.611	0.802	0.640
power	40	233	123.525	24	13.675	0.561	0.339	0.935	0.963	0.723
PP-Decagon	20	5649	953.250	614	56.650	1.000	0.550	0.445	0.710	0.568
PP-Pathways	17	4626	1265.941	652	97.941	1.000	0.475	0.386	0.715	0.446
Yeast	24	258	92.667	32	15.417	1.000	0.482	0.587	0.845	0.729
LFR $_{20\ 500\ 2000}$	48	1161	208.333	88	41.208	1.000	0.586	0.659	0.764	0.587
LFR $_{7\ 60\ 4000}$	45	456	222.000	67	37.222	0.534	0.338	0.578	0.614	0.387

The table shows very high agreement between communities and zones for the communications networks (Email-Enron, Brightkite, anobii) and the video-sharing network (com-youtube), and high agreement for the facebook and com-dblp networks. In these networks, the zones correspond well to the communities. This means that in these networks, communities, similarly to zones, form around egos – those nodes on which the nodes in their surroundings are directly or indirectly dependent. However, it should be noted that many other zones were detected in these networks, predominantly nested in the communities and their corresponding zones

The smaller size of zones in comparison with communities raises the question of whether communities are not formed as a union of several zones. However, the question is not an easy one; to answer it, it would be necessary to find egos whose zones could form the community (e.g., similarly to multi-ego-centered communities as described in Danisch et al. (2013)).

An interesting result is that at least one zone that exactly matched one of the communities ($MCC = 1$) was found for almost all networks. The average match values range from 0.339 (power network) to over 0.8 (e.g., Email-Enron network). Compared to the other networks, there is a very high agreement between communities and zones for three communications networks, including Brightkite network (0.856) and anobii network (0.879), which will be described below in “Zones and ground-truth communities” section. Similarly, the facebook network is 0.727 and the com-dblp network is 0.716. In these networks, therefore, the zones correspond very well to the communities. It can, therefore, be assumed that in these cases, communities are formed around egos of the corresponding zones.

In the last two columns of Table 16, the average embeddedness for the detected communities and the corresponding zones is shown. In general, the average quality of communities is higher than the quality of the zones. In most cases, however, the difference in the average embeddedness is not large. The lower quality of the zones can be attributed to another view on a group of nodes. While communities prefer their density and weak interconnections with other neighboring communities, the zones are based on non-symmetric dependencies between nodes. These dependencies are reflected in a much more comprehensive view, in which they play a role concerning the inside of the group (inner-zone) and the dependency outward (outer-zone).

Zones and ground-truth communities

Our approach to detecting groups (zones) presented in this article is based on the assumption that groups existing for some purpose which is represented by an ego (a central node) can be extracted from the structure of a (weighted) network. In our experiments in "Zones in real-world networks" section, we worked with two networks com-amazon and com-dblp; the authors of the article (Yang and Leskovec 2015) identified in these networks ground-truth communities, otherwise referred to as real functional groups, and provided lists of the 5000 ground-truth communities with the highest quality. In addition to these two networks, in the experiment below, we have added two more networks with identified ground-truth communities. The first one is the com-youtube social network, with 1134890 nodes, 2987624 edges, and 5000 communities. The second one is a directed communication network, anobii, which we have transformed into an undirected one so that an edge between nodes only exists if there are mutually directed edges between neighbors. After this transformation and the removal of outliers (isolated nodes), the network has 158330 nodes, 785939 edges, and 4797 communities. For details see Appendix B.

We prepared a similar experiment for these networks as with the non-overlapping communities above. The aim was to find out exactly how the ground-truth communities correspond to the zones and whether we are able to use zones to describe at least part of these communities existing for some real purpose and thus performing some function. The experiment is performed in two steps for each network. First, for every ground-truth community, the zone that best matches this community is found (measured by MCC, see Equation 7). Note that after this step, multiple communities can match the same zone. In the second step, the best-matching community is then selected for each zone from the first step. Theoretically, after the first step, different zones may correspond to different communities, but it may also be that a single zone corresponds to more communities. If the reality is close to the first case, then the communities are unique from the perspective of zones, and vice versa, in the latter case, the communities are very similar (redundant in terms of the agreement of more communities with the same zone). We also conducted this experiment with both the LFR networks, for which we know what overlapping communities were generated. The community properties of these networks, together with the results, are presented in Table 17.

It can be seen that the com-amazon network contains redundant communities from the perspective of zone detection; more than 71% of the communities have no zones that they would match better than other communities. This is probably due to the fact that the network is created from products hierarchically organized into categories and co-purchased together; the categories of products in these hierarchies, which are considered communities, may not be very different in many cases. For the other networks, there is a unique zone for almost every community. The two real-world networks (com-amazon and com-dblp) show a very high match $MCC > 0.9$ for more than half of the detected unique zones corresponding to ground-truth communities. It is worth noting that for the com-dblp network there is an almost 40% perfect match ($MCC = 1$) within 4850 out of all the 5000

Table 17 Zones in networks with ground-truth communities

Network	Communities			Matching zones			MCC	MCC = 1		MCC >0.9	
	Count	Min	Max	Count	Percent	Avg size	Avg	Percent	Avg size	Percent	Avg size
anobii	4797	1	7062	4560	95.06	2.502	0.337	3.49	1.119	3.51	1.150
com-amazon	5000	3	328	1442	28.84	9.549	0.853	44.52	6.467	57.49	7.943
com-dblp	5000	6	7556	4850	97.00	8.721	0.869	39.75	7.487	55.53	7.810
com-youtube	5000	2	2217	4368	87.36	5.485	0.648	18.10	2.217	18.407	2.281
LFR $_{20\ 500\ 2000}$	186	7	500	175	94.09	25.514	0.703	5.71	8.000	16.57	9.690
LFR $_{7\ 60\ 4000}$	1000	3	60	928	92.80	8.179	0.628	3.66	3.147	4.85	3.822

There is a low match of zones and corresponding communities for the anobii network. On the other hand, for the com-amazon and com-dblp networks, the match is high. For anobii and both the LFR networks, only a low proportion of well-matching zones exists; the average size of the matching zones is low for anobii and high for the first LFR Network. Table 16 shows, however, that the match for the com-amazon and LFR networks is much better than for non-overlapping communities detected by the Louvain algorithm. Conversely, for the anobii and com-youtube networks, the match for non-overlapping communities is higher

identified ground-truth communities. Figure 15 shows the relationships between the size of the zone and the average match accuracy (MCC) with the corresponding community for all four real-world networks with ground-truth communities, as well as the frequency of occurrence of zones for the given match accuracy. The corresponding results for the LFR networks are shown in Fig. 16.

In the experiments with communities, we showed two key results. The first one is related to the Observation 1. Even though zones cannot be considered communities, in some networks (especially communication networks, see Table 16), most zones correspond very accurately to non-overlapping communities. E.g., in the anobii network, the zones well describe non-overlapping communities detected by Louvain algorithm; however, they do not well represent the ground-truth communities in this network. The second result shows that the zones provide high potential in describing ground-truth communities in some networks. In the com-amazon and com-dblp networks, zones that

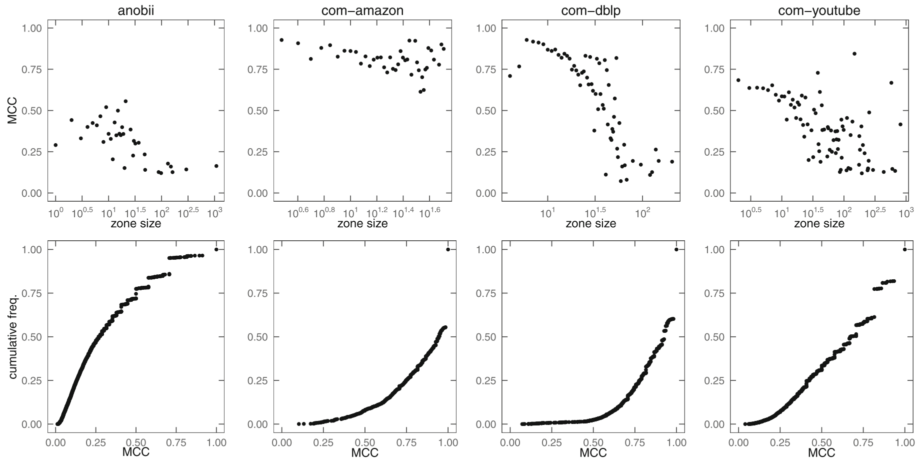

Fig. 15 Match of zones and ground-truth communities. In the upper row of plots, it can be seen that in all cases the average accuracy of the zone and ground-truth community match decreases with zone size. This trend is most noticeable with the com-dblp network; the best-matching zones have around ten nodes. For large zones, the average accuracy of the match in this network is low. The plots in the lower row show the cumulative distributions of the frequencies of zones for a given match accuracy with the best-matching community. Except for the anobii network, the frequency of the zones is balanced or increases with higher match accuracy. This is especially true for the com-dblp network, where there is only a very low number of zones that have a match accuracy below 0.5

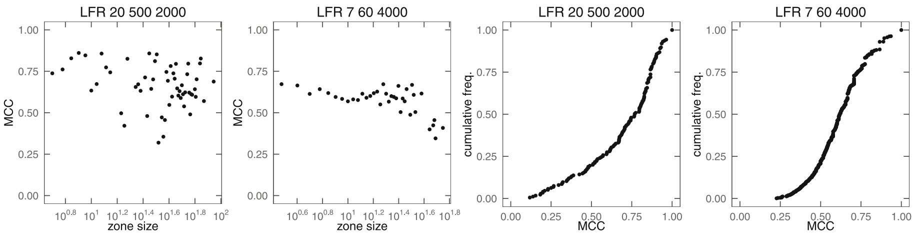

Fig. 16 Match of zones and communities of LFR networks. In LFR networks, there are minimum differences in average match accuracy for different zone sizes. In LFR $_{20\ 500\ 2000}$ network, there is a large variation in zone and community match accuracy. The frequency of zones in LFR networks is increasing for a lower match accuracy; however, it is decreasing for a very high match. For both networks, most of the zones have a match accuracy above 0.5

correspond very closely to a considerable number of identified ground-truth communities were detected. Here it is necessary to remember that each zone is the result of the exact analysis of the surroundings of a selected ego. Therefore, we may conclude that especially the smaller ground-truth communities in these networks are grouped around egos; thus, these egos determine the resulting communities very precisely.

On the other hand, especially for large ground-truth communities, there are often no zones that match with sufficient accuracy. In this case, communities could be, e.g., formed as a union of several zones generated by different egos. The analysis of this case was not, however, the subject of our research.

The results show that our approach cannot be seen as universal from the perspective of ground-truth communities. However, although zone detection is intended for weighted networks, it has been applied to unweighted networks in experiments with real-world networks. The question for further research is what the results of the comparison of zones and ground-truth communities would be in the case of weighted networks.

Summary

One of the interesting results of our experiments with real-world networks is the characteristics of these networks based on the dependency and properties of the detected zones. The Table 18 recapitulates selected outputs from experiments.

Key characteristics of our approach can be summarized into seven points that characterize the description and detection of zones.

1 The relationships between pairs of nodes in zones differ and are not symmetric (non-symmetric similarity).

2 The zone has a clear outer boundary, beyond which nodes are not considered to be members of this zone.

3 The zone contains nodes in different roles. The role of the node is mainly related to its linking in-and-out (i.e., beyond the boundary) of this zone.

4 Zones may have different sizes. They can be both large (dozens or hundreds of nodes and more) and small (triads, dyads and trivial with a single node).

5 Zones may overlap (one node may be in multiple zones), and overlaps may be large (i.e., overlap size is close to the size of overlapping zones).

6 In most cases, overlaps of zones have a higher density than the zones of the same size.

Table 18 Summary of network characteristics

Networks	NetDep	Weak-Prominents (%)	Co-liaisonship Avg	Embed.	CC	Modular.	Non-zero TwDep neighbors	Overlap size	Good match
Collaboration astro-ph cond-mat cond-2005	High	High	High	High	High	High	High	Middle	Low Louvain
Social artist facebook new sites	Low	Low*	Low*	Low	Low*	Middle*	Low	Low	Middle Louvain
Communication Brightkite Email-Enron	Low	Low	Low	Low	Unclear	Middle	Middle	Middle	High Louvain
Technological as-22july06 power	Highest	Low	Low	Unclear	Low	Unclear	Middle	Unclear	Unclear Louvain
Biological ChCh-Miner PP-Decagon PP-Pathways Yeast	Low	Low	Unclear	Low	Low	Low	Lowest	High**	Middle Louvain
Ground-truth com-amazon com-dblp	High	High	Middle	High	Unclear	High	High	Middle	Highest ground-truth
LFR benchmark LFR $_{20\ 500\ 2000}$ LFR $_{7\ 60\ 4000}$	Low	Low	Low	High	Middle	Middle	Low	Unclear	High ground-truth

* except facebook, ** except Yeast

In the individual columns in the table, there are values of selected parameters of networks that we analyzed. Each value of individual parameters comes from the experiments performed; if 'unclear' is listed, the networks are different in this parameter; the value of 'Non-zero TwIndep neighbors' represents the proportion of nodes that have at least one neighbor with which they are mutually dependent; 'Overlap size ' is estimated based on both maximum and average values.

There is a similarity between collaboration and ground-truth networks; however, this is expected because the com-dblp network is also one of the collaboration networks. Social, communication, and biological networks show further similarity; communication networks are distinguished by a high match of zones with non-overlapping communities, and, on the other hand, large overlaps exist in biological networks. Technological networks and LFR benchmark networks differ from these three network groups; technological networks have a high value of NetDep, the total network dependency; the LFR benchmark networks differ in having high embeddedness value. Anobii and com-youtube networks are not listed because they were not the subject of all experiments

7 Overlaps of larger zones with non-trivial structures (e.g., other than small cliques) contain in most cases other zones that can also overlap.

Conclusions

As far as we know, we are the first to focus on analyzing the structures of both weighted and unweighted undirected networks through the non-symmetric similarity between the nodes, which we call dependency. This dependency allows us to clearly describe groups of nodes in the network structure which are organized around one of the group nodes – the ego. To distinguish such groups from communities, we call them ego-zones and examine them both locally and globally.

The local view extends the possibilities of traditional methods for analyzing ego-networks or ego-communities, respectively. Our approach contributes to this, in particular, by preferring weighted networks, exploring the wider surroundings of the ego, and working with nodes in newly defined roles in the zone. Especially in networks with social interaction, thanks to the zones detected for each ego, the analyst gets comprehensive information about the internal non-symmetric dependencies in the zones and thus about the influence of each ego on its surroundings and, vice versa, the influence of its surroundings on it.

From a global perspective, our approach brings, in particular, a typology of network nodes that takes into account their importance on the basis of their structural independency – prominency. Prominent nodes are either entirely or, at least, partly independent on their neighbors and have a significant impact on the node dependencies in their surroundings. The way the zones and prominent nodes are defined using non-symmetric dependencies contributes to understanding why large and dense overlaps of node groups emerge in networks. We consider this finding to be a significant contribution to the analysis of community network structures. Our experiments also show that, in terms of dependencies and overlapping zones, different types of real-world networks have different properties that distinguish them from each other and from generated networks.

The experimental results that we presented also raise questions for future research. Above all, it is a question of more in-depth analysis of the relationships between zones exactly determined by their egos and overlapping or non-overlapping communities in the traditional concept of many different detection methods. Furthermore, it also involves a detailed assessment of how zones and their egos correspond to the existence of ground-truth communities in various types of, in particular, weighted and directed networks.

Appendix A: Zone detection method

The method for detecting all zones in a network with n nodes and m edges consists of two steps. The first step is to calculate the dependency matrix. The second step is the detection of inner and outer-zones for each node (ego). It is not easy to precisely define the time complexity for each step, because the calculations are dependent on the network representation, its complex structure, density, and dependencies between nodes. For this reason, we only describe some of the cases and the result of an experiment providing an estimation of complexity based on the analysis of real-world networks.

To compute the dependency matrix, two dependencies must be calculated for each pair of network nodes. This calculation is related to the detection of the common neighbors of two nodes. In general, a dense network is the worst case and the time complexity of finding the common neighbors of two nodes is $O(n^2)$. Thus, the computation of the dependency matrix has, in the worst-case scenario, the complexity of $O(mn^2)$. For sparse networks, finding common neighbors is related to the average degree of network d, and the time complexity, in this case, is $O(md^2)$. Note, however, that we use the IsDependent relationship to detect zones which, thanks to the threshold 0.5, needs to take into account only a sufficient number of common neighbors in the numerator in Eq. 1 to calculate the dependency. The use of this property also affects the time complexity.

Thanks to the IsDependent relationship, the resulting dependency matrix is a binary one and, therefore, it is the adjacency matrix representing the original network after its transformation into its unweighted directed variant (see Fig. 2a for Karate Club network).

Zone detection is described in Algorithm 1 and, as can be seen, consists of inner-zone and outer-zone detection, and then the zone is a union of both of them.

The time complexity of the inner-zone detection again depends on the network parameters, especially on the density and the number of iterations necessary to be performed. The key is that in each iteration it is necessary to find and test all neighbors of the nodes added in the previous step into the inner-zone. To simplify, it is possible to work with two cases to consider complexity. The first and the worst case is a complete network in which each pair of nodes is two-way dependent. The second case is the sequence of one-way dependent nodes.

In the case of a complete unweighted network, we need two iterations to find one inner-zone. In the first iteration, because of mutual dependency with ego, all nodes of the network (except for the ego) are added into the inner-zone. In the second iteration, for all the added nodes, it is necessary to test whether their neighbors are outside the inner-zone. The time complexity of the detection is, therefore, $O(n^2)$ and also $O(m)$ because the network is dense.

In the case of the sequence of one-way dependent nodes, the number of iterations corresponds to the position of the node in the sequence. However, at most, we need n and a minimum of 1 iteration always working with only one neighbor. In this case, the time complexity of the inner-zone detection is $O(n)$.

The time complexity of the outer-zone detection is based on the previously detected inner-zone for whose nodes it is necessary to find the nodes on which they are dependent and which are outside the inner-zone (and thus form the outer-zone). In the worst case, we can assume that the network is dense, and the inner-zone contains half of the nodes of the network and the remaining nodes of the network form the outer-zone. In this case, it is necessary for each of $\frac{n}{2}$ nodes in inner-zone to test $\frac{n}{2}$ nodes outside the inner-zone; therefore, the outer-zone detection will have the time complexity $O(n^2)$, i.e., not higher than the worst case for inner-zone detection in dense network ($O(m)$).

Assuming that we detect each zone separately, the time complexity of detection zones for all network nodes is, in the worst-case scenario, $O(nm)$ for dense networks. However, we must also consider cases where many nodes in the network can be, e.g., isolated after the transformation of the original network into its unweighted directed variant. In this case, the zone contains only the ego node, and the time complexity of the inner and outer-zone detection is $O(1)$. Therefore, we can expect that the time complexity will be lower for real-world networks.

Figure 17 shows the time needed for the computation of the dependency matrix and the time of detection of zones for each node in the network. Moreover, the green dotted

Algorithm 1: Zone detection

```
Input   : Network $G = (V, E)$. Node $ego \in V$
Output: NodeSet $zone, innerZone, outerZone$

create NodeSet $zone$
create NodeSet $innerZone$
create NodeSet $outerZone$

// detection of inner-zone
create NodeSet $nodesLastAdded$
create NodeSet $nodesOutside$
AddNode($ego \rightarrow innerZone$)
AddNode($ego \rightarrow nodesLastAdded$)

while IsNotEmpty($nodesLastAdded$) do
    foreach Node $v \in nodesLastAdded$ do
        foreach Node $adj \in$ NeighborsOf ($v$) do
            if IsDependent($adj$ on $v$) and $adj \notin innerZone$ then
                AddNode($v \rightarrow nodesOutside$)
            end if
        end foreach
    end foreach
    AddAllNodes($nodesOutside \rightarrow innerZone$)
    Clear($nodesLastAdded$)
    AddAllNodes($nodesOutside \rightarrow nodesLastAdded$)
    Clear($nodesOutside$)
end while

// detection of outer-zone
foreach Node $v \in innerZone$ do
    foreach Node $adj \in$ NeighborsOf ($v$) do
        if IsDependent($v$ on $adj$) and $adj \notin innerZone$ then
            AddNode($adj \rightarrow outerZone$)
        end if
    end foreach
end foreach

// composition of zone
AddAllNodes($innerZone \rightarrow zone$)
AddAllNodes($outerZone \rightarrow zone$)
```

line shows that the estimated time complexity $O(m \log n)$ corresponds well to the dependency matrix computation for the real-world networks that were analyzed. Similarly, the blue dotted line shows the estimated time complexity $O(m)$ for the zones detection in all of the networks (that is $O(\frac{m}{n}) = O(d)$ to detect one zone). Therefore, both the time complexities of the dependency matrix calculation and the zones detection can be assumed to be considerably lower for analyzed real-world networks than the above-mentioned worst cases.

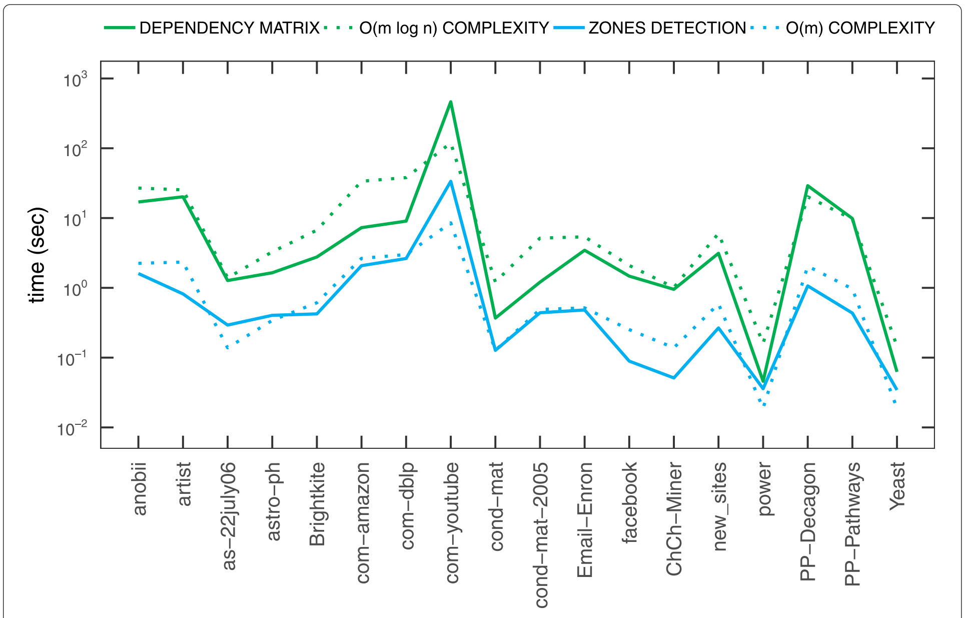

Fig. 17 Estimated time complexity. The dotted lines represent estimated time complexities approximating computation time of the dependency matrix calculation (green) and detection of all zones (blue) in the networks that were analyzed

Appendix B: Publicly archived datasets

Zachary's karate club Zachary (1977) - network of friendships between the 34 members of a karate club at a US university. Available at http://www-personal.umich.edu/~mejn/netdata/

Les Misérables Knuth (1993) - coappearance network of characters in the novel Les Misérables. Available at http://www-personal.umich.edu/~mejn/netdata/

Net Science Newman (2006) - a coauthorship network of scientists working on network theory and experiment. Available at http://www-personal.umich.edu/~mejn/netdata/

American College football Girvan and Newman (2002) - network of American football games. Available at http://www-personal.umich.edu/~mejn/netdata/

as-22july06 - a symmetrized snapshot of the structure of the Internet at the level of autonomous systems, reconstructed from BGP tables posted by the University of Oregon Route Views Project. Available at http://www-personal.umich.edu/~mejn/netdata

com-dblp Yang and Leskovec (2015) - a co-authorship network where two authors are connected if they publish at least one paper together. Publication venue, e.g, journal or conference, defines an individual ground-truth community. Available at https://snap.stanford.edu/data/com-DBLP.html

Email-Enron Yang and Klimmt (2004); Leskovec et al. (2009) - communication network that covers all the email communication within a dataset of around half million emails. Available at https://snap.stanford.edu/data/email-Enron.html

PP-Decagon_ppi Zitnik et al. (2018) - a protein-protein association network that includes direct (physical) protein-protein interactions, as well as indirect (functional) associations between human proteins. Available at https://snap.stanford.edu/biodata/datasets/10008/10008-PP-Decagon.html

artist Rozemberczki et al. (2018) - mutual like networks among verified Facebook pages – the types of sites included TV shows, politicians, athletes and artists among others. Available at https://snap.stanford.edu/data/gemsec-Facebook.html

astro-ph Newman (2001) - weighted network of coauthorships between scientists posting preprints on the Astrophysics E-Print Archive. Available at http://www-personal.umich.edu/~mejn/netdata/

Brightkite Cho et al. (2011) - undirected friendship network of Brightkite users (Brightkite was a location-based social networking website). Available at https://snap.stanford.edu/data/loc-Brightkite.html

com-amazon Yang and Leskovec (2015) - based on "Customers Who Bought This Item Also Bought" feature of the Amazon website. If a product i is frequently co-purchased with product j, the graph contains an undirected edge from i to j. Each product category provided by Amazon defines each ground-truth community. Available at https://snap.stanford.edu/data/com-Amazon.html

cond-mat Newman (2001) - network of coauthorships between scientists posting preprints on the Condensed Matter E-Print Archive. Available at http://www-personal.umich.edu/~mejn/netdata/

cond-mat-2005 Newman (2001) - update network of coauthorships between scientists posting preprints on the Condensed Matter E-Print Archive. Available at http://www-personal.umich.edu/~mejn/netdata/

facebook_combined Leskovec et al. (2009) - dataset consists of 'circles' (or 'friends lists') from Facebook,
https://snap.stanford.edu/data/egonets-Facebook.html

ChCh-Miner_drugbank-chem-chem Wishart et al. (2017) - network of interactions betweeen drugs, https://snap.stanford.edu/biodata/datasets/10001/10001-ChCh-Miner.html

new_sites Rozemberczki et al. (2018) - datasets represent blue verified Facebook page networks of different categories. Nodes represent the pages and edges are mutual likes among them. Available at https://snap.stanford.edu/data/gemsec-Facebook.html

Power grid Watts and Strogatz (1998) - unweighted network representing the topology of the Western States Power Grid of the United States. Available at http://www-personal.umich.edu/~mejn/netdata/

PP-Pathways_ppi Agrawal et al. (2018) - protein-protein interaction network that contains physical interactions between proteins that are experimentally documented in humans https://snap.stanford.edu/biodata/datasets/10000/10000-PP-Pathways.html

Yeast Bu et al. (2003) - protein-protein interaction network in budding yeast. Available at http://vlado.fmf.uni-lj.si/pub/networks/data/bio/Yeast/Yeast.htm

anobii Aiello et al. (2012) - social network (aNobii.com) of book recommendation. Two types of networks are available. Network composed by union of friendship and neighborhood links is the first. The second one is communication network representing message exchanges. Available (on request) at https://www.icwsm.org/2016/datasets/datasets/

com-youtube Mislove et al. (2007) - social network representing a video-sharing web site. Available at http://snap.stanford.edu/data/com-Youtube.html

Appendix C: Supplementary figures

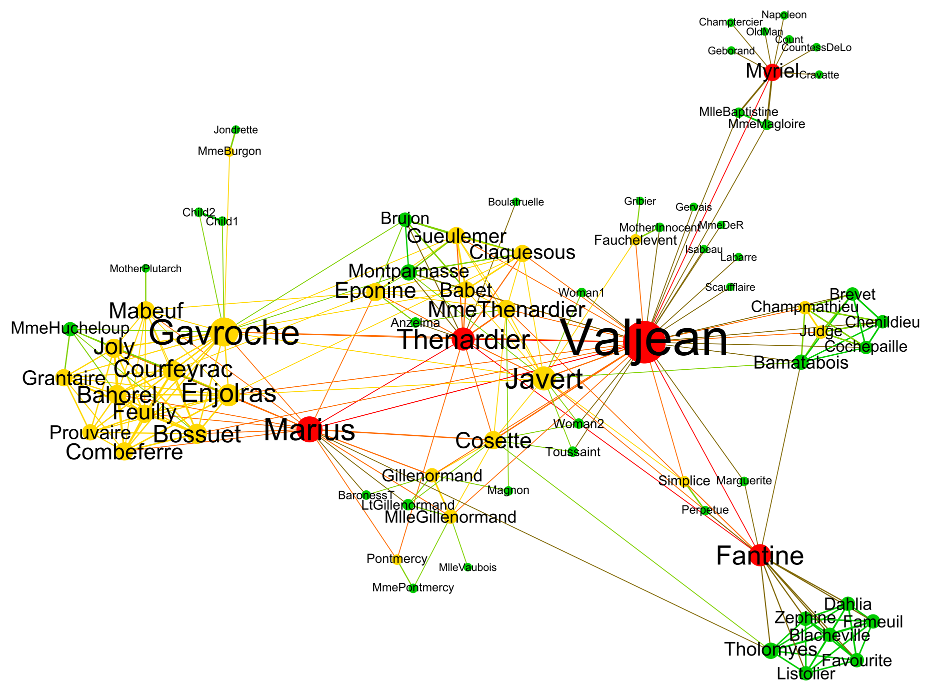

Fig. 18 Les Misérables network. Strongly-prominent nodes (red) link densely connected groups in the network. Weakly-prominent nodes (yellow) occur in different situations where they can be mutually dependent or dependent on some strongly-prominent nodes. However, at least one other node must be one-way dependent on each of them. On the left, there is a structure composed of predominantly weakly-prominent nodes; the zone network in Fig. 7b shows that in such a structure, there are several multi-ego and alternative zones

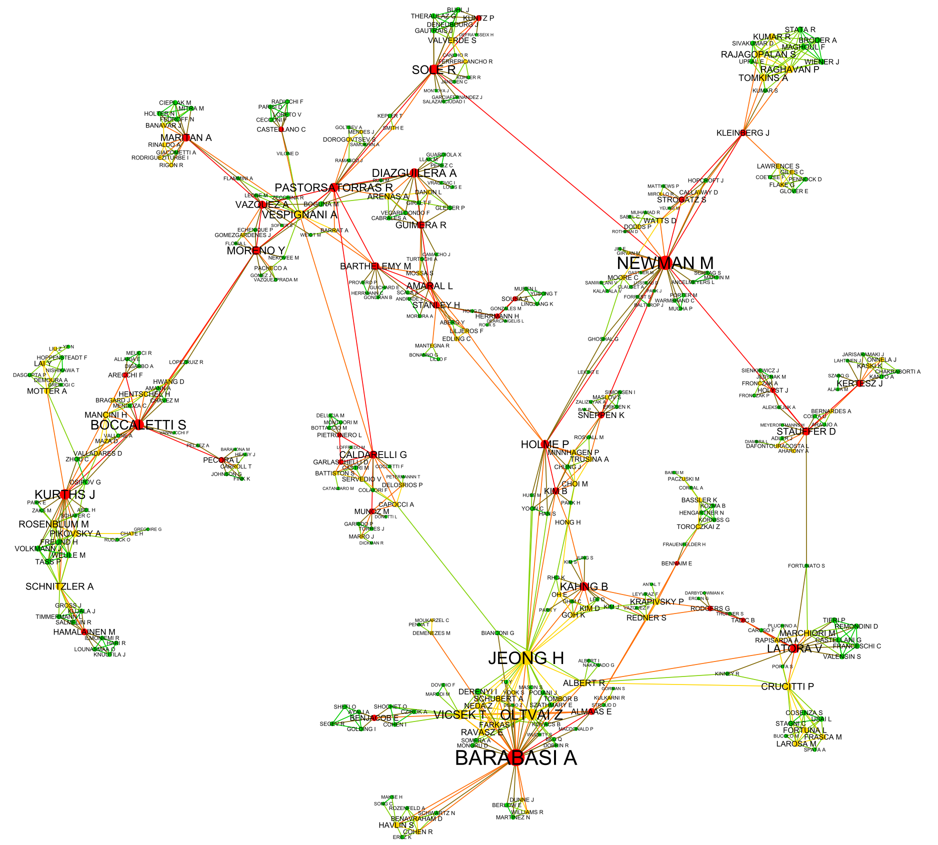

Fig. 19 The largest connected component of Net Science network. The researchers represented by strongly-prominent nodes (red) are independent on the other people in the network. Key researchers can be recognized by their high degree, e.g., Barabási and Newman. There are also weakly-prominent nodes (yellow) with a high degree, e.g., Jeong, Oltvai, and Vicsek. However, all of them are one-way dependent on Barabási

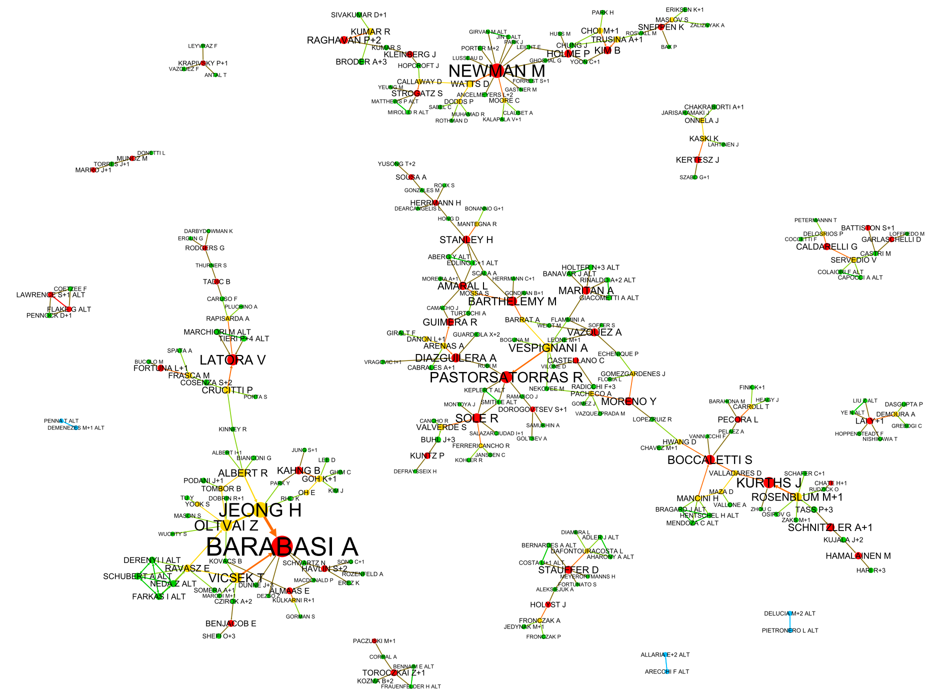

Fig. 20 Net Science zone network. The zone network provides information on the hierarchy of sub-zones, multi-ego zones, alternative zones, and zones in overlaps. Thus, the nodes represent zones and the directed edges point from the sub-zone to its closest super-zone (or to more super-zones when the zone is inside their overlap). The node size corresponds to the number of nodes in the zone; the edge weight corresponds to the number of nodes in the sub-zone. It can be seen that the Net Science network, from the perspective of the zones, has partitioned into differently-sized clusters which uncover the structure of ego-centered cooperating teams and their sub-teams. Two clusters have distinct centers with large zones (Barabási and Newman). On the contrary, the large cluster in the middle contains rather small zones

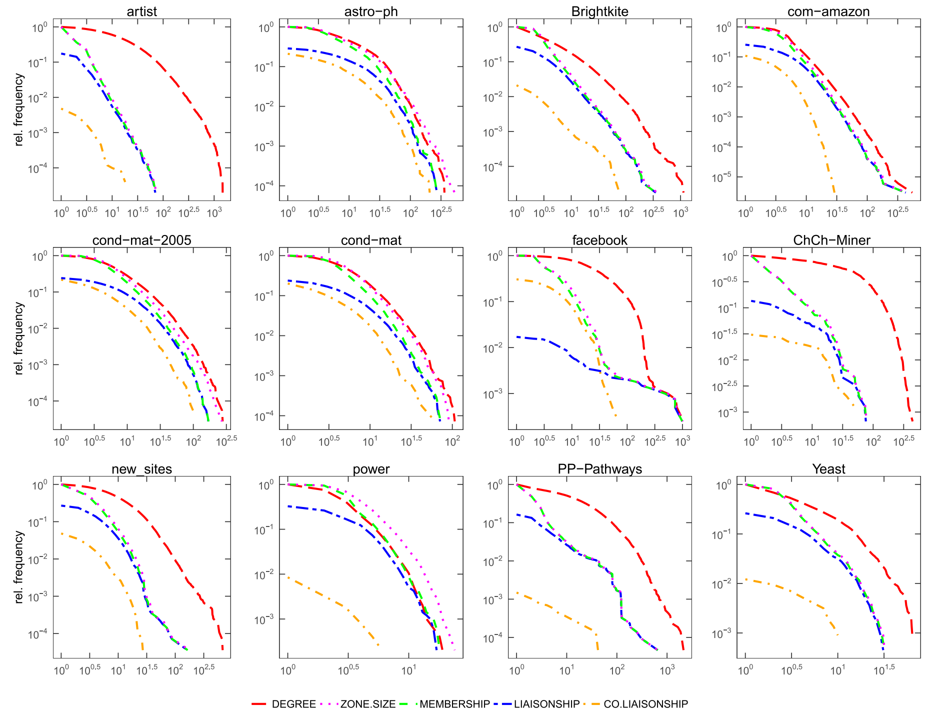

Fig. 21 Node and zone properties in real-world networks

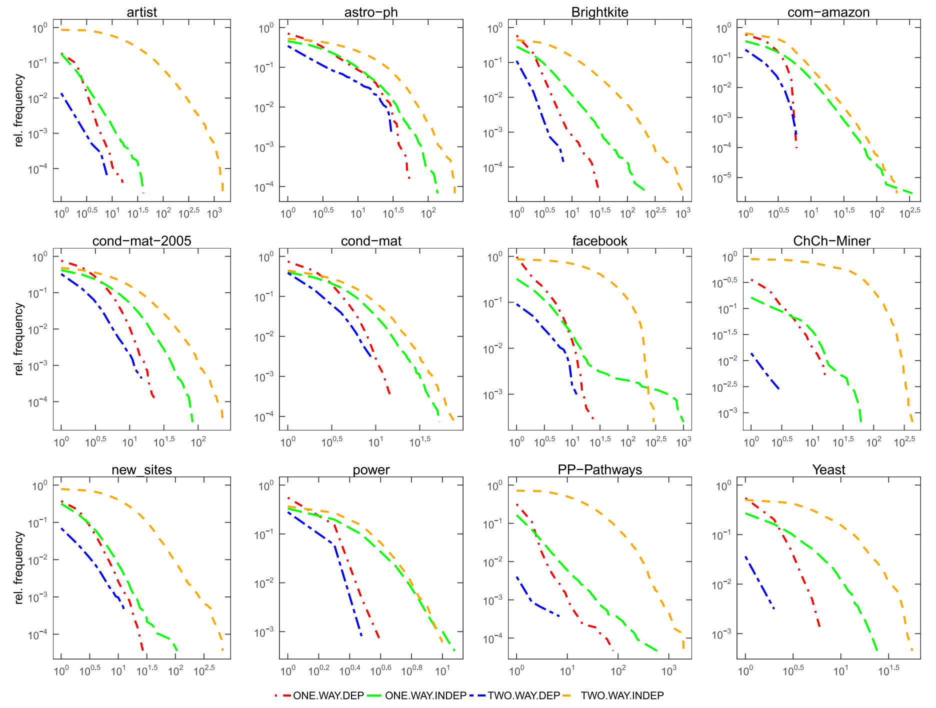

Fig. 22 Dependency properties in real-world networks

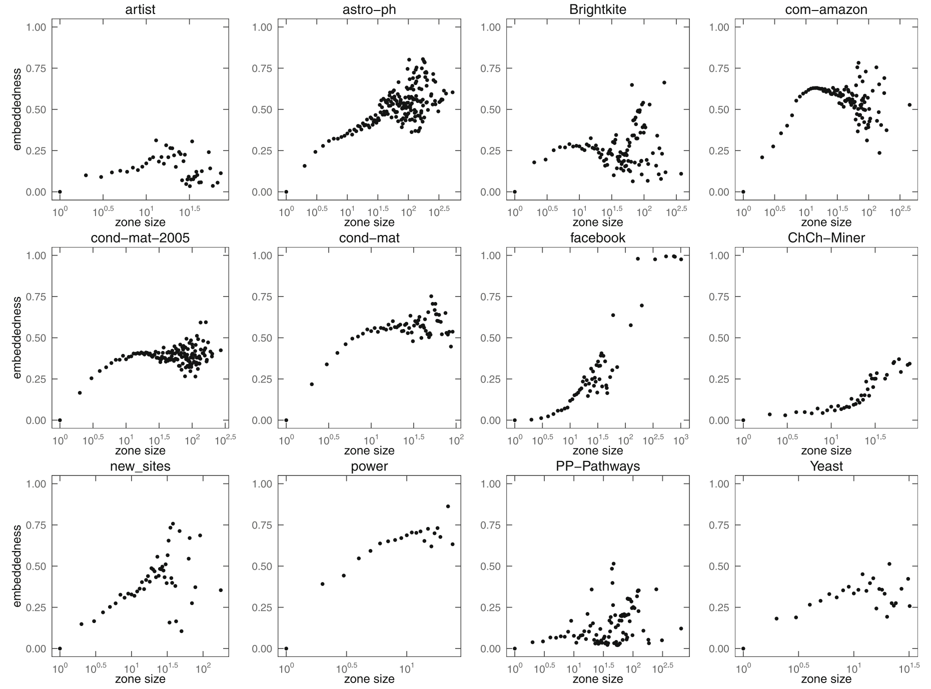

Fig. 23 Embeddedness in real-world networks

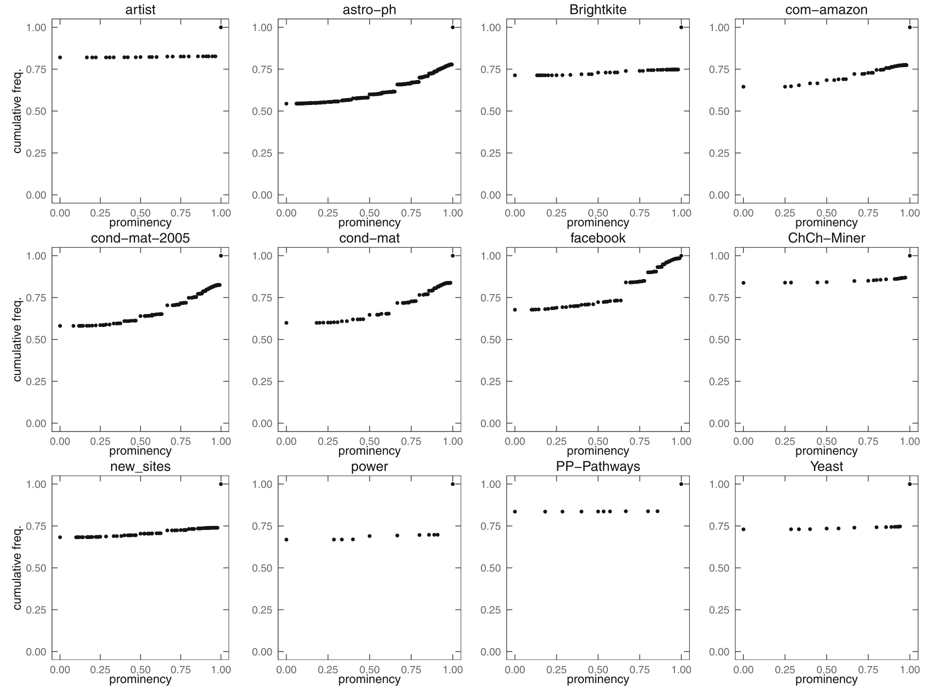

Fig. 24 Prominency in real-world networks

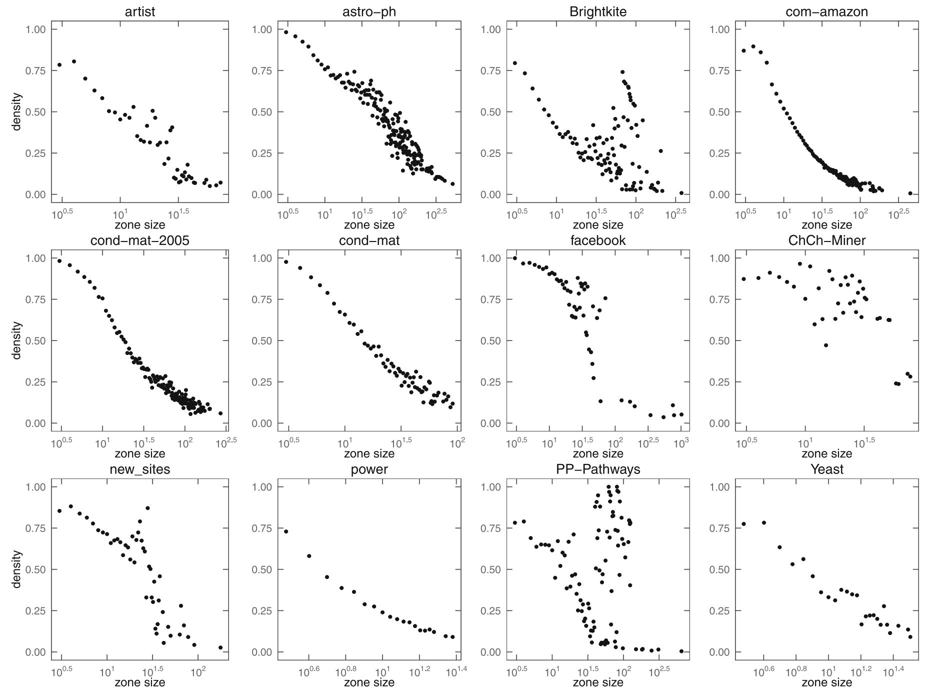

Fig. 25 Density of zones in real-world networks

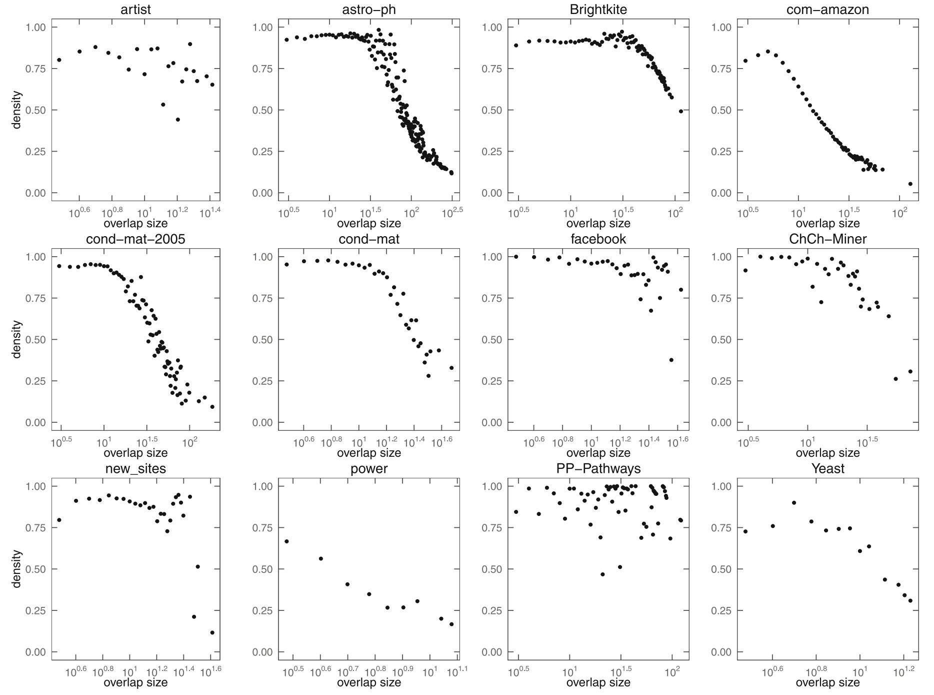

Fig. 26 Density of zone overlaps in real-world networks

Abbreviations

BA model: Barabási-Albert model; CC: clustering coefficient; ER model: Erdös-Rényi model; MCC: Matthews correlation coefficient; NetDep: network dependency; OwDep: one-way dependency; OwIndep: one-way independency; TC: triadic-closure based model; TwDep: two-way dependency; TwIndep: two-way independency

Acknowledgements

Not applicable.

Authors' contributions

We use the following notation for different types of contribution: A - algorithm design, I - implementation, P - paper writing, E - experimental evaluation. The authors contributed as follows: MK (A, I, P, E), EO (E, P), JP (I, E), SZ (P). All authors read and approved the final manuscript.

Funding

This work is partially supported by SGS, VSB-Technical University of Ostrava, under the grants no. SP2018/130 and SP2019/14 and Ministry of Health of the Czech Republic under grants no. MZ CR VES16-31852A and MZ VES16-32339A.

Availability of data and materials

The generated network data used in this study are available at https://homel.vsb.cz/~kud007/ego_zones_files/. Only publicly archived datasets (see Appendix B) or datasets generated according the known models were used during the study.

Competing interests

The authors declare that they have no competing interests.

Received: 13 February 2019 Accepted: 23 August 2019
Published online: 17 October 2019

References

Abbasi A, Chung KSK, Hossain L (2012) Egocentric analysis of co-authorship network structure, position and performance. Inf Process Manag 48(4):671–679

Agrawal M, Zitnik M, Leskovec J, et al (2018) Large-scale analysis of disease pathways in the human interactome. Pac Symp Biocomput 23:111–122. World Scientific

Ahn Y-Y, Bagrow JP, Lehmann S (2010) Link communities reveal multiscale complexity in networks. Nature 466(7307):761

Aiello LM, Deplano M, Schifanella R, Ruffo G (2012) People are strange when you're a stranger: Impact and influence of bots on social networks. In: ICWSM'12: Proceedings of the 6th AAAI International Conference on Weblogs and Social Media. AAAI

Albert R, Barabási A-L (2002) Statistical mechanics of complex networks. Rev Mod Phys 74(1):47

Bagrow JP, Bollt EM (2005) Local method for detecting communities. Phys Rev E 72(4):046108

Barnes ER (1982) An algorithm for partitioning the nodes of a graph. SIAM J Algebraic Discret Methods 3(4):541–550

Bashan A, Parshani R, Havlin S (2011) Percolation in networks composed of connectivity and dependency links. Phys Rev E 83(5):051127

Baumes J, Goldberg MK, Krishnamoorthy MS, Magdon-Ismail M, Preston N (2005) Finding communities by clustering a graph into overlapping subgraphs. IADIS AC 5:97–104

Bianconi G, Darst RK, Iacovacci J, Fortunato S (2014) Triadic closure as a basic generating mechanism of communities in complex networks. Phys Rev E 90(4):042806

Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. J Stat Mech Theory Exp 2008(10):10008

Bu D, Zhao Y, Cai L, Xue H, Zhu X, Lu H, Zhang J, Sun S, Ling L, Zhang N, et al (2003) Topological structure analysis of the protein–protein interaction network in budding yeast. Nucleic Acids Res 31(9):2443–2450

Chakraborty T, Dalmia A, Mukherjee A, Ganguly N (2017) Metrics for community analysis: A survey. ACM Comput Surv (CSUR) 50(4):54

Cho E, Myers SA, Leskovec J (2011) Friendship and mobility: user movement in location-based social networks. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '11. ACM. pp 1082–1090. https://doi.org/10.1145/2020408.2020579

Clauset A (2005) Finding local community structure in networks. Phys Rev E 72(2):026132

Danisch M, Guillaume J-L, Le Grand B (2013) Towards multi-ego-centred communities: a node similarity approach. Int J Web Based Communities 9(3):299–322

Erdös P, Rényi A (1959) On random graphs, i. Publ Math (Debrecen) 6:290–297

Evans TS, Lambiotte R (2010) Line graphs of weighted networks for overlapping communities. Eur Phys J B 77(2):265–272

Fortunato S (2010) Community detection in graphs. Phys Rep 486(3-5):75–174

Fortunato S, Hric D (2016) Community detection in networks: A user guide. Phys Rep 659:1–44

Freeman LC (1977) A set of measures of centrality based on betweenness. Sociometry 40(1):35–41. https://doi.org/10.2307/3033543

Girvan M, Newman ME (2002) Community structure in social and biological networks. Proc Natl Acad Sci 99(12):7821–7826

Gregory S (2010) Finding overlapping communities in networks by label propagation. New J Phys 12(10):103018

Guimera R, Sales-Pardo M, Amaral LAN (2004) Modularity from fluctuations in random graphs and complex networks. Phys Rev E 70(2):025101

Hric D, Darst RK, Fortunato S (2014) Community detection in networks: Structural communities versus ground truth. Phys Rev E 90(6):062805

Jacob Y, Winetraub Y, Raz G, Ben-Simon E, Okon-Singer H, Rosenberg-Katz K, Hendler T, Ben-Jacob E (2016) Dependency network analysis (d ep na) reveals context related influence of brain network nodes. Sci Rep 6:27444

Kernighan BW, Lin S (1970) An efficient heuristic procedure for partitioning graphs. Bell Syst Tech J 49(2):291–307
Khorasgani RR, Chen J, Zaïane OR (2010) Top leaders community detection approach in information networks. In: 4th SNA-KDD Workshop on Social Network Mining and Analysis. ACM
Knuth DE (1993) The Stanford GraphBase: A Platform for Combinatorial Algorithms. In: Proceedings of the Fourth Annual ACM-SIAM Symposium on Discrete Algorithms, Philadelphia. pp 41–43
Kudelka M, Zehnalova S, Horak Z, Kromer P, Snasel V (2015) Local dependency in networks. Int J Appl Math Comput Sci 25(2):281–293
Lancichinetti A, Fortunato S (2009) Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities. Phys Rev E 80(1):016118
Lancichinetti A, Fortunato S, Kertész J (2009) Detecting the overlapping and hierarchical community structure in complex networks. New J Phys 11(3):033015
Leskovec J, Lang KJ, Dasgupta A, Mahoney MW (2009) Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. Internet Math 6(1):29–123
McAuley J, Leskovec J (2014) Discovering social circles in ego networks. ACM Trans Knowl Discov Data (TKDD) 8(1):4
Mislove A, Marcon M, Gummadi KP, Druschel P, Bhattacharjee B (2007) Measurement and analysis of online social networks. In: Proceedings of the 7th ACM SIGCOMM conference on Internet measurement. ACM, New York. pp 29–42
Newman ME (2001) The structure of scientific collaboration networks. Proc Natl Acad Sci 98(2):404–409
Newman ME (2006) Finding community structure in networks using the eigenvectors of matrices. Phys Rev E 74(3):036104
Newman ME, Girvan M (2004) Finding and evaluating community structure in networks. Phys Rev E 69(2):026113
Palla G, Derényi I, Farkas I, Vicsek T (2005) Uncovering the overlapping community structure of complex networks in nature and society. Nature 435(7043):814
Parshani R, Buldyrev SV, Havlin S (2011) Critical effect of dependency groups on the function of networks. Proc Natl Acad Sci 108(3):1007–1010
Pons P, Latapy M (2005) Computing communities in large networks using random walks. In: International Symposium on Computer and Information Sciences. Springer. pp 284–293. https://doi.org/10.1007/11569596_31
Poulin V, Théberge F (2018) Ensemble clustering for graphs. In: International Conference on Complex Networks and their Applications. Springer. pp 231–243
Rosvall M, Bergstrom CT (2008) Maps of random walks on complex networks reveal community structure. Proc Natl Acad Sci U S A 105(4):1118–1123
Rozemberczki B, Davies R, Sarkar R, et al. (2018) Gemsec: Graph embedding with self clustering. arXiv preprint arXiv:1802.03997
Tatti N, Gionis A (2013) Discovering nested communities. In: Machine Learning and Knowledge Discovery in Databases. Springer, Berlin. pp 32–47
Tversky A (1977) Features of similarity. Psychol Rev 84(4):327
Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world'networks. Nature 393(6684):440
Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, Sajed T, Johnson D, Li C, Sayeeda Z, et al (2017) Drugbank 5.0: a major update to the drugbank database for 2018. Nucleic Acids Res 46(D1):1074–1082
Xie J, Kelley S, Szymanski BK (2013) Overlapping community detection in networks: The state-of-the-art and comparative study. ACM Comput Surv (csur) 45(4):43
Yang J, Leskovec J (2012) Community-affiliation graph model for overlapping network community detection. In: 2012 IEEE 12th International Conference on Data Mining. IEEE. pp 1170–1175. https://doi.org/10.1109/icdm.2012.139
Yang J, Leskovec J (2015) Defining and evaluating network communities based on ground-truth. Knowl Inf Syst 42(1):181–213
Yang Y, Klimmt B (2004) The Enron corpus: A new dataset for email classification research. In: European Conference on Machine Learning. Springer. pp 217–226
Zachary WW (1977) An information flow model for conflict and fission in small groups. J Anthropol Res 33(4):452–473
Zhang J, Cheng J, Su X, Yin X, Zhao S, Chen X (2018) Correlation Analysis of Nodes Identifies Real Communities in Networks. arXiv preprint arXiv:1804.06005
Zitnik M, Agrawal M, Leskovec J (2018) Modeling polypharmacy side effects with graph convolutional networks. Bioinformatics 34(13):i457–i466

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.