Stanford Infolab Technical Report

Overlapping Communities Explain Core-Periphery Organization of Networks

Jaewon Yang, Jure Leskovec*
Stanford University

*To whom correspondence should be addressed;
E-mail: crucis@cs.stanford.edu, jure@cs.stanford.edu

October 14, 2014

Overlapping Communities Explain Core-Periphery Organization of Networks

Jaewon Yang and Jure Leskovec

Abstract—Networks provide a powerful way to study complex systems of interacting objects. Detecting network communities—groups of objects that often correspond to functional modules—is crucial to understanding social, technological, and biological systems. Revealing communities allows for analysis of system properties that are invisible when considering only individual objects or the entire system, such as the identification of module boundaries and relationships or the classification of objects according to their functional roles. However, in networks where objects can simultaneously belong to multiple modules at once, the decomposition of a network into overlapping communities remains a challenge.

Here we present a new paradigm for uncovering the modular structure of complex networks, based on a decomposition of a network into any combination of overlapping, non-overlapping, and hierarchically organized communities. We demonstrate on a diverse set of networks comping from a wide range of domains that our approach leads to more accurate communities and improved identification of community boundaries. We also unify two fundamental organizing principles of complex networks: the modularity of communities and the commonly observed core-periphery structure. We show that dense network cores form as an intersection of many overlapping communities. We discover that communities in social, information, and foodweb networks have a single central dominant core while communities in protein-protein interaction as well as product co-purchasing networks have small overlaps and form many local cores.

Index Terms—Networks, Community detection, Ground-truth communities, Core-periphery structure.

1 Introduction

Networks provide a way to represent systems of interacting objects where nodes denote objects (people, proteins, webpages) and edges between the objects denote interactions (friendships, physical interactions, links). Nodes in networks organize into communities [1], which often correspond to groups of nodes that share a common property, role or function, such as functionally related proteins [2], social communities [3], or topically related webpages [4]. Communities in networks often overlap as nodes might belong to multiple communities at once. Identifying such overlapping communities in networks is a crucial step in studying the structure and dynamics of social, technological, and biological systems [2], [3], [4], [5]. For example, community detection allows us to gain insights into metabolic and protein-protein interactions, ecological foodwebs, social networks like Facebook, collaboration networks, information networks of interlinked documents, and even networks of co-purchased products [6], [7], [8], [9], [10], [11], [12]. In particular, communities allow for analysis of system properties that cannot be studied when considering only individual objects or the entire system, such as the identification of module boundaries and relationships and the classification of objects according to their functional roles [13], [14], [15], [16], [17].

Here we explore the community structure of a number of networks from many domains. We distinguish between *structural* and *functional* definitions of communities [18]. Communities are often *structurally* defined as sets of nodes with many connections among the members of the set and few connections to the rest of the network [1]. Communities can also be defined *functionally* based on the function or role of its members. For example, functional communities may correspond to social groups in social networks, scientific disciplines or

- *J. Yang is with the Department of Electrical Engineering, Stanford Univeristy, Stanford, CA. E-mail: jayang@cs.stanford.edu*
- *J. Leskovec is with the Department of Computer Science, Stanford Univeristy, Stanford, CA. E-mail: jure@cs.stanford.edu*

Manuscript received March 18, 2014; revised August 1, 2014.

research groups in scientific collaboration networks, and biological modules in protein-protein interaction networks. The premise of community detection is that these functional communities share some degree share some common structural signature, which allows us to extract them from the network structure.

Based on this distinction one can state that the goal of community detection is to build a bridge between network structure and function. That is, to identify communities based on the network structure with the aim that such *structurally* identified communities would correspond to *functional* communities. Thus, the aim is to use community detection to identify functional communities based on their structural connectivity patterns.

In this paper we build on this view of network community detection and identify networks where we can obtain reliable external labels of functional communities. We refer to such explicitly labeled functional communities as *ground-truth* communities [18]. We study structural properties of such ground-truth functional communities and find that they exhibit a particular structural pattern. We discover that the probability of nodes being connected increases with the number of ground-truth communities they share. Our observation means that nodes residing in overlaps of ground-truth communities are more densely connected than nodes in the non-overlapping parts of communities. Interestingly, we also find that assumptions behind many existing overlapping community detection methods lead to the opposite conclusion that the more communities a pair of nodes shares, the less likely they are to be connected [6], [7], [8], [9], [10], [11]. Thus, as a consequence many overlapping community detection methods may not be able to properly detect ground-truth communities.

Based on the above observations we develop a new overlapping community detection method *Community-Affiliation Graph Model (AGM)*, which views communities as overlapping "tiles" and the tile density corresponds to edge density [19]. Figure 1 illustrates the concept. Our methodology decomposes the network into a combination of overlapping, non-overlapping, and hierarchically organized communities. We compare AGM to a number of widely-used overlapping and non-overlapping community detection methods [6], [7], [10], [20] and show that AGM leads to more accurate functional communities. On average, AGM gives 50% relative improvement over existing methods in assigning

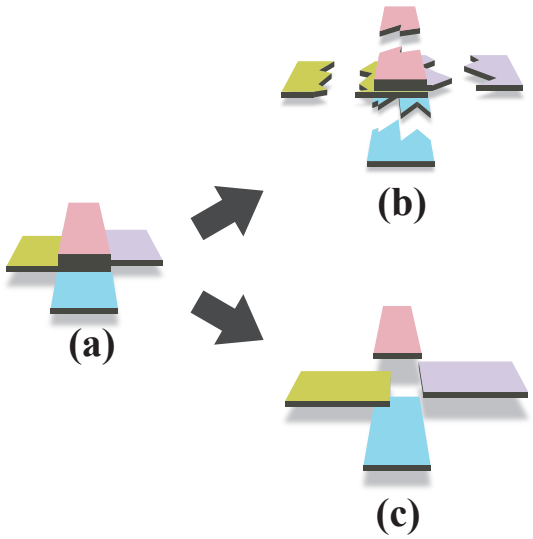
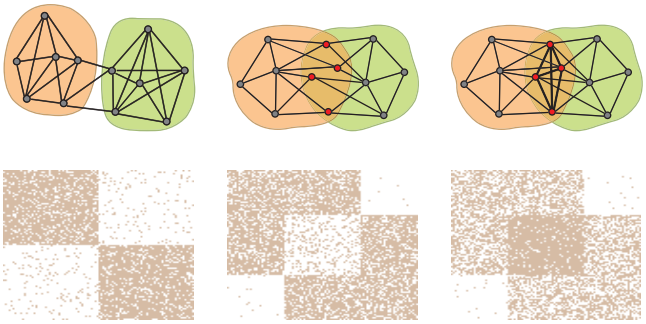


Fig. 1: **Communities as tiles. (a)** Communities in networks behave as overlapping tiles. **(b)** Many methods view communities as clusters with a homogeneous edge density and thus they may break the tiles. **(c)** Our AGM methodology successfully decomposes the network into different tiles (communities).

nodes to their ground-truth communities in social, co-authorship, product co-purchasing, and biological networks.

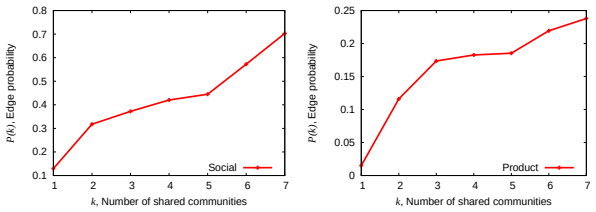
Finally, we unify two fundamental organizing principles of complex networks: overlapping communities and the commonly observed core-periphery structure. While network communities are often thought of as densely linked clusters of nodes, in core-periphery network structure, the network is composed of a densely connected core and a sparsely connected periphery [21], [22], [23]. Many large networks may exhibit core-periphery structure. The network core was traditionally viewed as a single giant community and therefore it was conjectured that the core lacks internal communities [24], [25], [26], [27]. We unify those two organizing principles and show that dense network cores form as a result of many overlapping communities. Moreover, we find that foodweb, social, and web networks exhibit a single dominant core while protein-protein interaction and product co-purchasing networks contain many local cores formed around the central core.

Our methodology to decompose networks into communities provides a powerful tool for studying social, technological, and biological systems by uncovering their modular structure. Our work represents a new way of studying networks of



(a) (b) (c)

Fig. 2: **Three structural definitions of network communities.** Networks (top) and corresponding adjacency matrices (bottom), where rows/columns denote nodes and dots denote edges: **(a)** two non-overlapping communities; **(b)** two overlapping communities where the overlap is less connected than the non-overlapping parts of communities; **(c)** two overlapping communities where the nodes in the overlap are better connected. Based on **(c)**, we structurally define communities as analogous to "tiles", where community overlaps lead to higher density of edges.



(a) Social network **(b) Product network**

Fig. 3: **Community overlaps have higher edge density than the non-overlapping parts of communities.** Edge probability $P(k)$ as a function of the number of common community memberships k in the social network **(a)** and in the product co-purchasing network **(b)** (Table 1). Results in **(a)** and **(b)** suggest that as nodes share multiple communities, they are more likely to be connected, which leads to higher edge density in community overlaps as illustrated in Figure 2c.

complex systems by bringing a shift in perspective from defining communities as densely connected nodes to conceptualizing them as overlapping tiles.

2 From Structural to Functional Definitions of Communities

The traditional structural view of network communities is based on two fundamental social network processes: *triadic closure* [28] and the *strength of weak ties* theory [29], [30]. Under this view, structural communities are often defined as corresponding to sets of nodes with many "strong" connections between the members of the community and few "weak" connections with the rest of the network (Figure 2a). However, in many domains nodes may belong to multiple communities at once, and thus the notion of structural communities has also been extended to include overlapping, hierarchical, and disassortative community structures [6], [31], [32], [33], [34].

Despite great progress in the field, we find that extending the traditional structural view to overlapping communities leads to an unnoticed consequence that nodes in community overlaps are less densely connected than nodes in the non-overlapping parts of communities (Figure 2b). (Refer to the extended version of the paper [35] for details.) We find this hidden consequence to be present in many existing approaches to overlapping community detection [6], [7], [8], [9], [10], [11].

We examine a diverse set of six networks drawn from a wide range of domains including social, collaboration, and co-purchasing networks for which we obtain explicitly labeled functional communities, which we refer to as the *ground-truth* communities [18]. For example, in social networks we take ground-truth communities to be social interest-based groups to which people explicitly join. In product networks, ground-truth communities correspond to product categories [35]. Note that ground-truth communities are not defined based on some observed node attribute or property (like, user's age or user's homework in a case of a social network). The idea behind ground-truth communities is that they would correspond to true functional modules in complex networks. While the obtained ground-truth labels may sometimes be noisy or incomplete, consistency and robustness of the results suggests that the ground-truth labels are overall reliable.[1]

By studying the structure of ground-truth communities we find that two nodes are more likely to be connected if they have multiple ground-truth communities in common (Figure 3). For example, in the LiveJournal online social network (Table 1), the edge probability jumps from $\sim 10^{-6}$ for nodes that share no ground-truth communities to 0.1 for nodes that have one ground-truth community in common and keeps increasing all the way to 0.7 as nodes

1. Networks with ground-truth communities can be downloaded from http://snap.stanford.edu/agm.

share more communities (Figure 3a). This implies that the area of overlap between two communities has a higher average density of edges than an area that falls in just a single community (Figure 2c).

Our observation is intuitive and consistent across several domains. For example, proteins belonging to multiple common functional modules are more likely to interact [2], people who share multiple interests have a higher chance of becoming friends [36], and researchers with many common interests are more likely to collaborate [36].

2.1 Defining Structural Communities as Tiles

We think of communities as analogous to overlapping "tiles". Thus, just as the overlap of two tiles leads to a higher tile height in the overlapping area, the overlap of two communities leads to higher density of edges in the overlap. (Figure 1 illustrates the concept.) The composition of many overlapping communities then gives rise to the global structure of the network.

Conceptually, our methodology represents a shift in perspective from structurally modeling communities as sets of densely linked nodes to modeling communities as overlapping tiles where the network emerges as a result of the overlap of many communities. Our structural definition of communities departs from the *strength of weak ties* theory [30] and is consistent with the earlier *web of group affiliations* social network theory [37], which postulates that edges arise due to shared community affiliations.

Our findings here also have implications for the understanding of homophily, which is one of the primary forces that shape the formation of social networks [36]. Homophily is the tendency of individuals to connect to others with similar tastes and preferences. Based on [30], it has been commonly assumed that homophily operates in "pockets" and thus nodes that have neighbors in other communities are less likely to share the attributes of those neighbors (as in Figures 2a, 2b). In contrast, our results are implying *pluralistic homophily* where the similarity of nodes is proportional to the number of shared memberships/functions, not just their similarity along a single dimension. In a multidimensional network, the most central nodes are those that have the most shared dimensions.

3 Decomposition of Networks into Communities

In order to model communities in a network we define a Community-Affiliation Graph Model (AGM) [19]. In our model, edges of the underlying network arise due to shared community memberships (Figure 4a) [38], [39]. The AGM parameterizes each community A with a single parameter p_A. Two nodes that belong to community A then form an edge in the underlying network with probability p_A. Each community A generates edges between its members independently; however, if two nodes have already been connected, then the duplicate edge is not included in the network.

The AGM naturally models communities with dense overlaps (Figures 4a, 4b). Pairs of nodes that belong to multiple common communities become connected in the underlying network with a higher probability, since for each shared community the nodes are given an independent chance of forming an edge.

The flexible nature of the AGM allows for modeling a wide range of network community structures, such as non-overlapping, hierarchically nested, and overlapping communities (Figures 4c, 4e, 4d). Given a bipartite community affiliation graph and a probability p_A for each community A, the AGM allows us to generate synthetic networks with realistic community structures, a procedure useful in and of itself.

Using the AGM, we can also identify and analyze community structure of real-world networks. We accomplish decomposition of a given network into communities by fitting the AGM to the network with tools of statistical inference. We combine a maximum-likelihood approach with convex optimization and a Monte Carlo sampling algorithm on the space of community affiliation graphs [19], [35], [40]. This technique allows us to efficiently search for the community affiliation graph that gives the observed network the greatest likelihood. To automatically determine the number of communities in a given network, we apply techniques from statistical regularization and sparse model estimation [35].

4 Accuracy of Detected Communities

Next, we aim to infer functional communities based only on the structure of a given unlabeled undirected network.

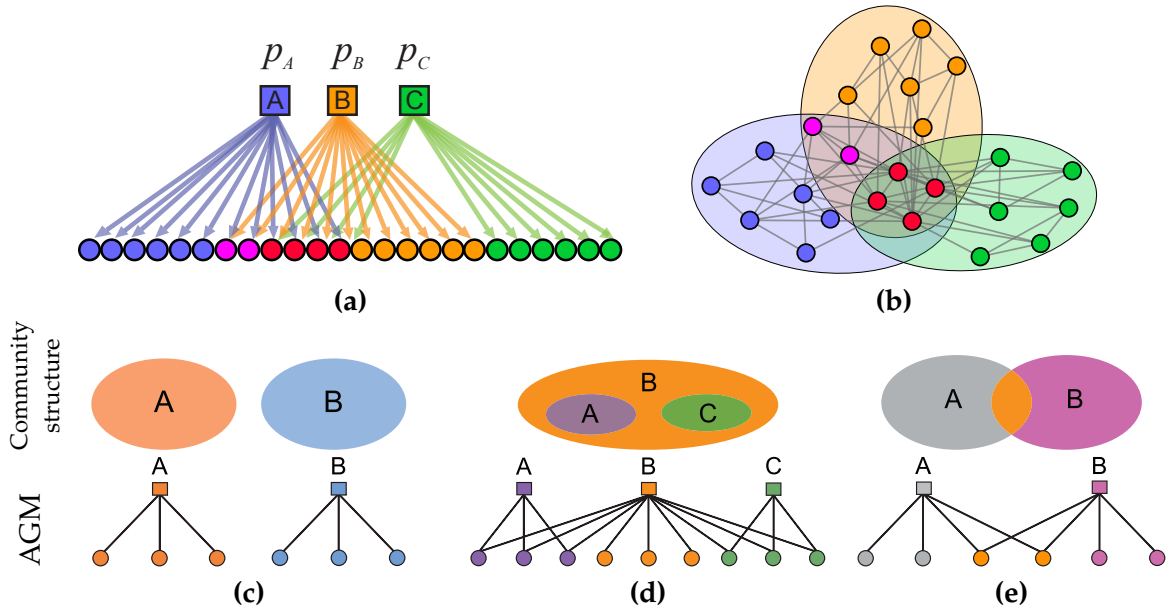


Fig. 4: **Community-Affiliation Graph Model (AGM) [19]. (a)** Squares represent communities and circles represent the nodes of a network. Edges represent node community memberships. For each community A that two nodes share they create a link independently with probability p_A. The probability that a pair of nodes u, v creates a link is thus $p(u,v) = 1-\prod_{A \in C_{uv}}(1-p_A)$, where C_{uv} is the set of communities that u and v share. If u and v do not share any communities, we assume they link with a small probability ε. **(b)** Network generated by the Community-Affiliation Graph Model in **(a)**. As pairs of nodes that share multiple communities get multiple chances to create edges, the AGM naturally generates networks where nodes in the community overlaps are more densely connected than the nodes in non-overlapping regions. **(c–e)** AGM is capable of modeling any combination of **(c)** non-overlapping, **(d)** hierarchically nested, as well as **(e)** overlapping communities.

4.1 Qualitative Evaluation

As an illustrative example, we consider a Facebook friendship network of a single user's friends (Figure 5a and Table 1). In order to obtain labels for ground-truth communities, we asked the user to manually organize his Facebook friends into communities. The user classified his friends into four communities corresponding to his high-school, workplace, and two communities of university friends. The visualization of the same network using communities in Figure 5b shows that the network in Figure 5a is in fact composed of the overlaps of the four communities. In this example, the goal of community detection is to identify the communities in Figure 5b based only on the connectivity structure of the network in Figure 5a.

Due to an implicit assumption that nodes in community overlaps are less densely connected than nodes in the non-overlapping parts of communities (Figure 2b), many overlapping community detection approaches [6], [7], [8], [9], [10], [11] fail to properly detect communities in this network. For example, Figures 5c, 5d, and 5e illustrate the result of applying Clique Percolation [10], Link Clustering [6], and Mixed-Membership Stochastic Block Model [7] to the Facebook network in Figure 5a. We also give a formal argument that explains the behavior of these methods in the Appendix A.1 and the extended version [35].

When we use the AGM to analyze the Facebook network, the AGM automatically detects four communities (Figure 6), which is the same as the number identified by the user. Moreover, the communities detected by the AGM nearly perfectly correspond to communities identified by the user. The AGM correctly determines community overlaps and community memberships for 94% of the user's friends.

4.2 Quantitative Evaluation

We also perform a large-scale quantitative evaluation on AGM on biological, social, collaboration, and product networks where functional communities are explicitly labeled [18]. The networks represent a wide range of sizes and edge densities, as well as amounts of community overlap. We compare the AGM to a number of widely used overlapping and non-overlapping community de-

Network	Properties of networks					Properties of detected communities		
	N	E	$\langle C \rangle$	D	$\langle k \rangle$	K	$\langle S \rangle$	$\langle A \rangle$
Facebook	183	2,873	0.56	2.80	31.40	4	70.8	1.5
Social network	3,997,962	34,681,189	0.28	6.47	17.35	29,774	83.3	0.6
Foodweb	128	2,075	0.33	1.90	32.42	5	54.4	2.1
Web graph	255,265	1,941,926	0.62	9.36	15.21	5,000	83.3	1.6
PPI network	1,213	2,556	0.33	10.50	4.21	40	31.6	1.0
Product network	334,863	925,872	0.40	15.00	5.53	9,020	50.0	1.3

TABLE 1: **Network statistics and properties of detected communities.** We consider the Facebook ego-network of a particular user, the full LiveJournal online social network, the Florida bay foodweb network, the Stanford University web graph, the literature-curated *Saccharomyces cerevisiae* protein-protein interaction (PPI) network, and the Amazon product co-purchasing network. Network statistics: N: Number of nodes, E: Number of edges, $\langle C \rangle$: Average clustering coefficient, D: Effective diameter, $\langle k \rangle$: Average node degree. Properties of detected communities: K: Number of communities, $\langle S \rangle$: Average detected community size, $\langle A \rangle$: Average number of community memberships per node. The networks vary from those with modular to highly overlapping community structure and represent a wide range of edge densities. While the number of communities detected by AGM varies, the average community size is quite stable across the networks. Average number of community memberships per node reveals that communities in the foodweb overlap most pervasively, while in PPI and social networks overlaps are smallest.

tection methods [6], [7], [10], [20] and quantify the correspondence between the explicitly labeled ground-truth communities and the communities detected by a given method. The performance metrics quantify the accuracy of the method in assigning nodes to their ground-truth communities. (Refer to Appendix A.2 for further details.)

On a set of social, collaboration, and product networsk AGM on average outperforms existing methods by 50% in four different metrics that quantify the accuracy in assigning nodes to their ground-truth communities (Figure 11a). In particular, AGM gives a 50% relative improvement over Clique Percolation [10]. Link Clustering [6] detects overlapping as well as hierarchical communities and AGM improves 61% over it. Similar levels of improvement are achieved when comparing AGM to other overlapping and non-overlapping methods [7], [20]. Furthermore, AGM gives a 14% relative improvement over Link Clustering using the same networks and same data-driven benchmarks as used in the Link Clustering work [6].

Furthermore, we also experiment with AGM on a set of four different biological protein-protein interaction networks. Remarkably, even though AGM was developed based on insights gained on primarily social networks, we find that AGM performs surprisingly well on biological networks as well. As performance metrics, we compute the average statistical significance of detected communities (p-value) for the three types of Gene Ontology (GO) [41]. We consider negative logarithm of average p-values for each of the three GO term types as three separate scores. On average, the AGM outperforms Link Clustering by 150%, CPM by 163%, Infomap by 148%, and MMSB for 12 times (Figure 11b). Further experimental details are in Appendix and [35].

Overall, the AGM approach yields substantially more accurate communities. The success of our approach relies on the AGM's flexible nature, which allows the AGM to decompose a given network into a combination of overlapping, non-overlapping, and hierarchical communities.

5 Communities, Pluralistic Homophily, and Core-Periphery Structure

The AGM also makes it possible to gain well-founded insights into the community structure of networks. In particular, we discover that overlapping communities lead to a global core-periphery network structure. Core-periphery structure captures the notion that many networks decompose into a densely connected core and a sparsely connected periphery [21], [22]. The core-periphery structure is a pervasive and crucial characteristic of large networks [23], [24], [42].

We discover that a network core forms as a result of *pluralistic homophily* where the connectedness of nodes is proportional to the number of shared community memberships, and not just their similarity along a single dimension or community. Thus, the network core forms as a result of many overlapping communities. The average number of community memberships of a node decreases with

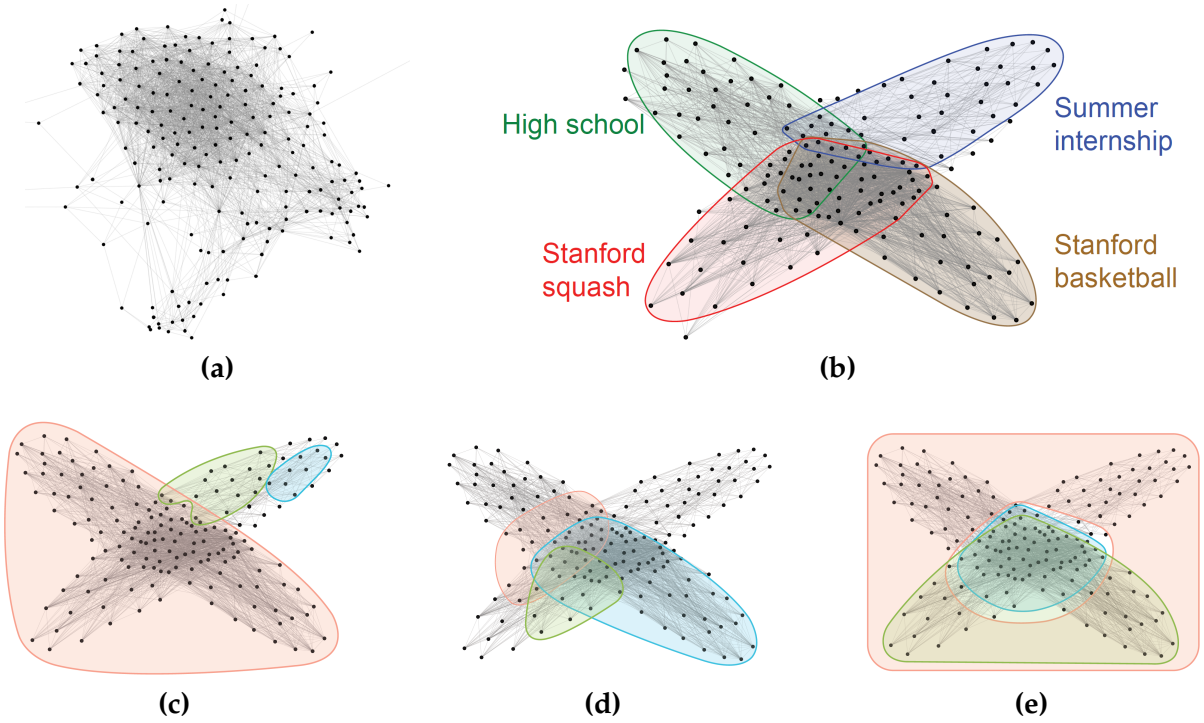


Fig. 5: **An example on a Facebook friendship network of a particular user. (a)** Facebook friendship network of a single user. **(b)** The same network but with communities explicitly labeled by the user: high-school friends, colleagues at the workplace, and university friends with whom the user plays basketball and squash. Communities are denoted by filled regions. Notice that nodes in the overlap of communities have higher density of edges. **(c-e)** Results of applying **(c)** Clique Percolation, **(d)** Link Clustering, and **(e)** Mixed-Membership Stochastic Block Model to the Facebook network.

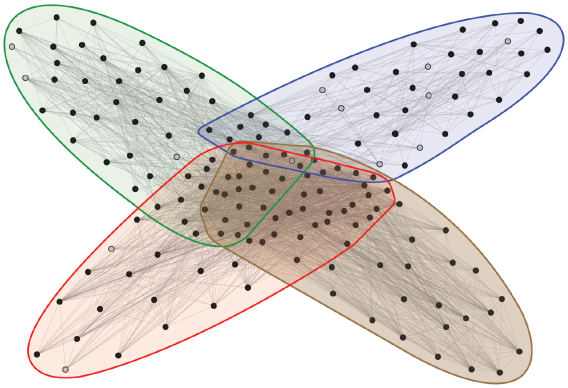
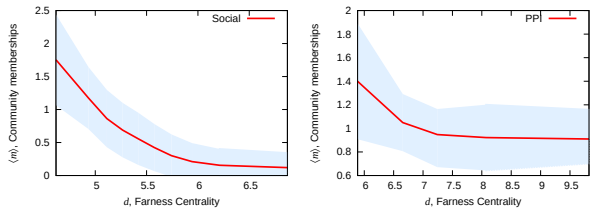


Fig. 6: **AGM on the Facebook network from Figure 5.** AGM successfully decomposes the network into different tiles (communities) and correctly determines community overlaps as well as community memberships for 94% of the nodes.

its distance from the center of the network (Figure 7). Moreover, the edge likelihood increases as a function of community memberships (Figure 3). Thus, the nodes in the center of the network have higher density of edges than nodes in the periphery. Therefore, we show that even in the presence of many communities, pluralistic homophily leads to dense community overlaps, which cause a global core-periphery network structure.

A further examination of the amount of community overlap reveals that social, web, and foodweb networks in Table 1 have a single central dominant core (Figure 8a). On the other hand, communities in protein and product networks have small overlaps and also form many local cores (Figure 8b). In particular, protein communities only slightly overlap and form local cores as well as a small global core (Figure 8d). Small overlaps of protein communities can be explained by the fact that communities act as functional modules, and it would be hard for the cell to independently control heavily overlapping modules [2], [6]. Communities of co-purchased products can also be thought of as functional modules since the products in a community are bought together for a specific purpose. On the other hand, foodweb communities overlap pervasively while forming a single dominant core. This leads to a flower-like overlapping community structure (Fig-



(a) Social network **(b) PPI network**

Fig. 7: **Overlapping communities lead to global core-periphery network structure.** The average (and the 10^{th} percentiles) of the number of community memberships $\langle m \rangle(d)$ as a function of its farness centrality d, defined as the average shortest path length of a given node to all other nodes of the network [3]. **(a)** LiveJournal social network, **(b)** *Saccharomyces cerevisiae* PPI network. Number of community memberships of a node decreases with its farness centrality. Nodes that reside in the center of the network (and have small shortest path distances to other nodes of the network) belong to the highest number of communities. This means that core-periphery structure forms due to community overlaps. Communities in the periphery tend to be non-overlapping while communities in the core overlap pervasively.

ure 8c), where tiles (communities) overlap to form a single core of the network. The heavily overlapping foodweb communities form due to the closed nature of the studied Florida bay ecosystem [43]. Web communities overlap moderately and form a single global core. Many of these communities form around common interests or topics, which may overlap with each other [4].

6 Conclusion

In closing, we note that our approach builds on the previous work on community detection [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16]. We examine an implicit assumption of sparsely connected community overlaps and find that regions of the network where communities overlap have higher density of edges than the non-overlapping regions.

We then rethink classical structural definitions of communities and develop the AGM, which models structural communities as overlapping tiles. Using our well-founded approach we find that all networks considered in this study exhibit a core-periphery structure where nodes that belong to multiple communities reside in the core of the network. However, networks have different kinds of core-periphery structure depending on the mechanism for community formation in the networks. Dense community overlaps also explain the mixed success of present community detection methods when applied to large networks [24], [27].

Our work also enhances our understanding of homophily as one of the most fundamental social forces. Homophily in networks has been traditionally thought to operate in small pockets/clusters. Thus, nodes that have neighbors in other communities were considered less likely to share properties of those neighbors. In contrast, our results are implying pluralistic homophily where the similarity of nodes' properties is proportional to the number of shared community memberships. In a network, the most central nodes are those that have the most shared properties/functions/communities with others. More generally, our work provides a shift in perspective from conceptualizing communities as densely connected sets of nodes to defining them as overlapping tiles and represents a new way of studying complex systems.

Acknowledgments. We thank R. Sosič, P. Mason, M. Macy, S. Fortunato, D. McFarland, and H. Garcia-Molina for invaluable discussions and feedback. Supported by NSF Career Award IIS-1149837, DARPA XDATA and GRAPHS, Alfred P. Sloan Fellowship, and the Microsoft Faculty Fellowship.

Appendix A

A.1 Detecting Densely Overlapping Communities

We next show that three popular community detection methods Clique Percolation [10], [44]; Link Clustering [6]; and Stochastic Block Model [7], [45] cannot properly detect communities with dense overlaps.

A.1.1 Clique Percolation

First, we analyze the Clique Percolation method and show that it may not properly detect two overlapping communities from Figure 2c. Clique Percolation Method (CPM) has a single input parameter k which determines the size of the maximal cliques that the algorithm looks for. For example, Figure 9 shows the result of CPM on the network of Figure 2c where the overlap between the two communities is denser than the individual communities. When $k = 3$, CPM finds a community that covers the whole network because the clique in the overlap connects the cliques in the left community and the right community, whereas CPM finds a community of the overlap when $k = 4$.

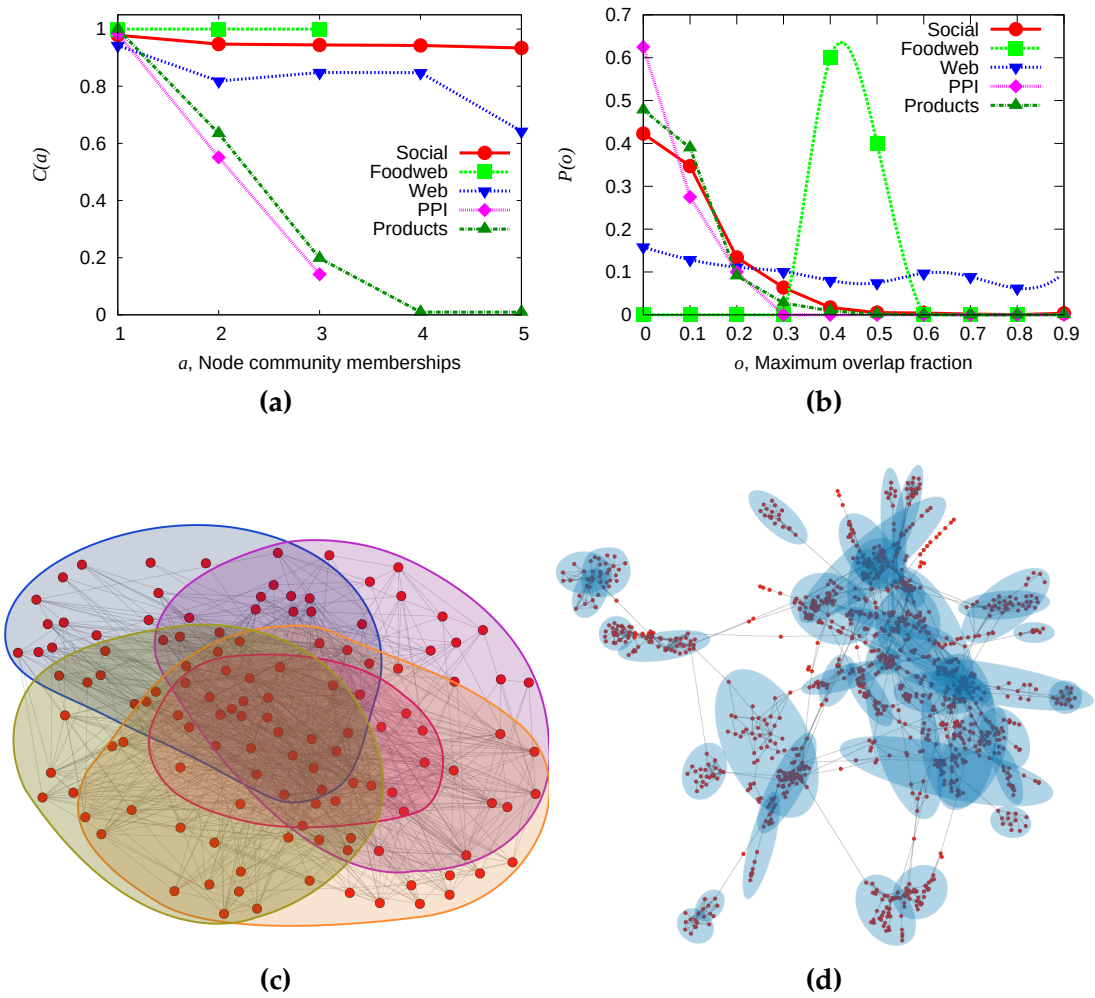
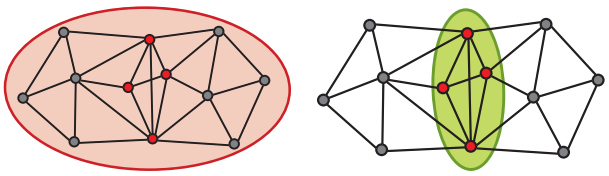


Fig. 8: **Primary and secondary cores in networks. (a)** The fraction of nodes $C(a)$ in the largest connected component of the induced subgraph on the nodes who belong to at least a communities. By thinking of a network as a valley where peaks correspond to cores and peripheries to lowlands, our methodology is analogous to flooding lowlands and measuring the fraction of the largest island. A high $C(a)$ means that there is a single dominant core (peak), while a low $C(a)$ suggests the existence of nontrivial secondary cores. **(b)** Probability density $P(o)$ of the maximum overlap o. Maximum overlap o_A of a given community A is defined as the fraction of A's nodes that are in the overlap with any other community. Communities in the PPI, social, and product co-purchasing networks are mostly non-overlapping whereas the communities in the foodweb and the web graph are pervasively overlapping. **(c)** Communities detected by the AGM in the foodweb form a single central core. **(d)** Communities in the PPI network form many secondary cores.

In addition to Clique Percolation Method, there are many other overlapping community detection methods that are based on expanding the maximal cliques. These methods (for example, Greedy clique expansion [46] and EAGLE [47]) also suffer from the same problem.

A.1.2 Stochastic Block Models

We show that three variants of stochastic block models are unable to correctly discover communities with dense overlaps: the traditional Stochastic Block Model [45], the Degree-Corrected Stochastic Block Model [48] and the Mixed-Membership Stochastic Block Model [7]. Based on the input matrix from Figure 2c, all three models identify three blocks as illustrated in Figure 10. The reason for this is that the edge probability between two nodes that belong to communities A and B is weighted average of $P(A, A)$ and $P(B, B)$, where $P(X, Y)$ is an edge probability between a node in community X and a node in community Y. This means that the edge probability between the two



(a) $k_{CPM} = 3$ **(b)** $k_{CPM} = 4$

Fig. 9: **Clique Percolation method cannot detect communities with dense overlaps.** Given a network with two communities and a dense overlap, Clique Percolation method would report a community that (depending on the parameter settings) either **(a)** includes both communities, or **(b)** it would find a small community consisting only of the overlap.

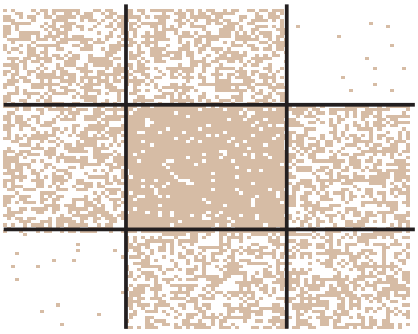


Fig. 10: **The result of a Stochastic Block Model and the Mixed-Membership Stochastic Block model on a network of two communities with dense overlap.** The adjacency matrix of the network in Figure 2c is shown and the bold lines denote the three partitions discovered by the stochastic block models, where the overlap is confused as a separate community.

nodes that share multiple communities is smaller than the maximum of $P(A, A)$ and $P(B, B)$ (due to the weighted summation). Therefore, the edge probability between overlapping nodes cannot be higher than the edge probability between nodes in an individual community. We also note that in principle one could apply post-processing of communities detected by stochastic block models to identify which of the detected structural communities actually correspond to overlaps of functional communities. However, it is not immediately clear how to develop such post-processing method.

A.1.3 Link Clustering

Lastly, we show that the Link Clustering [6] is not able to correctly detect overlapping communities with dense overlaps. Link Clustering performs hierarchical clustering with the edges of the given network based on the Jaccard similarity between the adjacent nodes of the edges. Since edge density in the area of community overlap is higher, this means that the Jaccard similarity between the adjacent nodes will be higher, which in turn means that Link Clustering will identify the edges in the overlap as a separate community. (Refer to the extended version [35] for details.)

A.2 Metrics of Community Detection Accuracy

We focus the evaluation of community detection methods on their ability to correctly identify overlapping communities.

To quantify the performance, we measure the level of agreement between the detected and the ground-truth communities. Given a network $G(V, E)$, we consider a set of ground truth communities C^* and a set of detected communities $\hat{C}$, where each ground-truth community $C_i \in C^*$ and each detected community $\hat{C}_i \in \hat{C}$ is defined by a set of its member nodes. To compare $\hat{C}$ and C^*, we use four performance metrics:

Average F1 score [49]: We compute $F_g(C_i) = \max_j F1(C_i, \hat{C}_j)$ for each ground-truth community C_i and $F_d(\hat{C}_i) = \max_j F1(C_j, \hat{C}_i)$ for each detected community $\hat{C}_i$, where $F1(S_1, S_2)$ is the harmonic mean of precision and recall between two node sets S_1, S_2. The average F1 score is $\frac{1}{2}(\bar{F}_g + \bar{F}_d)$ where $\bar{F}_g = \frac{1}{|C^*|}\sum_i F_g(C_i)$ and $\bar{F}_d = \frac{1}{|\hat{C}|}\sum_i F_d(\hat{C}_i)$.

Omega Index [50]: For each pair of nodes $u, v \in V$, we define C_{uv} to be the set of ground-truth communities to which both u and v belong and $\hat{C_{uv}}$ to be the set of detected communities to which the both nodes belong. Then the Omega Index is $\frac{1}{|V|^2}\sum_{u,v \in V} \mathbf{1}\{|C_{uv}| = |\hat{C_{uv}}|\}$.

Normalized Mutual Information [12]: We compute $1 - \frac{1}{2}(H(C^*|\hat{C}) + H(\hat{C}|C^*))$, where $H(A|B)$ is the extension of entropy when A, B are sets of sets [12].

Accuracy in the number of communities: $1 - \frac{||C^*| - |\hat{C}||}{|C^*|}$, which is the relative error in predicting the number of communities.

A.3 Applying AGM to Social, Product, and Collaboration Networks

Figure 11a displays the composite performance of each of the 5 methods over the six networks with ground-truth communities. Overall, we notice that AGM gives best overall performance on all networks except the Amazon, where it ties with MMSB. Furthermore, AGM detects highest quality communities for most individual performance

metrics in all networks. On average, the composite performance of AGM is 3.40, which is 61% higher than that of Link Clustering (2.10), 50% higher than that of CPM (2.41), 30% higher than that of Infomap and 8% higher than that of MMSB (3.25). The absolute average value of Omega Index of AGM over the 6 networks is 0.46, which is 21% higher than Link Clustering (0.38), 22% higher than CPM (0.37), 5% higher than Infomap (0.44) and 26% higher than MMSB (0.36).

In terms of absolute values of scores, AGM archives the average F1 score of 0.57, average Omega index of 0.46, Mutual Information of 0.15 and accuracy of the number of communities 0.42. We also note that AGM also outperforms CPM with other values of k ($k = 3, 4, 6$).

A.4 Applying AGM to Biological Networks

We also evaluate the performance of AGM on the four types of protein-protein interaction (PPI) networks of *Saccharomyces cerevisiae* [6]. As performance metrics, we compute the average statistical significance of detected communities (p-value) for the three types of Gene Ontology (GO) terms (biological process, cellular component and molecular function) [41]. We consider negative logarithm of average p-values for each of the three GO term types as three separate scores.

Figure 11b displays the composite performance in the four PPI networks. We observe that the AGM attains the best composite performance in all four networks. On average, the composite performance of AGM is 3.00, which is 150% higher than that of Link Clustering (1.20), 163% higher than that of CPM (1.14), 148% higher than that of Infomap (1.21) and 12 times higher than that of MMSB (0.08). We further investigated the poor performance of MMSB on these networks and found it is due to the fact that MMSB tends to find very large communities consisting of more than 80% of the nodes.

REFERENCES

[1] S. Fortunato, "Community detection in graphs," *Physics Reports*, vol. 486, no. 3-5, pp. 75 – 174, 2010.

[2] N. Krogan, G. Cagney, H. Yu, et al., "Global landscape of protein complexes in the yeast saccharomyces cerevisiae," *Nature*, vol. 440, no. 7084, pp. 637–643, 2006.

[3] S. Wasserman and K. Faust, *Social Network Analysis*. Cambridge University Press, 1994.

[4] G. Flake, S. Lawrence, C. Giles, and F. Coetzee, "Self-organization and identification of web communities," *Computer*, vol. 35, no. 3, pp. 66–71, 2002.

[5] M. Newman, *Networks: An Introduction*. Oxford University Press, Inc., 2010.

[6] Y.-Y. Ahn, J. P. Bagrow, and S. Lehmann, "Link communities reveal multi-scale complexity in networks," *Nature*, vol. 466, pp. 761–764, Oct. 2010.

[7] E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing, "Mixed membership stochastic blockmodels," *Journal of Machine Learning Research*, vol. 9, pp. 1981–2014, 2007.

[8] M. Sales-Pardo, R. Guimerà, A. Moreira, and L. A. N. Amaral, "Extracting the hierarchical organization of complex systems," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, pp. 18 874–18 874, 2007.

[9] I. Psorakis, S. Roberts, M. Ebden, and B. Sheldon, "Overlapping community detection using bayesian non-negative matrix factorization," *Physical Review E*, vol. 83, p. 066114, 2011.

[10] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," *Nature*, vol. 435, no. 7043, pp. 814–818, 2005.

[11] T. S. Evans and R. Lambiotte, "Line graphs, link partitions, and overlapping communities," *Physical Review E*, vol. 80, p. 016105, 2009.

[12] A. Lancichinetti and S. Fortunato, "Community detection algorithms: A comparative analysis," *Physical Review E*, vol. 80, no. 5, p. 056117, 2009.

[13] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, p. P10008, 2008.

[14] M. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Physical Review E*, vol. 69, p. 026113, 2004.

[15] P. J. Mucha, T. Richardson, K. Macon, M. A. Porter, and J.-P. Onnela, "Community structure in time-dependent, multiscale, and multiplex networks," *Science*, vol. 328, no. 5980, pp. 876–878, 2010.

[16] C. Granell, S. Gómez, and A. Arenas, "Hierarchical multiresolution method to overcome the resolution limit in complex networks," *International Journal of Bifurcation and Chaos*, vol. 22, no. 7, 2012.

[17] B. Ball, B. Karrer, and M. E. J. Newman, "Efficient and principled method for detecting communities in networks," *Physical Review E*, vol. 84, p. 036103, 2011.

[18] J. Yang and J. Leskovec, "Defining and evaluating network communities based on ground-truth communities," in *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, 2012, pp. 745–754.

[19] ——, "Community-affiliation graph model for overlapping community detection," in *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, 2012.

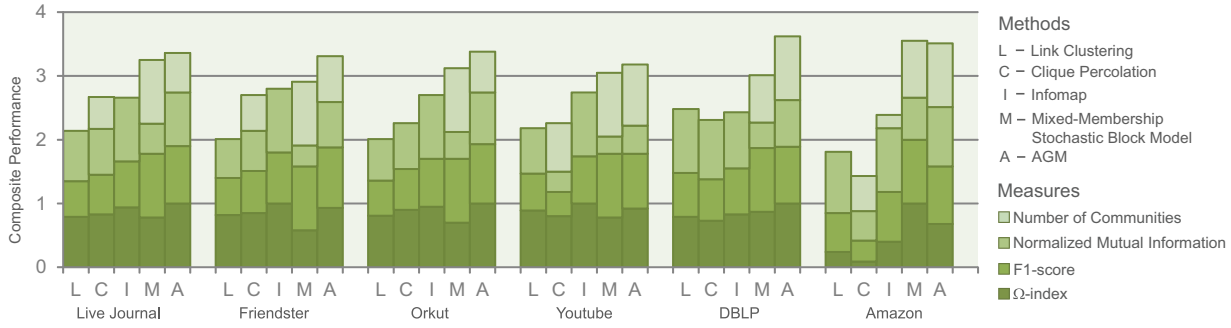
[20] M. Rosvall and C. T. Bergstrom, "Maps of random walks on complex networks reveal community structure," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, pp. 1118–1123, 2008.

[21] S. P. Borgatti and M. G. Everett, "Models of core/periphery structures," *Social Networks*, vol. 21, pp. 375 – 395, 1999.

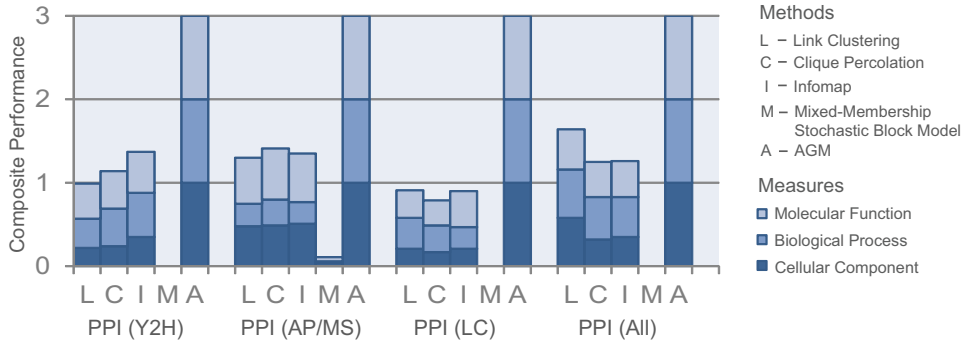
[22] P. Holme, "Core-periphery organization of complex networks," *Physical Review E*, vol. 72, p. 046111, 2005.

[23] F. D. Rossa, F. Dercole, and C. Piccardi, "Profiling core-periphery network structure by random walkers," *Scientific Reports*, vol. 3, 2013.

[24] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney, "Community structure in large networks: Natural cluster



(a) Social, collaboration, and product networks



(b) Biological networks

Fig. 11: The composite performance of the community detection methods on: **(a)** six networks with externally labeled ground-truth communities and **(b)** four biological networks.

sizes and the absence of large well-defined clusters," *Internet Mathematics*, vol. 6, no. 1, pp. 29–123, 2009.

[25] A. Clauset, M. Newman, and C. Moore, "Finding community structure in very large networks," *Physical Review E*, vol. 70, p. 066111, 2004.

[26] M. Coscia, G. Rossetti, F. Giannotti, and D. Pedreschi, "Demon: a local-first discovery method for overlapping communities," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2012, pp. 615–623.

[27] J. Leskovec, K. Lang, and M. Mahoney, "Empirical comparison of algorithms for network community detection," in *Proceedings of the International Conference on World Wide Web (WWW)*, 2010.

[28] D. Watts and S. Strogatz, "Collective dynamics of small-world networks," *Nature*, vol. 393, pp. 440–442, 1998.

[29] J. A. Davis, "Clustering and Structural Balance in Graphs," *Human Relations*, vol. 20, pp. 181–187, 1967.

[30] M. S. Granovetter, "The strength of weak ties," *American Journal of Sociology*, vol. 78, pp. 1360–1380, 1973.

[31] A. Clauset, C. Moore, and M. Newman, "Hierarchical structure and the prediction of missing links in networks," *Nature*, vol. 453, no. 7191, pp. 98–101, 2008.

[32] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi, "Defining and identifying communities in networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 9, pp. 2658–2663, 2004.

[33] J. Yang and J. Leskovec, "Community detection in networks with node attributes," in *ICDM '13: Proceedings of the IEEE International Conference on Data Mining*, 2013.

[34] J. J. Yang, McAuley and J. Leskovec, "Detecting cohesive and 2-mode communities in directed and undirected networks," in *WSDM '14: Proceedings of the ACM International Conference on Web Search and Data Minig*, 2014.

[35] J. Yang and J. Leskovec, "Structure and overlaps of communities in networks." Stanford InfoLab, Technical Report, October 2014. The data and code that were used for the experiments are available at `http://snap.stanford.edu/agm`.

[36] M. McPherson, L. Smith-Lovin, and J. M. Cook, "Birds of a feather: Homophily in social networks," *Annual Review of Sociology*, vol. 27, pp. 415–444, 2001.

[37] G. Simmel, *Conflict: the Web of Group Affiliations. Trans. by Kurt H. Wolff and Reinhard Bendix*. Free Press, 1955.

[38] R. L. Breiger, "The duality of persons and groups," *Social Forces*, vol. 53, no. 2, pp. 181–190, 1974.

[39] S. Lattanzi and D. Sivakumar, "Affiliation networks," in *Proceedings of the 41st annual ACM Symposium on Theory of Computing*, 2009, pp. 427–434.

[40] J. Yang and J. Leskovec, "Overlapping community detection at scale: A non-negative factorization approach," in *WSDM '13: Proceedings of the ACM International Conference on Web Search and Data Minig*, 2013.

[41] E. Boyle, S. Weng, J. Gollub, H. Jin, D. Botstein, J. Cherry, and G. Sherlock, "GO::TermFinder - open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes," *Bioinformatics*, vol. 20, no. 18, pp. 3710–3715, 2004.

[42] M. P. Rombach, M. A. Porter, J. H. Fowler, and P. J. Mucha, "Core-periphery structure in networks," *SIAM Journal of Applied Mathematics*, vol. 74, no. 1, pp. 167–190, 2014.

[43] R. E. Ulanowicz, C. Bondavalli, and M. S. Egnotovich,

"Network analysis of trophic dynamics in south florida ecosystem, FY 97: The florida bay ecosystem," *Annual Report to the United States Geological Service Biological Resources Division*, pp. 98–123, 1998.

[44] S. Lehmann, M. Schwartz, and L. K. Hansen, "Biclique communities," *Phys. Rev. E*, vol. 78, p. 016108, 2008.

[45] P. W. Holland, K. B. Laskey, and S. Leinhardt, "Stochastic blockmodels: First steps," *Social Networks*, vol. 5, no. 2, pp. 109–137, 1983.

[46] C. Lee, F. Reid, A. McDaid, and N. Hurley, "Detecting highly overlapping community structure by greedy clique expansion," in *Proceedings of the Fourth international workshop on Advances in social network mining and analysis*, 2010.

[47] H. Shen, X. Cheng, K. Cai, and M.-B. Hu, "Detect overlapping and hierarchical community structure in networks," *Physica A: Statistical Mechanics and its Applications*, vol. 388, no. 8, pp. 1706 – 1712, 2009.

[48] B. Karrer and M. Newman, "Stochastic blockmodels and community structure in networks." *Physical Review E*, vol. 83, p. 016107, 2010.

[49] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[50] S. Gregory, "Fuzzy overlapping communities in networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2011, no. 02, p. P02017, 2011.

Stanford Infolab Technical Report

Overlapping Communities Explain Core-Periphery Organization of Networks

Jaewon Yang, Jure Leskovec*
Stanford University

*To whom correspondence should be addressed;
E-mail: crucis@cs.stanford.edu, jure@cs.stanford.edu

October 14, 2014

Contents

S1 Data description: Networks with ground-truth communities

One of the challenges of the research on network community detection is in evaluation of obtained network communities. The main challenge stems from the fact that labeled ground-truth communities are require significant effort to obtain. Thus, it is very hard to quantify on a large scale the performance of a given community detection method. In this sense identifying a diverse set of networks where ground-truth communities are explicitly labeled can have significant impact on the field.

We identified networks where nodes explicitly state their ground-truth community memberships. We first describe the source of labels for ground-truth communities and then argue why they correspond to "real" underlying communities.

We consider a set of 6 large social, collaboration and information networks, where for each network we identify a graph and a set of explicitly labeled ground-truth communities. We identify networks where nodes explicitly state their ground-truth community memberships. We did our best to identify networks in which such ground-truth communities can be reliably defined and identified based on *functional roles* of the nodes. In particular, we define ground-truth communities based on common affiliations, social circles, roles, activities, interests, functions, or some other properties around which networks organize into communities. Network sizes range from hundreds of thousand to hundreds of millions of nodes and billions of edges. Even though our networks come from a diverse set of domains and the labels of individual ground-truth communities may include some noise, the results are surprisingly robust and consistent.

Social networks. First we consider four very different online social networks (the LiveJournal blogging community [5], the Friendster online network [42], the Orkut social network [42], and the Youtube social network [42]) where users create explicit *groups* which other users then join. Such groups serve as organizing principles of nodes in social networks and are focused on specific topics, interests, hobbies, affiliations, and geographical regions. Groups range from small to very large and are created based on specific topics, interests, hobbies and geographical regions. For example, LiveJournal categorizes groups into the following types: culture, entertainment, expression, fandom, life/style, life/support, gaming, sports, student life and technology. Overall, there are over three hundred thousand explicitly defined groups in LiveJournal. Similarly, users in Friendster as well as in Orkut and Youtube define topic-based groups that others then join. The Friendster and Orkut networks have more than a million explicitly defined groups and each user can join to one or more groups. We consider each such explicitly created group as a ground-truth community.

The LiveJournal network was provided to us by Lars Backstrom [5], the Friendster net-

	Network statistics					Ground-truth communities		
Network	N	E	$\langle C \rangle$	$\langle D \rangle$	$\langle k \rangle$	K	S	A
LiveJournal [5]	4,036,538	34,916,684	0.36	6.57	17.30	311,782	40.06	3.09
Friendster [42]	117,751,379	2,586,147,869	0.21	5.98	43.93	1,449,666	26.72	0.33
Orkut [42]	3,072,441	117,185,083	0.17	5.28	76.28	8,455,253	34.86	95.93
Youtube [42]	1,138,873	2,990,443	0.17	6.28	5.25	30,087	9.75	0.26
DBLP [5]	425,957	1,348,244	0.61	6.57	6.33	2,547	429.79	2.57
Amazon [37]	334,863	925,872	0.43	12.98	5.53	49,732	99.86	14.83

Table S1: **Networks with ground-truth communities.** N: Number of nodes, E: Number of edges, $\langle C \rangle$: Average clustering coefficient [61], $\langle D \rangle$: Average shortest path length, $\langle k \rangle$: Average node degree. Properties of ground-truth communities: K: Number of communities, S: Average community size, A: Community memberships per node. Additional networks used in this study are described in Table S2. All our networks are complete and publicly available at `http://snap.stanford.edu/agm`.

work was made public by the Internet Archive[1], and the Orkut and Youtube networks were kindly provided to us by Alan Mislove [42].

Amazon product co-purchasing network. The second type of network data we consider is the Amazon product co-purchasing network [37]. The nodes of the network represent products and edges link commonly co-purchased products. Each product (*i.e.*, node) belongs to one or more hierarchically organized product categories and products from the same category define a group which we view as a ground-truth community. This means members of the same community share a common function or role, and each level of the product hierarchy defines a set of hierarchically nested and overlapping communities. We crawled this network using the Amazon API [37].

DBLP collaboration network. We consider the collaboration networks of DBLP [5], where nodes represent authors and edges connect nodes that have co-authored a paper. We use publication venues as ground-truth communities which serve as proxies for highly overlapping scientific communities around which the network then organizes. This network was provided to us by Lars Backstrom.

Ground-truth network characteristics. Table S1 gives the dataset statistics. Observe that the size of the networks ranges between hundreds of thousands to hundreds of millions of nodes and billions of edges. The number of ground-truth communities varies from hundreds to millions and there is also a range in ground-truth community sizes and node membership distribution.

All our networks are complete and are publicly available at `http://snap.stanford.edu/agm`. For each of these networks we identified a sensible way of defining ground-truth communities that serve as organizational units of these networks.

Even though our networks come from very different domains and individual labels may be noisy or even incomplete, the results we present here are robust and consistent across all the datasets. Our work is consistent with the premise that is implicit in all network commu-

[1] `http://www.archive.org/details/friendster-dataset-201107`

nity literature: members of "real" communities share some (latent/unobserved) property or affiliation that serves as an organizing principle of the nodes and makes them well-connected in the network. Here we use groups around which communities organize to explicitly define ground-truth.

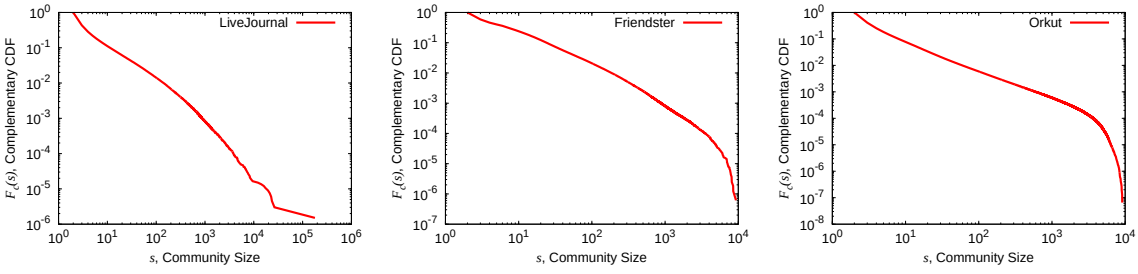
Data preprocessing. To represent all networks in a consistent way we drop edge directions and consider each network as an unweighted undirected static graph. Because members of the group may be disconnected in the network, we consider each connected component of the group as a separate ground-truth community. However, we allow ground-truth communities to be nested and to overlap (*i.e.*, a node can be a member of multiple groups at once).

Community size and membership size distribution. Next we present the distribution of the various properties of ground-truth communities. Our goal here is to investigate properties of ground-truth communities and demonstrate that such sets of nodes in fact correspond to "real" network communities.

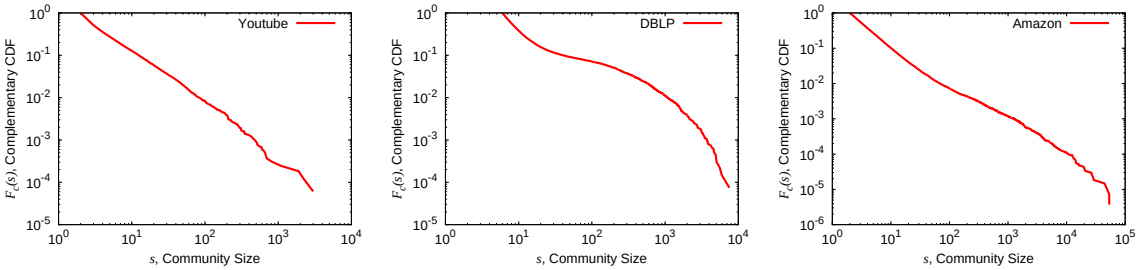
Previous literature found that the size of communities, *i.e.*, the number of the nodes in communities, follows a heavy-tailed distribution [1, 47, 65]. Figure S1 shows the CCDF (complementary cumulative distribution function) of the sizes of ground-truth communities in the 6 networks. The distribution appears to follow a heavy-tailed distribution, which for LiveJournal, YouTube and Amazon appears to be power-law.

Figure S2 shows that the CCDF of the distribution of the number of communities a node is member of. We observe it exhibits a power-law decay, but the distributions do not show a long tail in some data sets such as Orkut, DBLP, and Amazon. This is in accordance with Palla et al. [47] who reported that the distribution of node memberships, *i.e.*, the number of the communities that a node belongs to, tends to follow a power-law.

Last, we also examine the statistics of the community overlaps. We focus on overlaps between a pair of ground-truth communities and report the absolute and fractional size of the overlap between two communities. Figure S3 shows the distribution of the absolute overlap sizes. We observe that the distributions follow a power-law, as also observed by Palla et al. [47] on detected (not ground-truth) communities. In addition, we investigate how ground-truth communities overlap: Do ground-truth communities overlap in a nested structure? Or, do they overlap only for a small fraction of members? To do this, we measure the fraction f of the size of the overlap $A \cap B$ between two communities A, B to the size of the smaller community, $min(|A|, |B|)$ ($f = |A \cap B|/min(|A|, |B|)$). f being close to 1 means a nested structure where the larger community includes the smaller one, and small f means overlap in the fringe. Figure S4 plots the distribution of the overlap fraction f. The Amazon network shows high probability at $f = 1$ because the ground-truth communities form a nested structure by construction. In social networks and the DBLP network, most overlaps take a small fraction of individual communities, which is reasonable as each community has its own special interests.

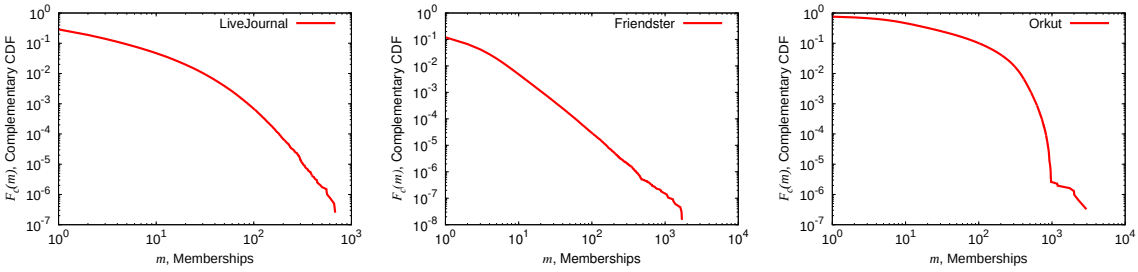


(a) LiveJournal **(b) Friendster** **(c) Orkut**

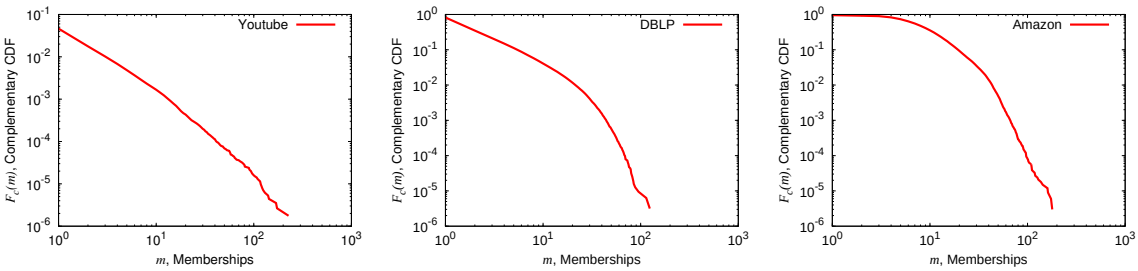


(d) Youtube **(e) DBLP** **(f) Amazon**

Figure S1: **Ground-truth community size distribution.** Complementary cumulative distribution function $F_c(s)$ of the size of ground-truth communities, s. The size of a ground-truth community denotes the number of nodes belonging to the ground-truth community.

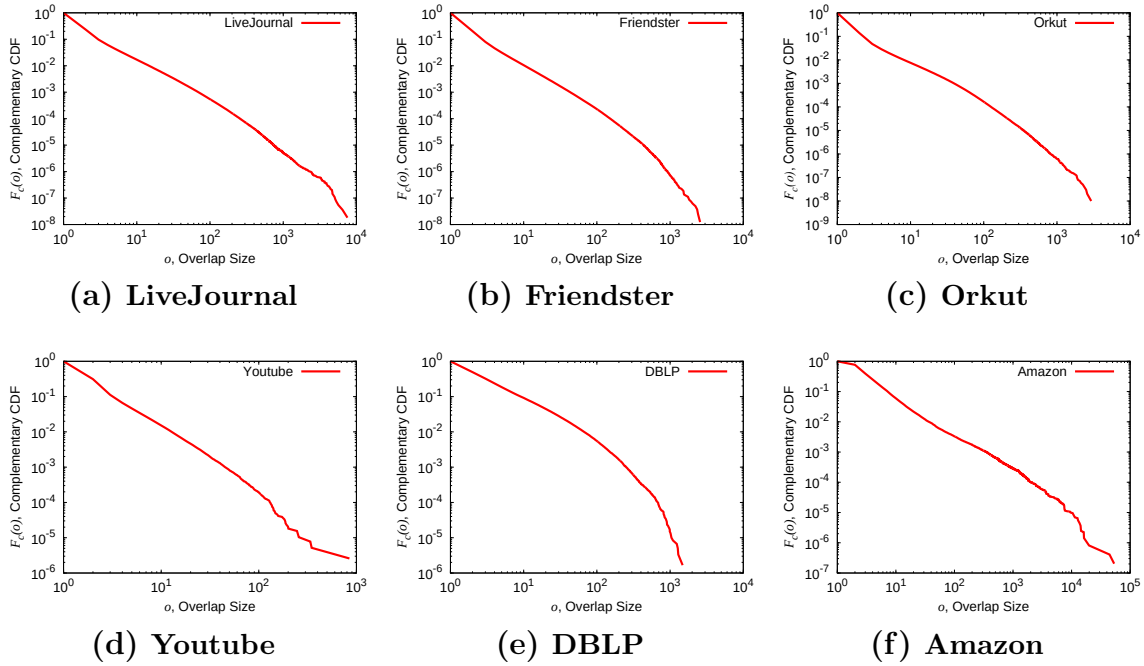


(a) LiveJournal **(b) Friendster** **(c) Orkut**



(d) Youtube **(e) DBLP** **(f) Amazon**

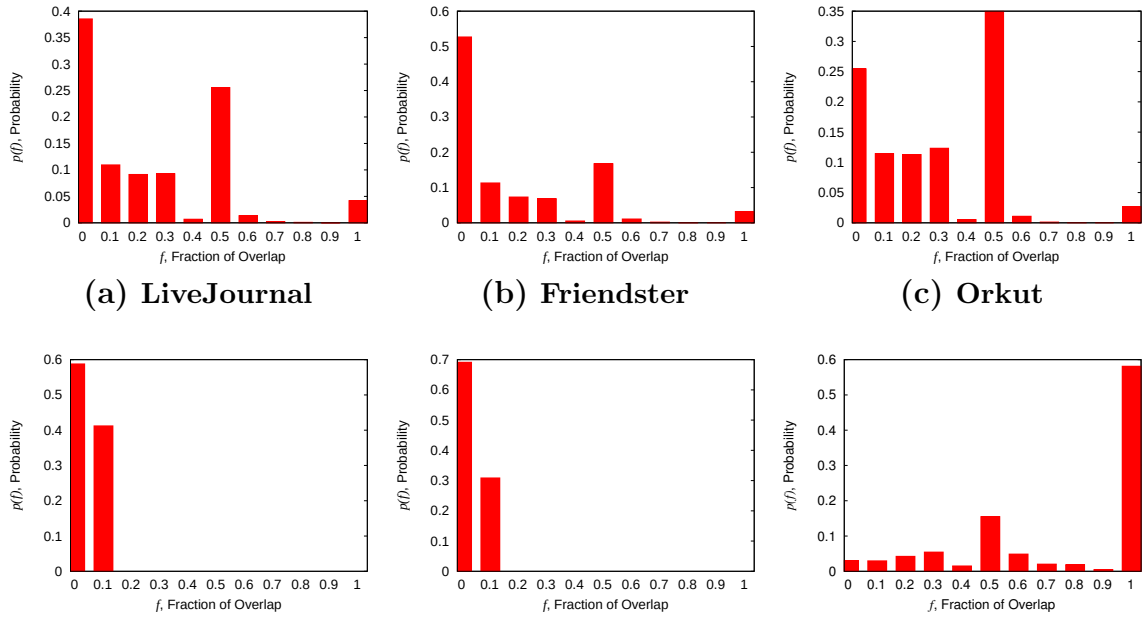
Figure S2: **Node membership distribution.** Complementary cumulative distribution function $F_c(m)$ of the number of communities m nodes belong to.



(a) LiveJournal **(b) Friendster** **(c) Orkut**

(d) Youtube **(e) DBLP** **(f) Amazon**

Figure S3: **Community overlap distribution.** Complementary cumulative distribution function $F_c(o)$ of the size of overlaps between pairs of ground-truth communities, o. The size of an overlap is the number of the nodes that belong to the overlap.



(a) LiveJournal **(b) Friendster** **(c) Orkut**

(d) Youtube **(e) DBLP** **(f) Amazon**

Figure S4: **Relative community overlap.** Histogram (probability) $p(f)$ of the fraction of the relative overlap size f. When ground-truth communities A, B overlap $A \cap B$ then $f = |A \cap B|/min(|A|, |B|)$, where $min(x, y)$ is the smaller of x and y.

S2 Empirical observation: Community overlaps have higher density of edges

The availability of reliable ground-truth communities allows us to empirically study their structure and enhance our understanding of how nodes organize themselves into communities. We analyze how nodes of ground-truth communities connect to each other, how they connect to the rest of the network, and how they overlap. This way we can empirically study on a large scale how real communities map on the underlying network structure. Based on empirical findings presented here, we will later develop a novel model and a new method for detecting overlapping communities in networks.

Empirical observation. Communities in networks form overlaps in a way that nodes belong to multiple communities simultaneously. As demonstrated in previous section ground-truth communities overlap and many nodes belong to multiple communities at once (Figures S2 and S3).

We study the structure of community overlap simply by asking what is the probability that a pair of nodes is connected if they share membership to k common ground-truth communities. Figure S5 plots this probability for all six datasets for which we have the ground-truth community data.

We notice that all curves are steeply increasing. This means that, the more communities a pair of nodes has in common, the higher the probability of them being connected. Notice the effect of shared memberships on the edge probability is very strong. For example, in LiveJournal, if a pair of nodes has 4 communities in common, the probability of friendship is nearly 50%. To appreciate how strong the effect of shared communities is on edge probability, one has to note that all of our networks are extremely sparse. The background probability of a random pair of nodes being connected is $\approx 10^{-5}$, while as soon as a pair of nodes shares two communities, their probability of linking increases by 4 orders of magnitude (from 10^{-5} to 10^{-1}).

We note that all other data sets exhibit similar behavior — the probability of a pair of nodes being connected approaches 1 as the number of common communities increases. While in online social networks the edge probability exhibits a diminishing-returns-like growth, in DBLP, it appears to follow a threshold-like behavior.

In retrospective, the above result is very intuitive: People sharing multiple interests have a higher chance of becoming friends [41], researchers with many common interests are more likely to work together [49], and proteins belonging to multiple common functional modules are more likely to interact [22, 32].

Implications for community detection. Our finding in Figure S5 suggests communities overlap as illustrated in Figure S6a. In particular, we can think of communities as being analogous to tiles, where tile overlaps lead to higher thickness of tiles, that is, community overlaps lead to higher density of edges (Figure S6a illustrates the concept).

Even though this notion of network communities is very intuitive, it is also very different from present literature that mostly defines communities as clusters of densely connected nodes. In particular, the predominant view of network communities today is based on two fundamental social network processes: triadic closure [61] and "strength of weak ties" [25]. This leads to the picture of network communities as illustrated in Figure S6b. Applying

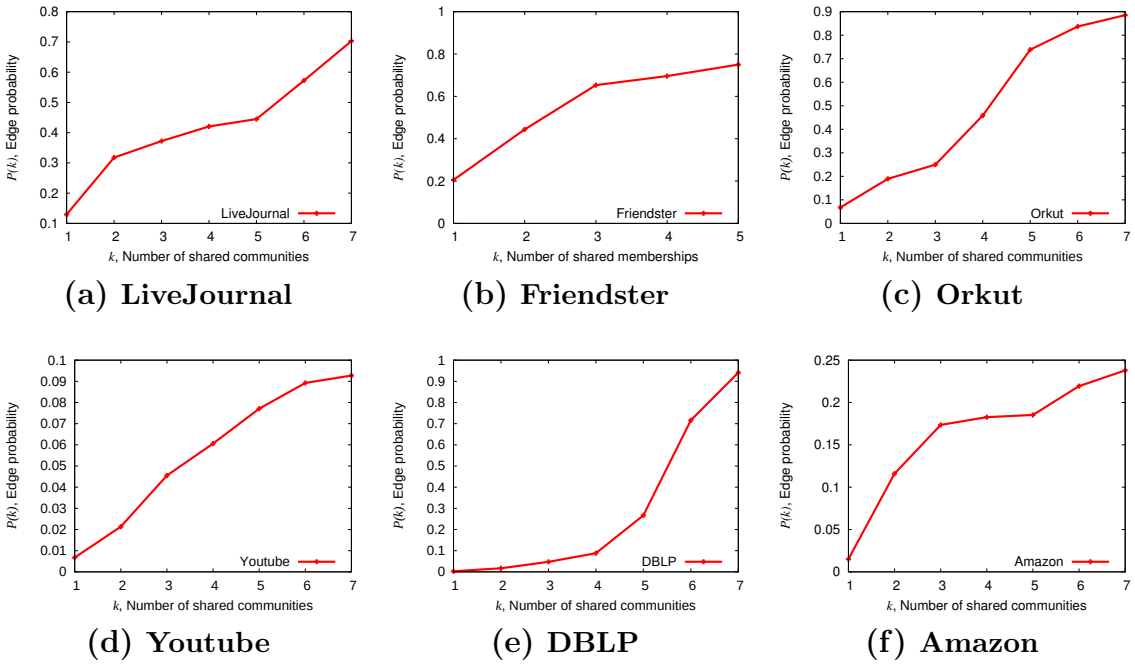


Figure S5: **Community overlaps are more densely connected than the non-overlapping parts of communities.** Edge probability $P(k)$ between two nodes given that two nodes share k communities. We observe that $P(k)$ is an increasing function of k in all the networks. For the purpose of this plot, we use the 5,000 ground-truth communities that are most cohesive in each data set [62].

this view to the case of overlapping communities leads to the (arguably unnatural) structure of community overlaps as illustrated in Figure S6c. On the other hand, our novel view of network communities as overlapping tiles is consistent with works of Simmel [57] on web of affiliations, and Feld [19] on focused organization of social ties. In both of these views networks consist of overlapping "tiles" or "social circles" that serve as organizing principles of nodes in networks.

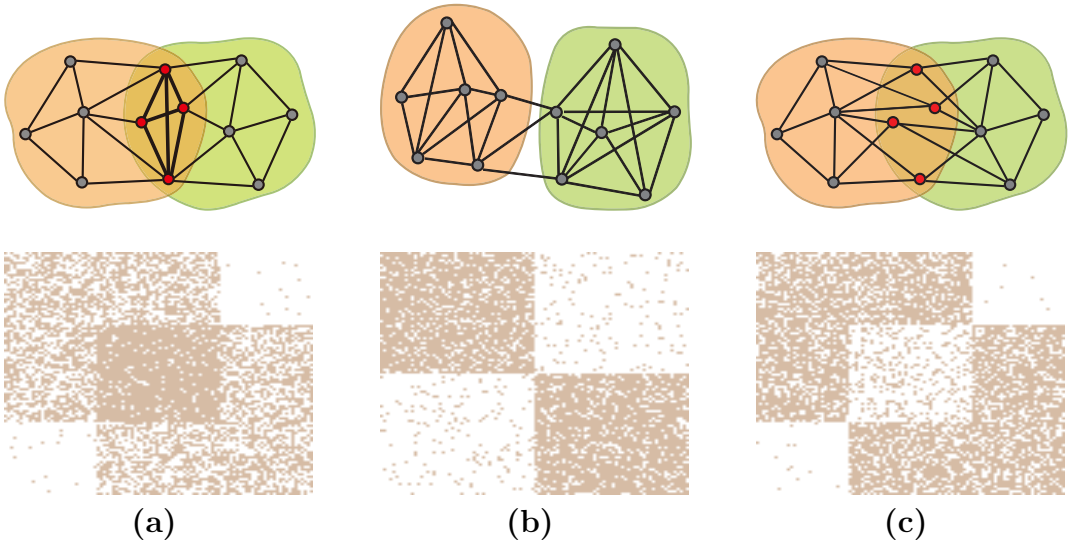
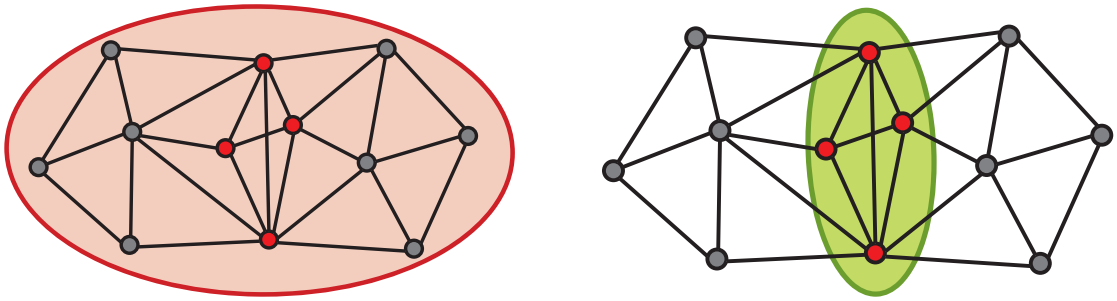


Figure S6: **Three different definitions of network communities.** Three networks (top) and corresponding adjacency matrices (bottom). We find **(a)** that as nodes share multiple communities, they are more likely to link which leads to densely connected community overlaps. However, most existing community detection methods either assume that **(b)** communities do not overlap or that **(c)** community overlaps are less well-connected than the non-overlapping parts of communities. Moreover, most existing community detection methods cannot properly detect communities with overlaps as in **(a)**.



(a) $k_{CPM} = 3$ **(b)** $k_{CPM} = 4$

Figure S7: **Clique percolation method cannot detect communities with dense overlaps.** Given a network with two communities and a dense overlap, Clique percolation method would report a community that (depending on the parameter settings) either include both communities (**(a)**), or it would find a small community consisting only of the overlap (**(b)**).

S3 Consequences for present community detection approaches

Having established that communities overlap as illustrated in Figure S6a we next show that present state-of-the-art community detection methods *fail* to properly detect communities with such overlaps. In particular, we show that three state-of-the-art overlapping community detection methods Clique percolation [47, 36], Link clustering [1], and Stochastic block model [2, 28] all fail to properly detect communities with dense overlaps (Figure S6a).

S3.1 Clique percolation

First we analyze the Clique percolation method and show that it fails to properly detect two overlapping communities as illustrated in Figure S6a.

Clique percolation method (CPM) has a single input parameter k which determines the size of the maximal cliques that the algorithm looks for. After finding all the k-cliques on the given network, the method merges two k-cliques if they share $k-1$ nodes. Overlaps can happen when the nodes in the overlaps belong to multiple k cliques that cannot be merged. When an overlap is denser, however, nodes in the overlap form many k-cliques themselves and the k cliques in the overlap would be likely to merge together. In this case, the method would either identify the overlap as a separate community, or merge adjacent communities through k-cliques in the overlap.

For example, Figure S7 shows the result of CPM on the network of Figure S6a where the overlap between the two communities is denser than the individual communities. When $k = 3$, CPM finds a community that covers the whole network because the clique in the overlap connects the cliques in the left community and the right community, whereas CPM finds a community of the overlap when $k = 4$.

In addition to Clique percolation method, there are many other overlapping community detection methods that are based on expanding the maximal cliques. These methods (for example, Greedy clique expansion [35] and EAGLE [56]) also suffer from the same problem.

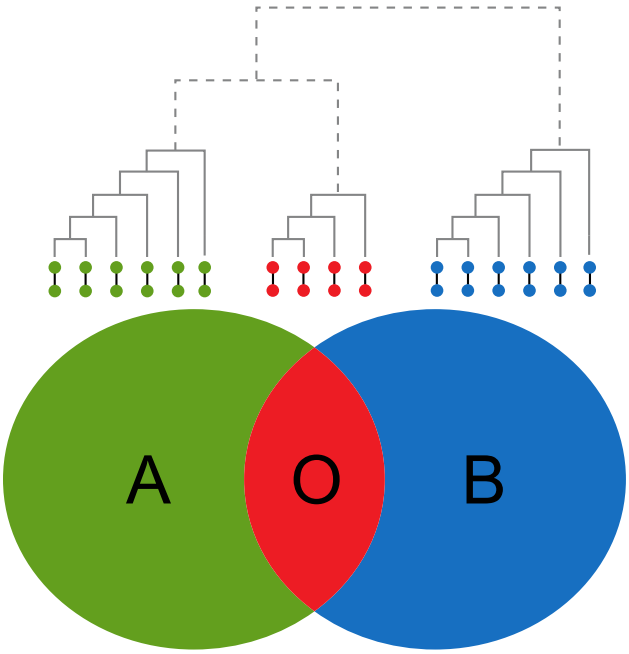


Figure S8: **Link clustering cannot detect communities with dense overlaps.** A network with two overlapping communities and the outcome of Link clustering. Link clustering builds a dendrogram with solid lines which finds the overlap as a separate community. The merger of the overlap and the single community regions (the right and the left communities), described by the dotted lines in the dendrogram, cannot happen as Link clustering will stop the algorithm because of the decrease in the partition density.

By the same reasoning that we used for Clique percolation, we can see that none of these clique expansion methods is able to discover densely overlapping communities. For example, in the networks in Figure S7, neither EAGLE nor Greedy clique expansion could correctly identify the red overlapping nodes. As all the red nodes form a maximal clique, both methods will regard the red nodes as a single community, and there is no way to tell that the red nodes belong to more than one community.

S3.2 Link clustering

Next we show that the Link clustering [1] also suffers from similar problems as the Clique percolation. Link clustering performs hierarchical clustering on the edges of the given network. For each pair of edges (i, k) and (j, k) that shares a single node k, Link clustering computes the Jaccard similarity $JAC(n(i), n(j))$ between the sets of neighbors $n(i)$ and $n(j)$ of node i and j and builds a dendrogram by merging the pair of edges with the highest Jaccard similarity. Finally, Link clustering cuts the dendrogram at the point where the partition density, a quality function proposed in [1], is maximized. In the following, we show that Link clustering does not discover the true communities when their overlaps are more densely connected than each individual community.

We consider a network with two overlapping communities A and B with their overlap O (Figure S8). Let A and B each contain $X + Y$ nodes and O contains X nodes. The total number of nodes in the network is thus $X + 2Y$. Moreover, assume that the nodes in an individual community are connected with probability p, and that the nodes in the overlap have a higher probability of being connected, say $2p$. Now let's consider the case where the number of nodes in the overlap is not larger than the number belonging a single community ($X \leq Y$).

Now consider that Link clustering computes Jaccard similarity between the neighbors of nodes u and v. Without the loss of generality we can have one of the four cases:

- (1) $u \in O$ and $v \in O$
- (2) $u \in A \setminus O$ and $v \in A \setminus O$
- (3) $u \in A \setminus O$ and $v \in O$
- (4) $u \in A \setminus O$ and $v \in B \setminus O$

We will now show that the Jaccard similarity between a pair of edges in case (1) is higher than in case (2), and that (2) is higher than (3), which is naturally higher than (4). This means that Link clustering will first merge edges between nodes in O, and only then merge the edges between nodes in $A \setminus O$ and edges between nodes in $B \setminus O$. Last, Link clustering will merge the edges with one endpoint in O and the other in $A \setminus O$ ($B \setminus O$). This process will produce the dendrogram illustrated in Figure S8. In particular, this means that regardless where one cuts the dendrogram, Link clustering will fail to correctly identify the community structure of the simple network in Figure S8.

To show this more formally we proceed as follows. Let's consider nodes $a_1, a_2 \in A \setminus O$ and nodes $o1, o2 \in O$, and their neighbors $n(a_1), n(a_2), n(o1), n(o2)$. We show that in expectation the following is true:

$$\frac{|n(o1) \cap n(o2)|}{|n(o1)| + |n(o2)|} > \frac{|n(a_1) \cap n(a_2)|}{|n(a_1)| + |n(a_2)|} \geq \frac{|n(a_1) \cap n(o1)|}{|n(a_1)| + |n(o1)|}$$

The above inequalities are equivalent to $JAC(n(o1), n(o2)) > JAC(n(a_1)$ and $n(a_2)) > JAC(n(a_1), n(o1))$. We have $|n(o1)| = |n(o2)| = 2pX + 2pY$, $|n(a_1)| = |n(a_2)| = pX + pY$ and we aim to compute the expected values of the sizes of the intersections between $n(o1)$, $n(o2)$, $n(a_1)$, and $n(a_2)$. For example, $|n(o1) \cap n(o2)|$ is $4p^2X + 2p^2Y$ in expectation because $o1$ and $o2$ have a common neighbor in single community regions ($2Y$ nodes) with probability p^2 and in overlap (X nodes) with probability $(2p)^2$. By the same logic, $|n(a_1) \cap n(a_2)| = p^2X + p^2Y$, and $|n(a_1) \cap n(o1)| = 2p^2X + p^2Y$. From these, we derive the following:

$$\frac{|n(o1) \cap n(o2)|}{|n(o1)| + |n(o2)|} = p\frac{2X + Y}{2X + 2Y} > \frac{p}{2} = \frac{|n(a_1) \cap n(a_2)|}{|n(a_1)| + |n(a_2)|} \geq p\frac{2X + Y}{3X + 3Y} = \frac{|n(a_1) \cap n(o1)|}{|n(a_1)| + |n(o1)|},$$

where the last inequality $1/2 \geq (2X + Y)/(3X + 3Y)$ comes from our assumption that $X \leq Y$. Therefore, Link clustering yields dendrogram in Figure S8, which first merges edges in O and then merges edges in the two non-overlapping parts ($A \setminus O, B \setminus O$) and only then merges edges between the overlapping and the non-overlapping parts ($O, A \setminus O$), ($O, B \setminus O$).

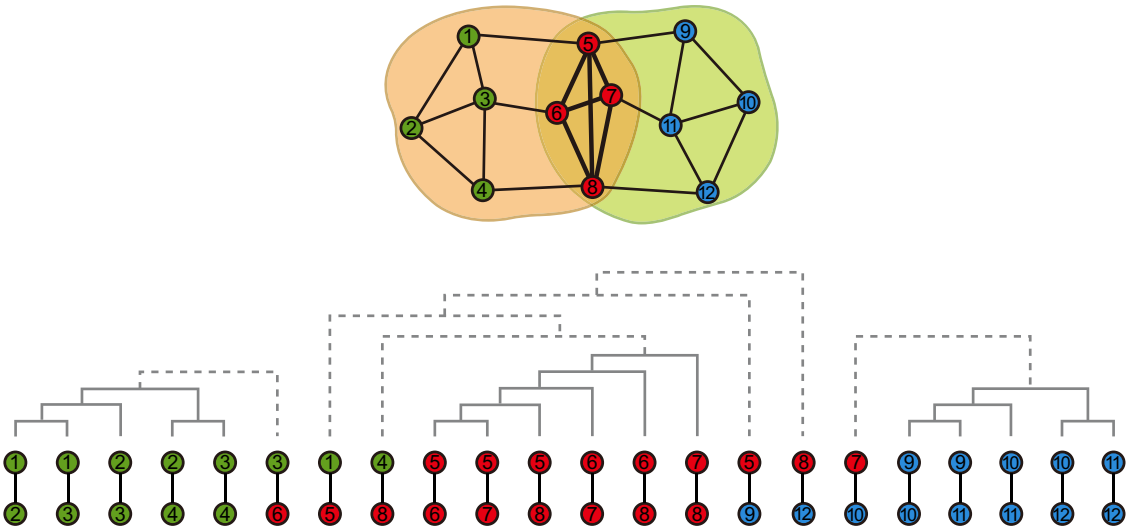


Figure S9: **Example of Link clustering on densely overlapping communities.** From the network at the top, Link clustering produces the dendrogram at the bottom. Since the method first groups the overlapping nodes (nodes $5, 6, 7, 8$) together, it is unable to find that these overlapping nodes belong to both communities.

Figure S9 gives a concrete example of a networks where the result of Link clustering gives counterintuitive results. Green nodes $(1, 2, 3, 4)$ belong to one community, blue nodes $(9, 10, 11, 12)$ belong to the second community and red nodes $(5, 6, 7, 8)$ belong to both communities. As we showed in our analysis, Link clustering merges the edges inside the overlaps (between the red nodes) together and then it merges the nodes in a non-overlapping part (between the green nodes or between the blue nodes). Consequently, Link clustering identifies the overlap of the communities as a separate community and is unable to find that the overlapping nodes belong to two communities at the same time.

S3.3 Stochastic block models

Last, we briefly mention that various kinds of stochastic block models [2, 28, 31] also will not be able to correctly discover communities with dense overlaps. We show this for three variants of stochastic block models: the traditional stochastic block model [28], the Degree-corrected stochastic block model [31] as well as the Mixed-membership stochastic block model [2].

The Stochastic block model [28] partitions a network into disjoint blocks and assigns an edge probability to each block. The only way for the model to increase the edge probability among the nodes in the community overlap is to regard overlaps as separate communities with higher edge density than the individual non-overlapping parts of communities. For example, Figure S10 illustrates the adjacency matrix of the network from Figure S6a and the block structure as discovered by the stochastic block model. Instead of two overlapping three communities are discovered.

The degree-corrected stochastic block model [31] relaxes the assumption that nodes in the same community have similar degrees. By assigning a high degree to nodes in community overlaps, it might be possible to extend the model to increase the edge probability between

overlapping nodes without treating the overlap as a separate community. However, the present version of the model assumes non-overlapping community structure and thus it cannot tell that the overlapping nodes belong to multiple communities.

Last, the mixed-membership stochastic block model can discover overlapping communities. However, the model cannot express dense community overlaps. In fact based on the input matrix from Figure S6a the model will identify three blocks as illustrated in Figure S10. The reason for this is that the edge probability between two nodes that belong to communities A and B is weighted average of $P(A, A)$, $P(A, B)$, and $P(B, B)$, where $P(X, Y)$ is an edge probability between a node in community X and a node in community Y. This means that the edge probability between the two nodes that share multiple communities is smaller than the maximum of $P(A, A)$ and $P(B, B)$ (due to the weighted summation). Therefore, the edge probability between overlapping nodes cannot be higher than the edge probability between nodes in an individual community.

S3.4 Other models of network communities

Our findings here present a new way of thinking about network communities as overlapping tiles. In particular, in the same way as thickness of tiles increases in the areas where the tiles overlap, the "thickness" (i.e., density) of edges increases in the areas where two or more communities overlap. To best of our knowledge our work here is the first to make these observations and build the conceptual understanding. However, we were able to identify a small number of other community detection methods that should in principle be able to correctly identify densely overlapping communities (as illustrated in Figure S6a). We note that none of these methods has been tested specifically whether it can detect densely overlapping communities.

First method that is likely to be able to detect densely overlapping communities is a statistical model of network communities by Ball et al. [6]. In particular, Ball et al. present an extension of stochastic block model where node community memberships are not modeled by a multinomial distribution but every node i maintains a factor $k_{i,c}$ that models the amount by which node i belongs to community c. This way one can think of node membership being described by a non-normalized multinomial distribution which allows for modeling an increased density of the edges in the community overlaps. However, due to model's generality the model inference method is not particularly scalable. Similarly, Mørup et al. [43] developed a non-parametric Bayesian multiple membership latent feature model for networks where edges of the network are generated by using a "soft-OR" function. And last, Gregory et al. [26] also developed a heuristic method for network community detection for which our analysis shows that it might be able to correctly identify densely overlapping communities.

S4 Mathematical model of communities: the AGM model

We just illustrated that most commonly used state-of-the-art community detection methods fail to properly detect communities with dense overlaps. And thus a new approach is needed.

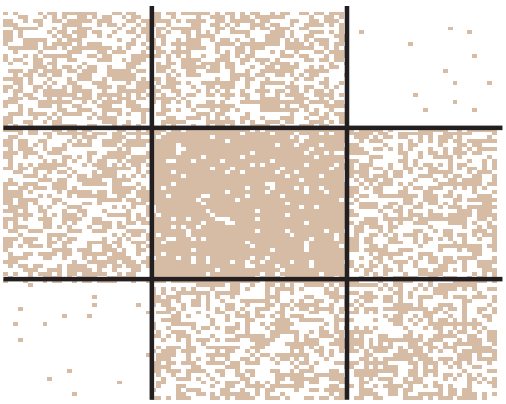


Figure S10: **The result of a Stochastic block model and the Mixed-membership stochastic block model on a network of two communities with dense overlap.** The adjacency matrix of the network in Figure S6a is shown and the bold lines denote the partitions discovered by stochastic block models. See main text for further discussion.

S4.1 The Community-Affiliation Graph Model

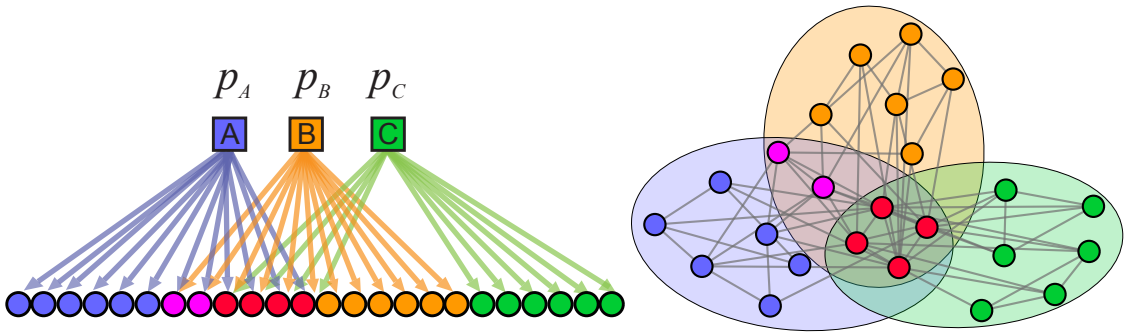
Next we develop a model-based community detection method that successfully detects overlapping, non-overlapping, as well as nested communities in networks. First, we present a simple, conceptual model of network formation, which naturally leads to densely overlapping communities. Then, we describe a method to 'fit' the model to a given network and this discover communities. The code of our approach to community detection is available at `http://snap.stanford.edu/agm`.

We build on Breiger's foundational work [11] where it has been recognized that communities arise due to shared group affiliations [11, 19, 57]. We present the *Community-Affiliation Graph Model* (AGM), a simple probabilistic generative model for networks that captures the observed phenomena and reliably reproduces the organization of networks into communities and the structure of community overlaps [**?**].

Consider a pair of people that are members of several different interest based communities. By having more interests in common, they are more likely to meet and link. Based on this example we require two ingredients: a way to capture memberships of nodes to communities and then a mechanism that gives nodes that share several communities multiple chances to create links among each other.

To capture memberships of nodes to communities we use a bipartite affiliation graph that links nodes of the network to communities that they belong to. We obtain the second ingredient by assigning a single parameter to each community. This parameter captures the probability that nodes belonging to that community create an edge. Nodes belonging to multiple common communities get multiple chances to form links. Thus, naturally, the more communities a pair of nodes shares, the higher is the probability of linking.

Figure S11a illustrates the essence of our model. We start with a bipartite graph where the nodes at the bottom represent the nodes of the social network and the nodes on the top represent communities. The edges between nodes of the network (bottom) and the communities (top) indicate node-community memberships. We denote the bipartite affiliation graph as $B(V, C, M)$, where V denotes the set of nodes of the original network G, C is a set



(a) Community Affiliation Network **(b) Overlap**

Figure S11: **Community-Affiliation Graph Model. (a)** Community-Affiliation Graph Model (AGM): Circles represent communities and squares represent the nodes of the observed network. Edges represent node community memberships. **(b)** Network generated by the Community-Affiliation Graph Model in **(a)**. The AGM **(b)** accurately models dense community overlaps.

of communities, and there is an edge $(u, c) \in M$ from node $u \in V$ to community $c \in C$ if node u belongs to community c.

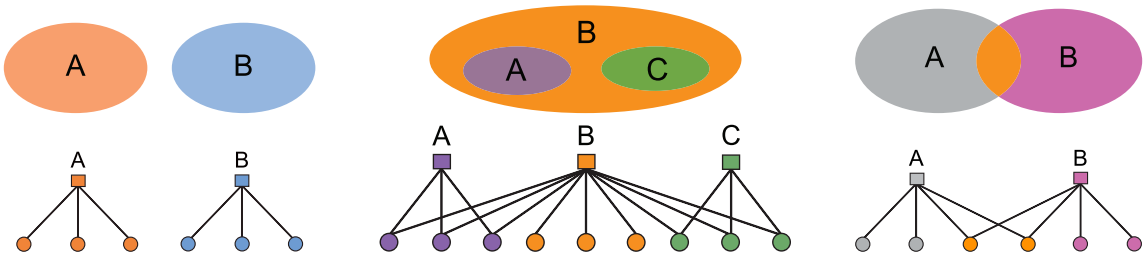
Now, given the affiliation network $B(V, C, M)$, we describe the process to generate the underlying network $G(V, E)$. To achieve this we need to specify the process that generates the edges E of G given the affiliation graph B. We consider a simple parameterization where we assign a parameter p_c to every community $c \in C$. The parameter p_c models the probability of an edge forming between two members of the community c. In other words, we simply generate an edge between a pair of nodes that belongs to community c with probability p_c. Each community c creates edges independently. However, if the two nodes have already been connected via some other common community membership, then the duplicate edge is not included in the graph $G(V, E)$.

Definition 1. *Let $B(V, C, M)$ be a bipartite graph where V is a set of nodes, C is a set of communities, and an edge $(u, c) \in M$ connects node $u \in V$ to community $c \in C$ if u belongs to community c in the network G. Also, let $\{p_c\}$ be a set of probabilities for all $c \in C$. Given the affiliation network $B(V, C, M)$ and $\{p_c\}$, the Community-Affiliation Graph Model generates a graph $G(V, E)$ where the node set V and the edge set E are defined as follows. For each pair of nodes $u, v \in V$, the AGM creates edge $(u, v) \in E$ with probability $p(u, v)$:*

$$p(u, v) = 1 - \prod_{k \in C_{uv}} (1 - p_k), \tag{1}$$

where $C_{uv} \subset C$ is a set of communities that u and v share ($C_{uv} = \{c | (u, c), (v, c) \in M\}$).

Note that this simple process already ensures that pairs of nodes that belong to multiple common communities are more likely to link. This is due to the fact that nodes that share multiple community memberships receive multiple chances to create a link. For example, pairs of nodes in the overlap of communities A and B (but not to C) in Figure S11a get two chances to create an edge. First they can be connected with probability p_A (due to their



(a) Non-overlapping (b) Nested (c) Overlapping

Figure S12: **Flexibility of AGM.** AGM allows for rich modeling of network communities: (a) non-overlapping, (b) nested, (c) overlapping. In (a) we can assume that nodes in disjoint communities connect with small probability ε which allows for sparse links between communities A and B.

membership in community A), and then also with probability p_B (due to membership in B). While pairs of nodes residing in the non-overlapping region of A link with probability p_A, nodes in the overlap link with probability $1-(1-p_A)(1-p_B) = p_A+p_B-p_Ap_B \geq p_A$, which means that overlaps of communities are more densely connected than the non-overlapping parts.

In the formulation of Equation 1, AGM does not account for the edges between the nodes that do not share any common communities. To account for this, we simply assume that nodes which have no communities in common link with a small probability ε. In all our experiments we simply set $\varepsilon = 1/|V|^2$, where $|V|$ is the number of nodes in a given network.

S4.2 Flexibility of AGM

Figure S12 illustrates the flexible nature of the Community-Affiliation Graph Model that allows for modeling any combination of network community structures: Traditional non-overlapping communities can be modeled by the affiliation graph where each network node links only to a single community node (Figure S12a). Hierarchically nested communities can be modeled by the affiliation graph where subsets of network nodes belong to more and more specialized communities (Figure S12b). Communities with overlaps can be modeled by the affiliation graph where some nodes belong to multiple communities while other nodes belong to only one community.

S4.3 Community detection with AGM

Having defined the AGM model we now explain how we detect network communities by fitting the AGM to a given network. Given network $G(V, E)$, we aim to detect communities by *fitting* the AGM (*i.e.*, finding affiliation graph B and parameters $\{p_c\}$) to the underlying network G by maximizing the likelihood $L(B, \{p_c\}) = P(G|B, \{p_c\})$:

$$\underset{B,\{p_c\}}{\operatorname{argmax}} L(B, \{p_c\}) = \prod_{(u,v)\in E} p(u, v) \prod_{(u,v)\notin E} (1 - p(u, v)) \tag{2}$$

To maximize the likelihood L we employ coordinate ascent strategy where we update $\{p_c\}$ fixing B and then we update B with $\{p_c\}$ fixed.

For now we assume the number of communities K ($K = |C|$) is known. We later show how to automatically determine K. We start the process by generating a random affiliation graph B on $|V|$ network nodes and K community nodes. We then proceed by updating $\{p_c\}$.

Updating $\{p_c\}$. With B fixed, we aim to find $\{p_c\}$ by solving the following optimization problem:

$$\arg\max_{\{p_c\}} \prod_{(u,v)\in E} (1 - \prod_{k\in C_{uv}} (1 - p_k)) \prod_{(u,v)\notin E} (\prod_{k\in C_{uv}} (1 - p_k)) \tag{3}$$

with the constraints $0 \leq p_c \leq 1$. Now, we show that we can convert it to a convex optimization problem.

We maximize the logarithm of the likelihood and change variables with $e^{-x_k} = 1 - p_k$:

$$\arg\max_{\{x_c\}} \sum_{(u,v)\in E} \log(1 - e^{-\sum_{k\in C_{uv}} x_k}) - \sum_{(u,v)\notin E} \sum_{k\in C_{uv}} x_k$$

The constraints $0 \leq p_c \leq 1$ become $x_c \geq 0$. This problem is a convex optimization of $\{x_c\}$ and can be thus solved efficiently using convex optimization techniques.

Updating B. Now, given fixed $\{p_c\}$ we aim to update B, while maximizing the likelihood. To this end we use the Metropolis-Hastings [14, 46] algorithm where we slowly update B using small local modifications to it. Given the current community affiliation graph $B(V, C, M)$, we consider three kinds of local modifications to generate a new community affiliation graph $B'(V, C, M')$. As V and C remain the same, we just need to update M to M':

- `LEAVE`: Choose a node-community pair $(u, c) \in M$ uniformly at random and let $M' = M \setminus \{(u, c)\}$.
- `JOIN`: Choose a node-community pair $(u, c) \notin M$ uniformly at random and let $M' = M \cup \{(u, c)\}$.
- `SWITCH`: Choose a node-community pair $(u, c1) \in M$, $(u, c2) \notin M$ at uniformly random and let $M' = (M \setminus \{(u, c1)\}) \cup \{(u, c2)\}$.

Once we have generated new community affiliation B', we accept B' with the Metropolis-Hastings rule. If B' achieves higher likelihood than B, we accept B'. Otherwise, we accept B' with probability $L(B', \{p_c\})/L(B, \{p_c\}))$.

We experimented with many synthetic and real-world networks, and found that the Markov chain of fitting AGM exhibits relatively quick convergence as the likelihood does not increase after roughly $O(|V|^2)$ steps. Although this is not a rigorous performance guarantee, results show that the fitting method works quite well in practice. The algorithm can fit AGM to the networks with up to a few thousand nodes in a reasonable amount of time. Figure S14b shows that how the log-likelihood of the Facebook ego-network (Figure 3a of the main paper) converges after about 10,000 iterations.

Identifiability of AGM. Our first test of the AGM community detection method is to examine whether our fitting method can recover the model parameters given a synthetic

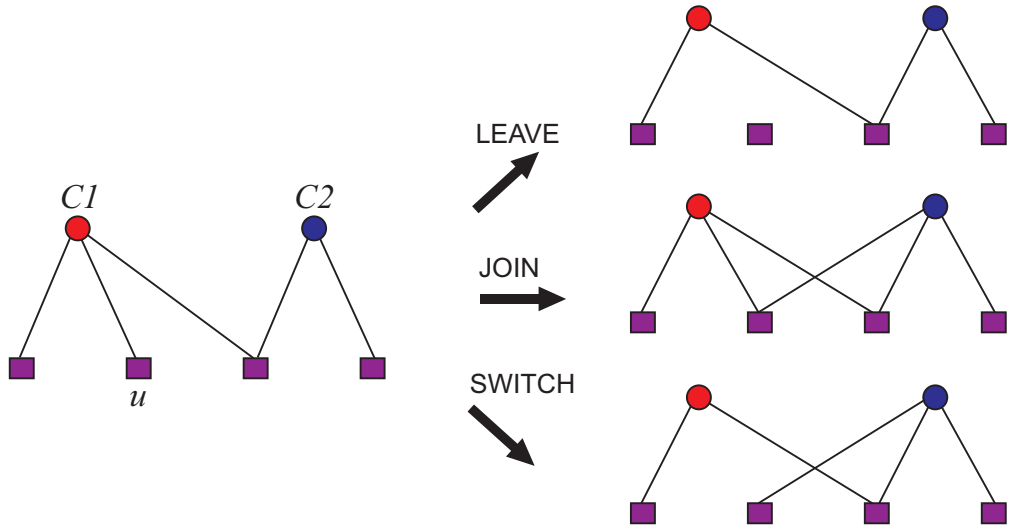


Figure S13: **Local modifications for updating community affiliation graph B.** Given the current community affiliation graph on the left, we consider three local modifications. `LEAVE` considers a node quitting a community. `JOIN` considers a node joining a new community. `SWITCH` causes a node to replace one of its current communities with a new one.

network that was generated by the AGM model itself. Assume a synthetic network G^* that is generated by AGM using some input parameters $B^*, \{p_c\}^*$. Now our goal is to recover $B^*, \{p_c\}^*$ based only on the network G^*.

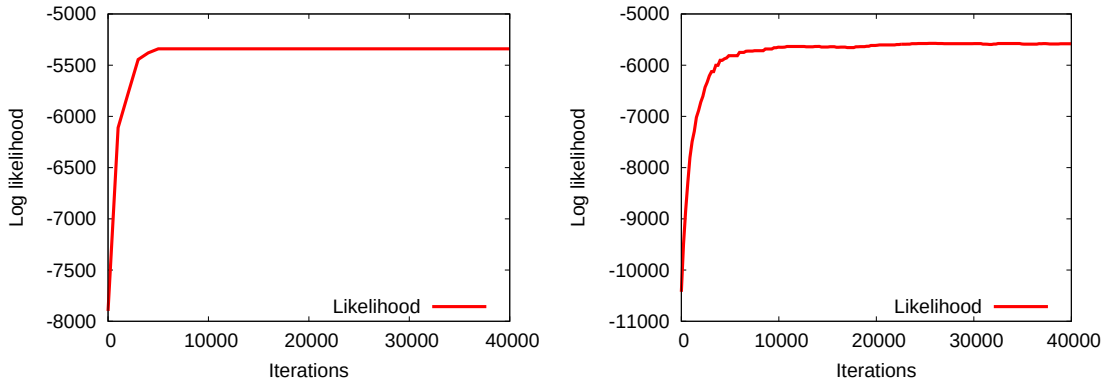
For example, we generated a network with the overlapping communities such that 100 nodes belong to community A, 100 nodes belong to community B, and 50 nodes belong to the both communities. We set $p_A^* = p_B^* = 0.3$ and generate the network G^*. Now given G^* we identify communities by fitting AGM to the synthetic network G. The AGM can discover communities A and B with perfect accuracy and estimate p_A, p_B very closely ($p_A = 0.30, p_B = 0.29$). Figure S14a shows the likelihood as a function of the number of iterations. After 3,000 iterations, the likelihood reaches a plateau and converges.

We also considered a more general cases where we generated some random B^* and randomly assigned parameters $\{p_c\}^*$. In nearly all cases our algorithm was successfully able to recover parameters B^* and $\{p_c\}^*$ given only the synthetic network G^*.

S4.4 Automatically finding the number of communities

To initialize $B(V, C, M)$, we need to set the number of communities $K = |C|$, which in practice is not know in advance. To resolve this, we develop a method to automatically estimate the number of communities. Our approach is based on statistical regularization techniques [58].

Our strategy begins with a candidate set of a large number of communities that might exist in a given network. We generate this candidate set by fitting AGM using a very large number of communities K. Then, we keep removing redundant communities using l_1-regularization until we observe a drop in the log-likelihood. We observe a threshold like behavior of the log-likelihood a as a function of the regularization parameter. We choose



(a) Synthetic network **(b) Facebook**

Figure S14: **Convergence of fitting the AGM to a given network.** The likelihood of AGM versus the number of iterations of Metropolis-Hastings measured on a synthetic network generated by AGM (a) and a Facebook ego-network (b). In both cases AGM reaches converges after around 10,000 iterations and accurately discovers communities.

minimum number k^* of communities before the log-likelihood drops. This way we find the minimum number k^* of communities that is still sufficient to model the structure of a given network.

More precisely we proceed as follows. First, we use a very large number of communities ($|C_0| = O(|V|)$) to fit AGM on the given network $G(V, E)$ and obtain the resulting bipartite community affiliation graph $B_0(V, C_0, M_0)$.

Note that not every community found in B_0 is important. This B_0 is a set of candidate communities from which we find a set of communities that are very likely to exist. Key intuition is that we can ignore communities if their corresponding p_c is 0. Therefore, we aim to reduce the number of communities in B_0 by forcing more and more parameters p_c to zero.

We apply l_1-regularization with parameter λ to a problem of fitting $\{p_c\}$ to B_0. At each value of λ, we solve the following problem:

$$\{\hat{p_c}(\lambda)\} = \underset{\{p_c\}}{\operatorname{argmax}} P(G|B_0, \{p_c\}) - \lambda \sum_c |p_c| \tag{4}$$

Solution $\{\hat{p_c}(\lambda)\}$ is a *sparse* vector with only few individual $\hat{p_c} \neq 0$. Non-zero $\hat{p_c}$ act as indicators of active communities in B_0. We construct $B(\lambda)(V, C(\lambda), M(\lambda))$ by taking the communities in B_0 with non-zero $\hat{p_c(\lambda)}$. Each such $B(\lambda)$ represents a set of active communities at the regularization intensity λ.

Now our goal is to find the value of regularization parameter λ such that we discover the true number of communities in the network. We achieve this by measuring how well $B(\lambda)$ can represent $G(V, E)$ by measuring its likelihood $L(B(\lambda)) = \max_{\{p_c\}} P(G|B(\lambda), \{p_c\})$. Likelihood $L(B(\lambda))$ tells us how well $G(V, E)$ can be explained when we use only $C(\lambda)$ communities. For example, Figure S15 plots $L(B(\lambda))$ and $K(\lambda)$ measured on a network that has $K^* = 2$ true communities.

Notice that whereas $K(\lambda)$ is an almost strictly decreasing function of λ, $L(B(\lambda))$ seems to be a step function which is flat until λ reaches some threshold and then suddenly drops.

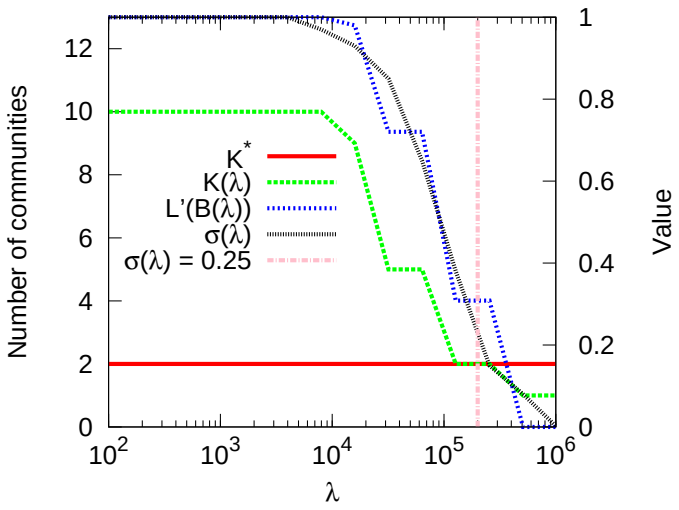


Figure S15: **Automatically determining the number of communities in a given network.** We plot various quantities as a function of regularization intensity λ using two Y-axes. K^*: The true number of communities (using the left axis). $K(\lambda)$: The number of communities we estimate under regularization intensity λ (using the left axis). $L'(B(\lambda))$: The normalized likelihood of $B(\lambda)$ (using right axis). $\sigma(\lambda)$: The sigmoid function fit to normalized $L(B(\lambda))$ (using right axis). Pink dotted vertical line: λ^* at which $\sigma(\lambda^*)$ falls below 0.25 (using the right axis). Red horizontal line, the estimated (as well as the true) number of communities.

This suggests that no more than $K(\lambda)$ communities are needed to explain $G(V, E)$ when λ is relatively small. In other words, $K(\lambda)$ with high $L(B(\lambda))$ gives us a upper bound for the number of communities that exist on the network. The tight upper bound happens at the point at which $L(B(\lambda))$ suddenly drops, and we report such $K(\lambda)$ measured at the quick drop of $L(B(\lambda))$ as our estimate for the number of communities.

Since we cannot examine all possible values of λ, detecting the exact value of λ at which $L(B(\lambda))$ suddenly drops is a challenging task. To find such λ accurately, we approximate $L(B(\lambda))$ by the sigmoid function $\sigma(\lambda) = \frac{1}{1+e^{\alpha\lambda+\beta}}$. We first normalize $L(B(\lambda))$ into $L'(B(\lambda))$ so that the maximum over λ is 1 and the minimum is 0, and then we fit the sigmoid function $\sigma(\lambda)$ to $L'(B(\lambda))$ by finding the optimal parameters α and β [12]. Then we compute λ^* such that $\sigma(\lambda^*) = \delta$ for some constant $\delta \ll 1$, and $K(\lambda^*)$ is our estimate for the number of communities. We experimented with various values of δ and found that setting $\delta = 0.25$ is a reasonable choice.

For example, in Figure S15, our strategy estimates the true number of communities correctly. In the experiments on the real-world networks, our strategy succeeds to estimate the true number of communities more accurately than other methods (Section S5).

The run time of this method mostly depends on fitting B_0 because solving Problem 4 can be done efficiently due to its convexity. Therefore the overhead of automatically finding the number of communities bring little computational overhead in practice. More importantly, with this method we can use the AGM without any parameters using the following two-step strategy. When a network is given, we first estimate the number of communities $\hat{K}$ in the

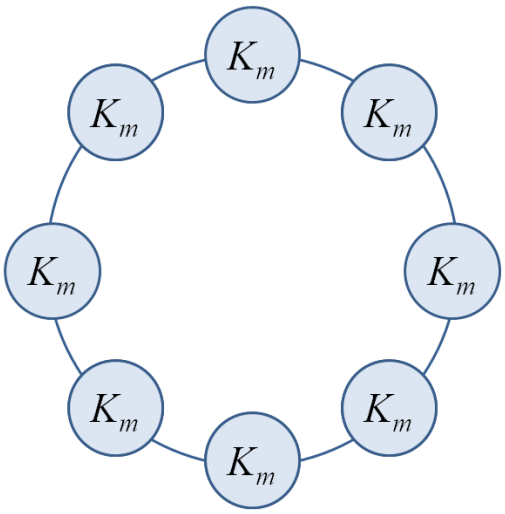


Figure S16: **Ring of cliques.** Ring of cliques K_m (a complete graph of m nodes) considered in [21], where the authors show that Modularity-based methods fail to detect each K_m as a separate community. In our experiments, AGM successfully discovers each K_m as a separate community.

network, and then fit the AGM using our estimate $\hat{K}$.

S4.5 AGM does not suffer from the "resolution" limit

Many community detection methods suffer from the "resolution limit" [8, 21, 24]. In Particular, Fortunato et al. [21] showed that Modularity has a resolution limit in a sense that Modularity cannot detect communities if they are too small.

A ring of cliques (illustrated in Figure S16) is an example of a graph where the resolution limit can be reliably studied. The ring of cliques consists of n cliques K_m, where each K_m is a complete graph on m nodes. The cliques are then connected into a ring by adding a single edge between two consecutive cliques. On such a graph we expect a given community detection method to n find communities—each K_m is a separate community. However, [21] proved that the Modularity-based methods fail to discover each K_m as a separate community if n is larger than the square root of the number of edges in the network, *i.e.*, Modularity-based methods fail for a network consisting of many small modules.

We run AGM for the same values of n, m used in [21] ($n = 30, m = 5$), and find that AGM correctly identifies both the number of communities n as well as detects each K_m as a separate community. We also experimented with many other values for n and m (*e.g.*, $n = 50, m = 10$, $n = 100, m = 10$) and observed that AGM perfectly identifies each K_m as a community. From these experiments we conclude that AGM does not suffer from the resolution limit.

S4.6 Anecdotal comparison between AGM and the existing methods

In the main text, we gave an example of Facebook network where AGM correctly identifies overlapping network communities while the existing methods fail. In this section, we show two more anecdotal evidences: Network of famous philosophers in Wikipedia [1], and the *E.Coli* metabolic network [1].

First, we apply the AGM to the network of Wikipedia pages for famous philosophers [1] (Table S2), where the nodes are the Wikipedia pages and the edges mean hyperlinks between the pages. For the sake of visualization, we show the communities that Francis Bacon belongs to (Fig. **??**). Our AGM detects two communities that Bacon belongs to: one with scientists and the other with philosophers (Fig. S17a), whereas existing methods fail to produce as interpretable results as AGM (Fig. S17b, S17c, and S17d).

Second, we also consider the *E.Coli* metabolic network [1] where nodes are metabolites and edges mean interactions. Here, the existing methods miss the communities shared by very important metabolites. For example, between H2O and CO2, the three existing methods that we consider report that the two metabolites [47, 1, 2] share only one community, whereas the AGM detects 18 communities shared by the two metabolites.

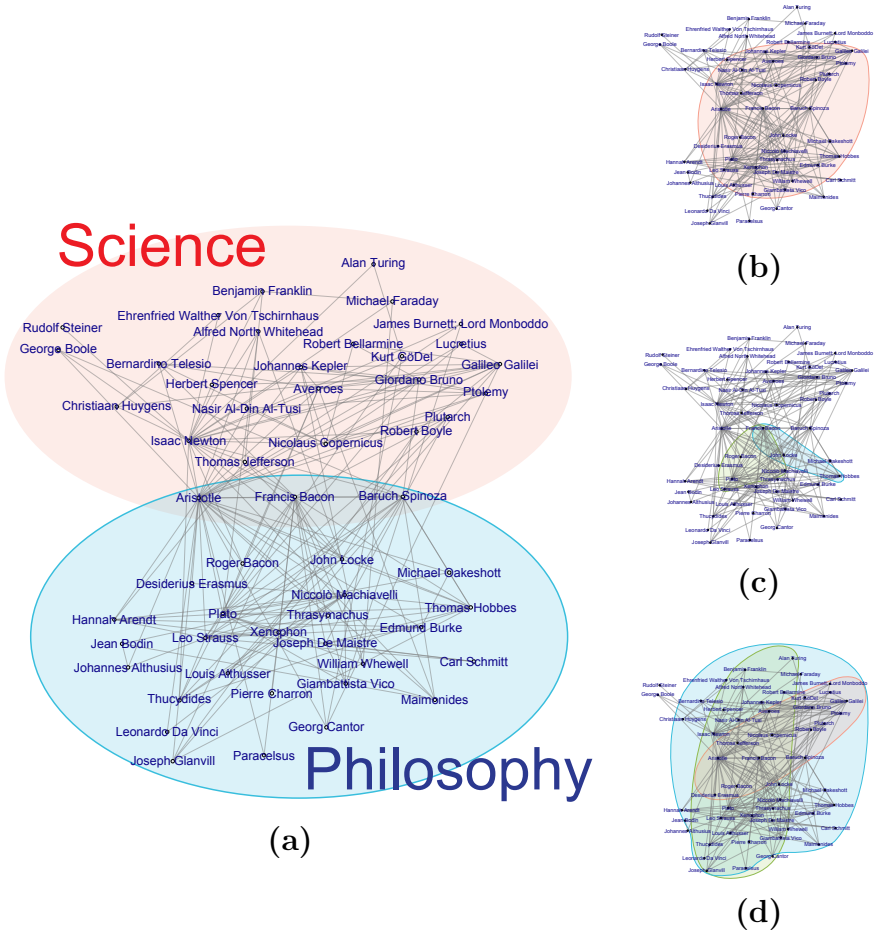


Figure S17: **Examples of detected communities in the network of Wikipedia articles for philosophers (a–d).** The detected communities that Francis Bacon belongs to are displayed by filled regions. **(a)** The AGM detects two communities which represent Bacon's connections to scientists and to philosophers respectively. On the other hand, clique percolation **(b)**, link clustering **(c)**, and mixed-membership block models **(d)** detect less interpretable communities.

S5 Experiments: Networks with ground-truth communities

In this section we evaluate the performance of AGM and compare it to the state-of-the-art community detection methods on a range of networks from a number of different domains and research areas. We perform experiments on the 6 networks described in Section S1 (Table S1) for which we have explicitly labeled ground-truth communities. Availability of ground-truth communities allows us to quantify the 'accuracy' of community detection methods by comparing the level of correspondence between the detected and the ground-truth communities.

S5.1 Experimental setup

We focus on the evaluation of community detection methods on their ability to correctly identify overlapping communities. With this purpose in mind, running community detection algorithms on a whole network is not an effective way for two reasons. First, for some nodes we have no ground-truth community labels. And more importantly, *none* of the community detection algorithms that we consider is scalable to the size of networks we consider here.

To resolve this we proceed by finding subnetworks with highly overlapping community structure from a given network $G(V, E)$. To obtain one such subnetwork we pick a random node $u \in V$ that belongs to more than one ground-truth community and then take the induced subgraph of G consisting of all the nodes that share at least one ground-truth community membership with u. Figure S18 illustrates how a subnetwork (right) is created from $G(V, E)$ (left) when the red node u is chosen. We identify all the member nodes of the communities that the red node belongs to and then construct the induced subgraph on the right. This way we obtain subgraphs that are of reasonable size and contain fully labeled overlapping communities.

We sample 500 subnetworks for each of the six networks from Table S1. We control the sampled networks to have similar number of ground-truth communities across the data sets. In particular, we sample 100 networks for each of 2, 3, 4, 5 communities on each network. And the last 100 sampled networks have more than 5 communities.

We also considered many alternative ways of obtaining small enough subnetworks so that community detection algorithms could be run. For example, we considered a strategy where given a network $G(V, E)$, we pick a random node $u \in V$ and find a set of nodes V_u that are less than 2-hop away from u. We then construct the induced subgraph of V_u and take communities that have more than 50% of their members in this induced subgraph. We also considered the approach where we created a random set of "connected" communities that either share an edge or a node. In all these cases the results we obtained were qualitatively similar and lead to same conclusions.

S5.2 Methods for comparison

In order to evaluate performance, we compare AGM to existing, state of the art community detection methods. We choose the four most prominent community detection methods:

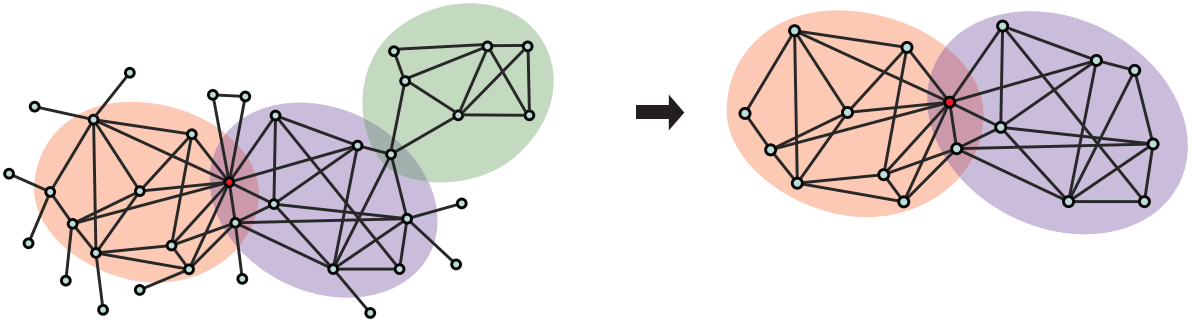


Figure S18: **Sampling subnetworks with community overlaps.** On the left is a part of a full network $G(V, E)$ and on the right is the subnetwork that we sample. We randomly pick a node (the red node on the left) and then construct a subnetwork consisting of the communities that the red node belongs to.

- Link clustering (LC) [1]
- Clique Percolation Method (CPM) [47]
- Infomap [55]
- Mixed membership stochastic blockmodel (MMSB) [2]

Link clustering, the Clique percolation method and Mixed membership stochastic blockmodels are regarded as state-of-the-art algorithms for detecting overlapping communities, and Infomap is a state-of-the-art method for detecting non-overlapping communities.

While Infomap and Link clustering are parameter-free, the Clique percolation method requires an input parameter k that determines the size of cliques to be percolated. We choose the clique size $k = 5$ as we find that CPM with $k = 5$ estimates the number of communities most accurately in all the networks. CPM with $k = 6$ tends to estimate too few communities and CPM with $k = 4$ detects too many communities. MMSB requires the number of communities to be given. To this end we use the Bayes Information Criterion (BIC) described by the authors [2] to determine the number of communities.

S5.3 Evaluation metrics

To quantify the performance we measure the level of agreement between the detected and the ground-truth communities. Given a network $G(V, E)$, we consider a set of ground truth communities C^* and a set of detected communities $\hat{C}$ where each ground-truth community $C_i \in C^*$ and each detected community $\hat{C}_i \in \hat{C}$ is defined by a set of its member nodes. To compare $\hat{C}$ and C^*, we use four performance metrics:

- **Average F1 score** [40] is the average of the F1-score of the best-matching ground-truth community for each detected community, and the F1-score of the best-matching detected community for each ground-truth community. In particular, we compute $F_g(C_i) = \max_j F1(C_i, \hat{C}_j)$ for each ground-truth community C_i and $F_d(\hat{C}_i) = \max_j F1(C_j, \hat{C}_i)$ for each detected community $\hat{C}_i$, where $F1(S_1, S_2)$ is the harmonic mean of precision

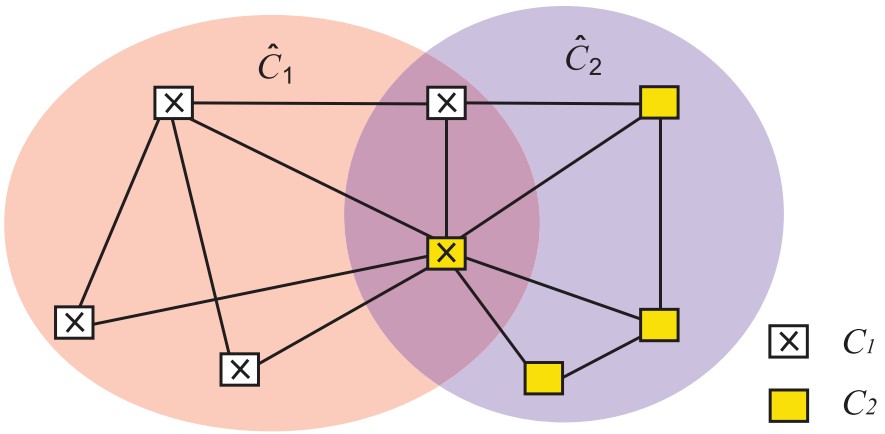


Figure S19: The example of two detected communities (shaded regions $\hat{C}_1$ and $\hat{C}_2$) and two ground-truth communities C_1, C_2 ("X"-marked nodes belong to C_1 and yellow nodes belong to C_2). In this example detected communities $\hat{C}_1$ and $\hat{C}_2$ achieve good correspondence to the ground-truth communities C_1, C_2 (only a single mistake is made).

and recall between two node sets S_1, S_2. The average F1 score is $\frac{1}{2}(\bar{F}_g + \bar{F}_d)$ where $\bar{F}_g = \frac{1}{|C^*|}\sum_i F_g(C_i)$ and $\bar{F}_d = \frac{1}{|\hat{C}|}\sum_i F_d(\hat{C}_i)$.

- **Omega Index** [27] is the accuracy on estimating the number of communities that each pair of nodes share. For each pair of nodes $u, v \in V$ we define C_{uv} to be the set of ground-truth communities to which both u and v belong and $\hat{C_{uv}}$ to be the set of detected communities to which the both nodes belong. Then the Omega Index is $\frac{1}{|V|^2}\sum_{u,v\in V} \mathbf{1}\{|C_{uv}| = |\hat{C_{uv}}|\}$.

- **Normalized Mutual Information** [33] adopts the criterion used in information theory. The Normalized Mutual Information is $1 - \frac{1}{2}(H(C^*|\hat{C}) + H(\hat{C}|C^*))$ where $H(A|B)$ is the extension of entropy when A, B are sets of sets [33].

- **Accuracy in the number of communities** is $1 - \frac{||C^*| - |\hat{C}||}{|C^*|}$, which is the relative error in predicting the number of communities.

Note that all performance metrics take values on the interval $[0, 1]$ and higher values correspond to better performance. In all metrics score 1.0 is achieved when the detected communities $\hat{C}$ are exactly the same as the ground-truth communities C^*.

Figure S19 gives a simple example of two ground-truth communities C_1, C_2 and two detected communities $\hat{C}_1, \hat{C}_2$. $\hat{C}_1$ and $\hat{C}_2$ are denoted by shared ellipses. The nodes marked with "X" belong to C_1 and the nodes with yellow background belong to C_2.

In this example detected community $\hat{C}_1$ perfectly corresponds to C_1 but the detected community $\hat{C}_2$ only partially corresponds to C_2. For this particular case the F1-score is 0.94 as we see $F_g(C_1) = 1.0, F_g(C_2) = 0.89$ and $F_d(\hat{C}_1) = 1.0, F_g(\hat{C}_2) = 0.89$. The Omega Index is 0.85 and the Normalized Mutual Information is 0.78. And the accuracy in the number of communities is 1.

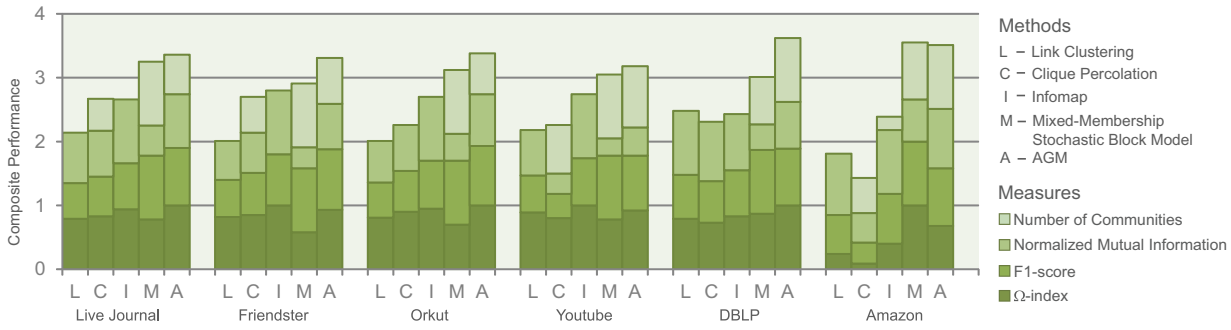


Figure S20: The composite performance of the community detection methods on six networks with ground-truth communities. The AGM gives overall best performance.

S5.4 Results

For each of 6 networks and the 500 subnetworks (3,000 subnetworks total) we run AGM as well as the four other methods. For each subnetwork and method we compute the four evaluation metrics. We then compute the average value of a given performance metrics for a given method and network. Now, for each metric, we normalize the scores of methods so that the best performing method for each score has the value of 1.0. Finally, we compute the composite performance by summing up the four normalized scores. If a method achieves better value than any other method in all the scores, then the composite performance of the method is 4.0.

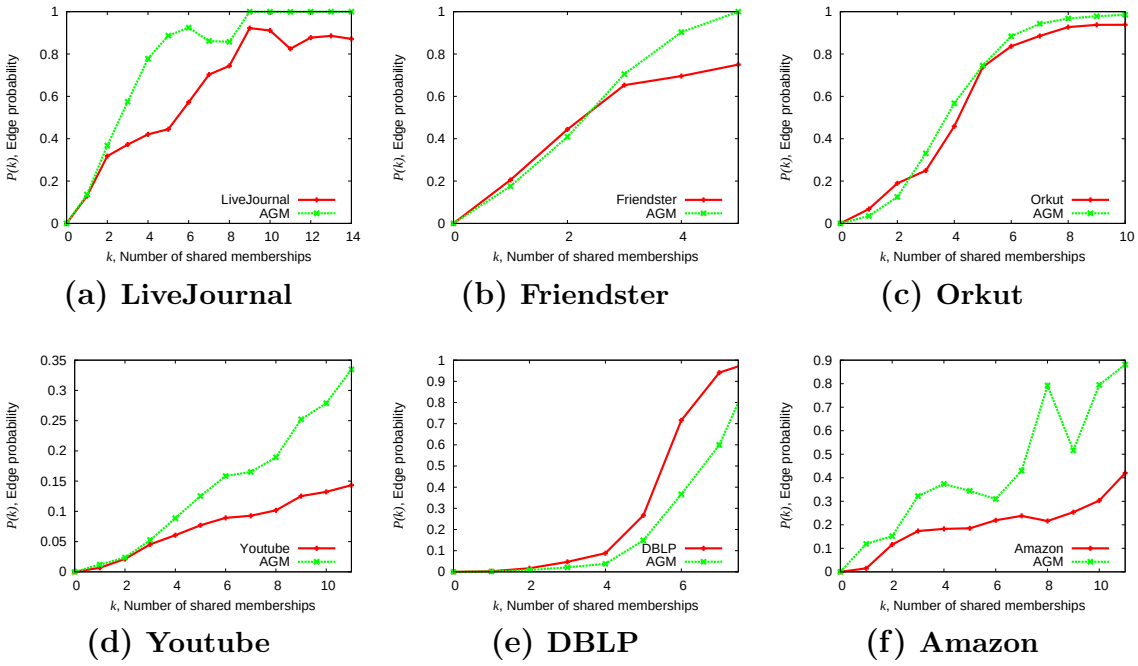
Figure S20 displays the composite performance of each of the 5 methods over the six networks with ground-truth communities. Overall, we notice that AGM gives superior overall performance on all networks except the Amazon, where it ties with MMSB. Furthermore, AGM detects highest quality communities for most individual performance metrics in all networks. On average, the composite performance of AGM is 3.40, which is 61% higher than that of Link clustering (2.10), 50% higher than that of CPM (2.41), 30% higher than that of Infomap and 8% higher than that of MMSB (3.25). The absolute average value of Omega Index of AGM over the 6 networks is 0.46, which is 21% higher than Link clustering (0.38), 22% higher than CPM (0.37), 5% higher than Infomap (0.44) and 26% higher than MMSB (0.36).

In terms of absolute values of scores, AGM archives the average F1 score of 0.57, average Omega index of 0.46, Mutual Information of 0.15 and accuracy of the number of communities 0.42. We also note that AGM also heavily outperforms CPM with other values of k (*e.g.*, CPM with $k = 3, 4, 6$).

S5.5 Experiments on modeling the network structure

Having shown that AGM reliably discovers community structure of real-world networks, we now proceed to evaluate how accurately AGM models *network* connectivity structure itself. For a given network G, we estimate input parameters of AGM by fitting AGM to G, and then we generate a synthetic network $\hat{G}$ with AGM using our estimates for input parameters (Equation 1). We then compare the structure of G with that of $\hat{G}$.

First, we verify that the edge probability between a pair of nodes in $\hat{G}$ is an increasing



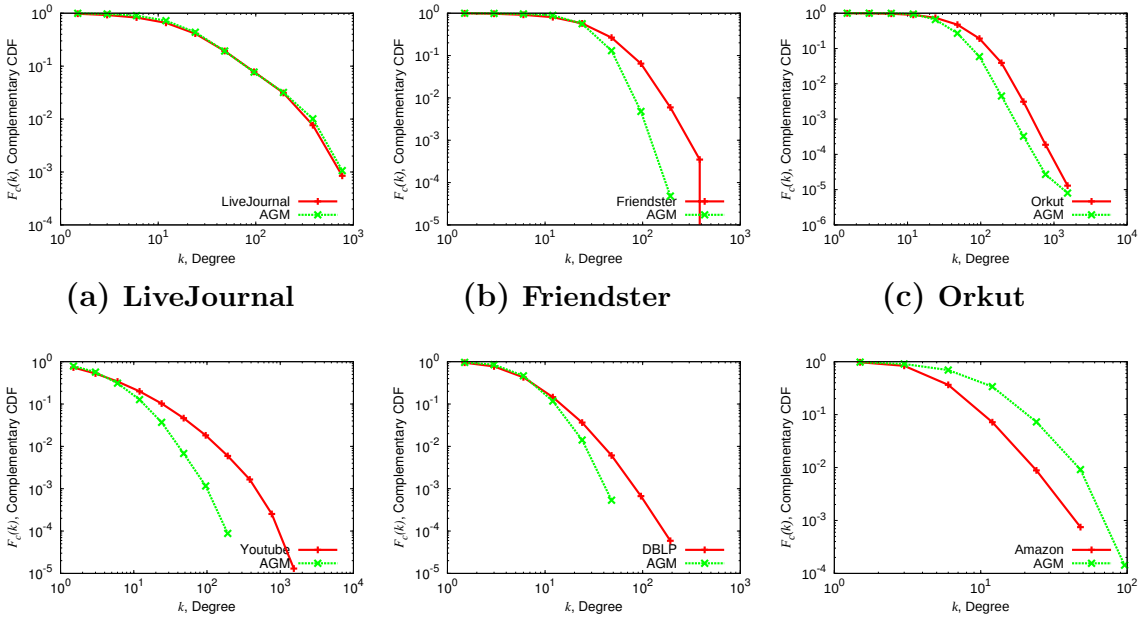
(a) LiveJournal **(b) Friendster** **(c) Orkut**

(d) Youtube **(e) DBLP** **(f) Amazon**

Figure S21: **Edge probability measured on the network reproduced by AGM.** Edge probability $P(k)$ as a function of the number of common community memberships k in the 6 networks (red) and as modeled by the AGM (green), which reliably captures the pattern.

function of the number of communities that the nodes share (Figure S5). Figure S21 plots the edge probability as a function of the number of common communities between a pair of nodes in G (red curves) and $\hat{G}$ (green curves). We note that the AGM successfully reproduces the increase in the edge probability, which means that AGM naturally produces a network with dense community overlaps.

Second, we study whether the AGM is able to generate overall realistic networks. We examine how well the global structural properties of the synthetic network $\hat{G}$ match the properties of the ground-truth network G. Figure S22 plots the node degree distribution. Red curves are the statistics of the ground-truth network and green curves are the statistics of the synthetic networks. Figures show that synthetic networks generated by AGM exhibit similar structural properties to real-world networks. Overall, these results demonstrate that the AGM is not only able to reliably discover the network communities but also accurately captures the structure of the underlying networks.



(a) LiveJournal **(b) Friendster** **(c) Orkut**

(d) Youtube **(e) DBLP** **(f) Amazon**

Figure S22: **Degree distribution of networks reproduced by AGM.** Complementary cumulative degree distribution function $F_c(k)$ as function of node degree k in the 6 networks (red) and as modeled by AGM (green). The synthetic networks generated by AGM exhibit a heavy-tailed degree distribution in a similar way as the real-world networks.

S6 Experiments: Small networks

As a point of comparison and a sanity check we also run the AGM on small networks that were thoroughly studied in literature [3, 6, 16, 20, 23, 34, 45]. Overall, out goal here is to show that communities obtained by the AGM agree with previous literature. Note that since these networks are known to have non-overlapping community structure, our experiments evaluate how well AGM discovers non-overlapping communities.

NCAA football network.

The first network we examine is the network of American college football studied in [23, 51]. This network represents the schedule of Division I games for the 2000 season: nodes are college teams and an edge between two nodes means that the two teams had a game in the season. By construction, this football network has a very well-defined community structure: The teams are divided into conferences in which most match-ups are made. On average, teams played 7 games in the conference that they belong to and 4 games outside the conference.

Figure S23 shows the football network and the communities detected by AGM. Node color specifies which conference a node belongs to and red circles display the communities detected by AGM. We observe that AGM discovers most conferences perfectly with only a few exceptions, which we find can in fact be nicely explained. For example, the 4 white nodes that AGM assigns to "any community" are "independent" teams that actually do not belong to any conference. And the grey nodes that are split into two communities belong to the Sunbelt conference, which did not enforce its members to play games against each other [23]. This result implies that AGM can accurately detect non-overlapping community structure.

Bottlenose dolphins of Doubtful Sound.

We also consider the social network of a community of 62 bottlenose dolphins living in Doubtful Sound, New Zealand (Figure S24). The network was compiled from seven years of field studies of the dolphins by Lusseau et al. [39]. Nodes are dolphins and edges mean statistically significant frequent association. Lusseau et al. observed the division of dolphins into two groups, represented by black and white nodes. The two shaded regions are two communities detected by AGM, which match the known division quite well. The AGM does not include some nodes to either community, which is plausible because those nodes are very sparsely connected to other nodes in the network, and thus violate the assumption that all nodes of the same community connect with a uniform probability.

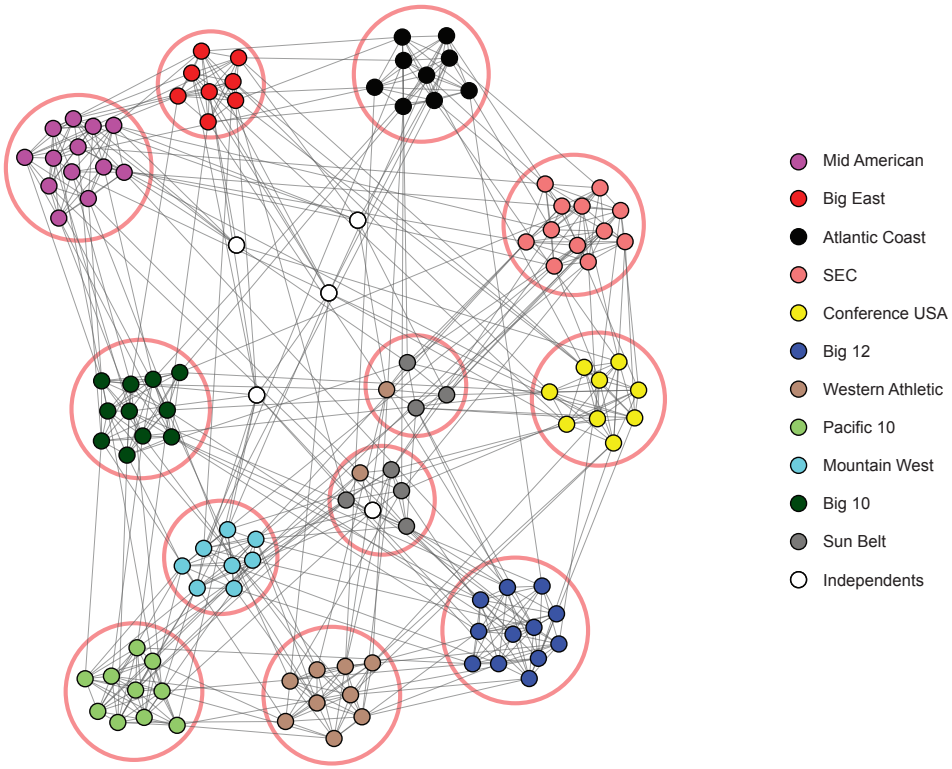


Figure S23: **AGM on the NCAA football network.** Community structure in a network reflecting the schedule of regular season Division I college football games for year 2000 [23]. Nodes are universities and edges represent games between universities. Node colors represent the NCAA conference that the node belongs to. AGM communities, specified by circles, correspond to NCAA conferences with high accuracy.

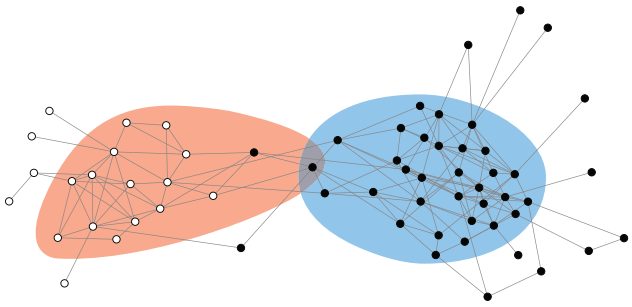


Figure S24: **AGM on a network of bottlenose dolphins of Doubtful Sound.** Community structure in the social network of bottlenose dolphins assembled by Lusseau et al. [39], detected by AGM. Lusseau et al. [39] reported that dolphins are separated into two groups (black and white nodes). AGM reliably captures the existence of two separate groups.

S7 Experiments: Biological networks

Having found that the AGM can discover the community structure on the social and information networks accurately, we now evaluate the AGM on biological networks.

S7.1 Dataset description

We consider the protein-protein interaction (PPI) networks of *Saccharomyces cerevisiae*, which are one of the most complete protein-protein interaction networks available today [63, 32, 1]. The PPI networks are compiled into three different genome-scale networks: yeast two-hybrid (Y2H), affinity purification followed by mass spectrometry (AP/MS), and literature curates (LC). Edges correspond to statistically significant interactions among proteins. In addition to these three networks already mentioned, we use also the union of the three networks (PPI (All)). Basic statistics of the networks are shown in Table S2.

S7.2 Evaluation metrics

We measure the quality of detected communities by using high quality node meta-data. We use the Gene Ontology terms (GO terms) as meta-data [4]. The GO terms provide the most elaborate annotations for the biological roles of groups of proteins in the protein-protein interaction network for three different types: Biological process, Cellular component, and Molecular function. Moreover, there are statistical tools [7, 10] that find the most relevant GO-term for a group of proteins. We quantify the correspondence between the protein communities detected by a method and the significance of the associated GO terms.

Given a PPI network, we detect communities $\hat{C}$ over the whole network. For each detected community $\hat{C}_i \in \hat{C}$, we find the most statistically relevant GO term and its p-value, $p_v(\hat{C}_i)$ using the GO term finder [10]. We then report the average p-value over the detected communities: $\bar{p}$ ($\bar{p} = \frac{1}{|\hat{C}|}\sum_i p_v(\hat{C}_i)$). We compute the average p-value $\bar{p}$ for the three types of GO terms (biological process, cellular component and molecular function) and use $-\log(\bar{p})$ for the each GO term type as a separate score. We take the negative logarithm of the p-value to transform it so that the performance score is a nonnegative increasing function of the quality. Finally, we normalize the scores so that the best method achieves the value of 1.0 as we did in Section S5.

S7.3 Results

Figure S25 displays the composite performance for biological networks for each of the five methods (LC, CPM, Infomap, MMSB, and AGM). Note that the AGM attains the best composite performance in all four networks by a huge margin. CPM is the second best, Link clustering is the third, and Mixed membership stochastic blockmodel scores the worse.

Since the comparison is made on the logarithms of p-values, this result suggests that communities detected by the AGM are far more statistically relevant than those detected by other methods. For example, the average p-value of the AGM communities over all the networks is 0.008, which is 13-times better than that of Link clustering (0.11), 15-times better than of Clique percolation (0.12), 12-times better than of Infomap (0.096) and also

Dataset	N	E	$\langle C \rangle$	$\langle D \rangle$	$\langle k \rangle$
Protein-protein interaction networks [1] (Section S7)					
PPI (Y2H) [63]	1,647	2,518	0.10	6.60	3.06
PPI (AP/MS) [15]	1,004	8,319	0.72	6.51	16.57
PPI (LC) [50]	1,213	2,556	0.46	8.56	4.21
PPI (All) [1]	1,647	12,784	0.41	6.24	8.60
Networks from Ahn et al. [1] (Section S8)					
Metabolic [18]	1,042	8,756	0.74	3.15	16.81
Philosophers [1]	1,218	5,972	0.30	4.25	9.81
Word Association [44]	5,018	55,232	0.19	4.04	22.01
Social networks (Section S9)					
MSN Finland	592,982	2,448,213	0.13	7.95	8.26
LinkedIn	254,151	482,286	0.09	7.18	3.80
Web graphs (Section S9)					
Web-Stanford	255,265	1,941,926	0.62	9.36	15.21
Web-Notre Dame	325,729	1,090,108	0.23	9.42	6.69
Web-Berkeley/Stanford	654,782	6,581,871	0.61	10.06	20.10
Foodweb networks (Section S9)					
Foodweb-wet	128	2,075	0.33	1.90	32.42
Foodweb-dry	128	2,106	0.33	1.90	32.90

Table S2: **Network statistics.** N: Number of nodes, E: Number of edges, $\langle C \rangle$: Average clustering coefficient [61], $\langle D \rangle$: Average shortest path length, $\langle k \rangle$: Average node degree. The 4 networks in the first block are the protein-protein interaction networks of *Saccharomyces cerevisiae* which are described in Section S7. The next 3 networks in the second block are the networks used in Ahn et al. [1] and we describe these networks in Section S8. The rest 3 blocks are the different types of networks where different kinds of core-periphery structures arise (Section S9). The 2 networks in the third block are social networks, the 3 networks in the fourth block are web graphs, and the 2 networks in the last block are foodweb networks.

119-times better than that of MMSB (0.95). We further investigated the poor performance of MMSB on this network and found it is due to the fact that MMSB tends to find very large communities, which in turn leads to very poor p-values.

In terms of absolute values of p-values the AGM performs quite well. For example, in the AP/MS network, the AGM achieves the average p-value of 1.9×10^{-5}, which suggest the high significance of the detected communities.

Discovering interactions among proteins still remains an active research area; only $\approx 20\%$ of all protein-protein interactions in yeast have been currently reported [63]. The high-quality protein communities detected by the AGM can suggest very plausible candidates which biologists can investigate for undiscovered protein-protein interactions [48, 52].

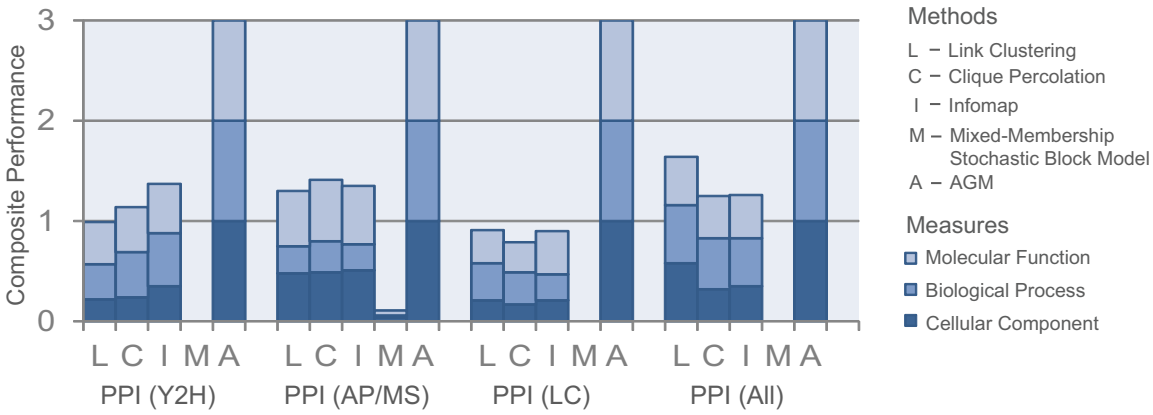


Figure S25: The composite performance of the algorithms on the protein-protein interaction networks. The AGM gives overall best performance by a large margin.

S8 Experiments: Networks in Ahn et al.

Finally, we also evaluate the performance of the AGM under exactly the same conditions as used in the original Link Clustering paper by Ahn et al. [1]. Ahn et al. [1] kindly shared with us the exact networks, metadata and the code. Ahn et al. provide objective evaluations of community detection methods with data-driven measures. We replicate the experiment in [1] with the same data sets and the same evaluation methodology.

S8.1 Dataset description

The seven networks used in [1] were kindly made available to us. We thank Sune Lehmann for generously providing data. We consider the 4 PPI networks that were described in Section S7. In addition, we test over the metabolic network of *E. coli* K-12 MG1655 strain (iAF1260), which is regarded as one of the state-of-the-art metabolic network reconstructions [18]. Two metabolites have an edge if they share a cellular reaction. In total, we have five biological networks.

In addition, we also examine other types of networks. We consider the network of famous philosophers constructed based on Wikipedia [1]. If the Wikipedia page of a philosopher has a hyperlink to the Wikipedia page of the other philosopher, then the two philosophers have an edge between them. We use the Word association network from the data sets from the University of South Florida and the University of Kansas [17, 44], which observed which words human subjects associate to given words. As the data set provides weighted, directed graph among words, we convert the graph into undirected and unweighted version [1, 47]. Basic statistics of the networks is in Table S2. Further details are in [1].

S8.2 Evaluation metrics

We adopt the 4 data-driven measures defined in [1]:

- The community coverage is the fraction of the nodes that belong to at least one detected community.
- The overlap coverage is the average value of the number of communities a node belongs to. If the method detects many communities that share large overlaps, then the overlap coverage will be high.
- The community quality assumes that the similarity of the two nodes $\mu(i, j)$ is available for any pair of nodes i and j. Given the similarities, the community quality is the average similarity between all pairs of nodes that share a community, divided by the average similarity between all pairs of nodes [1].
- The overlap quality requires for each node i the information $W(i)$ which is related to the number of true communities that i belongs to. On the protein-protein interaction networks, for example, a protein annotated with many GO terms is expected to belong many protein communities. On the word association network, words with many definitions are likely to belong to many communities of words. The overlap quality is

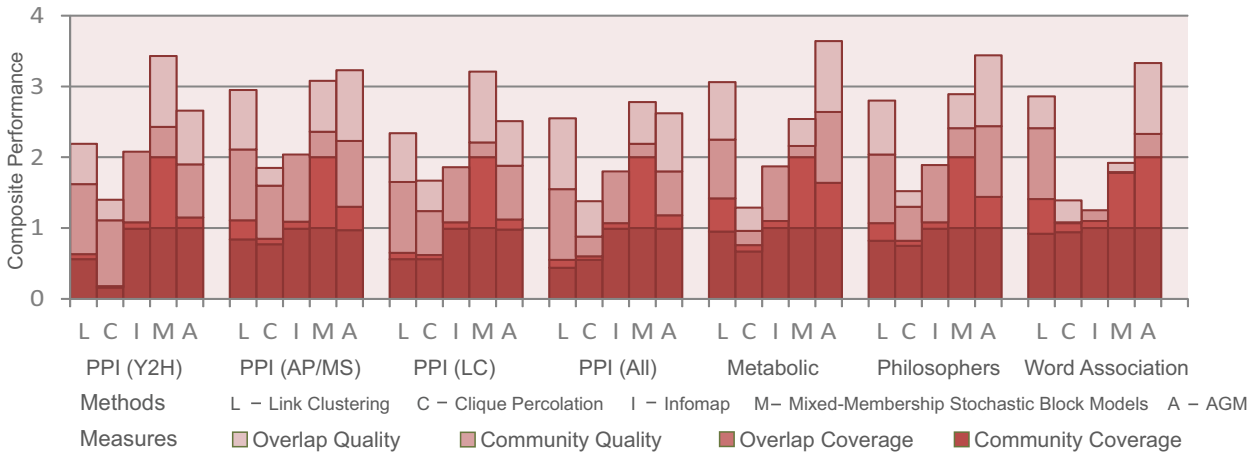


Figure S26: The data-driven benchmark presented in Y.Y. Ahn et al. [1]. Community coverage, Overlap coverage, Community quality, and Overlap quality measure the quality of the communities detected by the algorithms. The AGM gives overall best performance.

the mutual information between $W(i)$ and the number of detected communities that i belongs to.

For each network, we apply the AGM, Link clustering, Clique percolation, Infomap and Mixed-membership stochastic block model. For evaluation we use exactly the same meta data and the same parameters as in [1].

S8.3 Results

Similar to the previous experiments, we compute the composite performance by normalizing the scores the same way as we did in the experiments with ground-truth communities. Figure S26 shows the composite performance of the four methods. The AGM achieves best composite performance in the 3 networks (PPI (Y2H), PPI (LC) and Philosophers), Link clustering performs slightly better in the Word association and the metabolic network, and MMSB is the best in the PPI (Y2H) and PPI (All) networks. On average, the AGM achieves a composite performance score of 3.06, outperforming Link clustering (2.67) by 14%, Clique percolation (1.49) by 104%, Infomap (1.82) by 67% and MMSB (2.84) by 8%. Thus, AGM gives overall best performance on this diverse set of networks and evaluation metrics.

S9 Overlapping communities give rise to core-periphery network structure

In the main text, we showed how the core-periphery structure [9, 29, 54] arises in many different types of networks. In this section, we provide further evidence for our explanations by showing that same results hold for many other networks as well.

S9.1 Community overlaps lead to global core-periphery structure

We being by describing the rest of the networks that we consider for these experiments (Table S2):

- Three social networks. LiveJournal online social network is what we used in the main text, MSN Finland is the MSN network of users in Finland, and the LinkedIn network is the snapshot of the LinkedIn social network when it had 254,151 users.
- We consider the Amazon product network as we did in the main text.
- Three web graphs where nodes are web pages and edges mean hyperlinks [38]. Web-Stanford is a network of the web pages from Stanford University, Web-NotreDame is of the web pages from University of Notre Dame, and Web-BerkStan is of the pages from Stanford University and University of California Berkeley.
- 4 protein-protein interaction networks described in Section S7.
- 2 networks of the Florida Bay food web networks[59]: In wet season (Foodweb-wet) and in dry season (Foodweb-dry).

Global core-periphery structure. Given the network $G(V, E)$, we measure the average number of communities m that a node belongs to as a function of the *farness centrality* d of the node. The Farness Centrality [30, 60] d of node u is the average shortest path length from u to all other nodes in the network, *i.e.*, $d = \frac{1}{|V|}\sum_{v \in V} d(u, v)$ where $d(u, v)$ is the shortest path length from u to v.

Figure S27 displays the plots of m and d for all 10 networks. We use the community memberships detected by AGM to determine m. In all networks but the two web graphs, the number of community memberships of a node decreases with the farness centrality of a node, which implies that nodes residing in the center of the network, which have small shortest path distances to other nodes of the network, tend to belong to the highest number of communities. This result shows that our observation that overlapping communities lead to core-periphery structure of large networks (Figure 3b in the main text) generally applies to a wide range of real-world networks.

Emergence of local cores. We also examine the existence of local cores in the networks by measuring the fraction of the largest connected component (LCC) in the induced subgraph of the nodes who belong to at least l communities. Thinking of a network as a valley where peaks correspond to cores and peripheries to lowlands, our methodology is analogous to flooding lowlands and measuring the fraction of the largest island (which was a peak before flooding). High $c(l)$ means that there is a single dominant core (peak), while low $c(l)$ suggests the existence of nontrivial secondary cores.

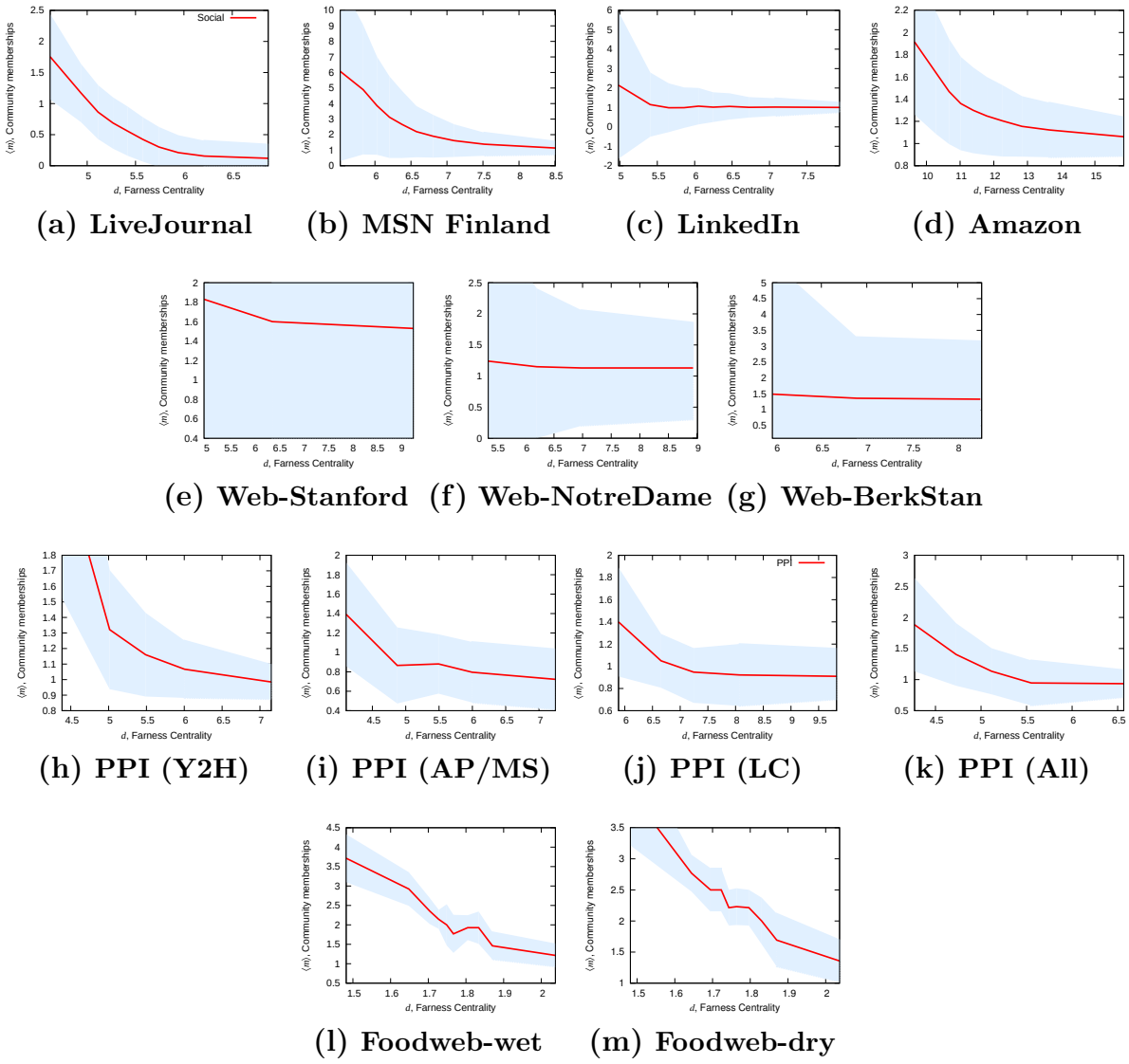
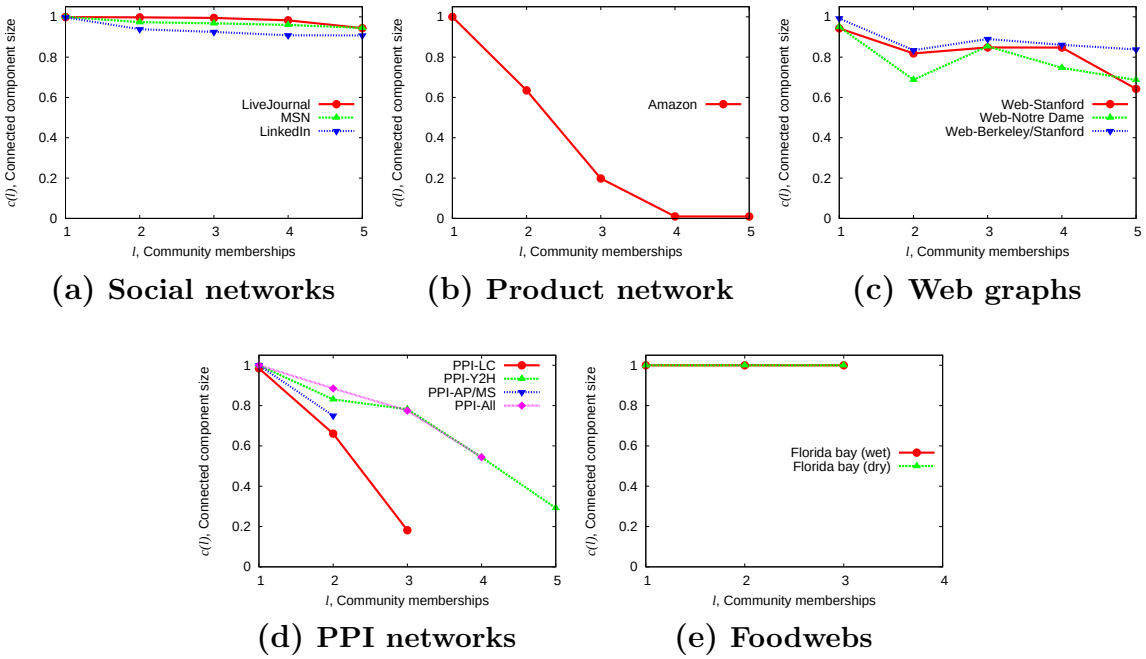


Figure S27: **Overlapping communities lead to global core-periphery network structure.** The average (and the 90-th percentile) of the number of community memberships m of a node as a function of the average shortest path length d to all other nodes of the network. The number of community memberships increases with the centrality of a node. Nodes that reside in the center of the network, and have small shortest path distances to other nodes of the network, tend to belong to the highest number of communities.



(a) Social networks **(b) Product network** **(c) Web graphs**

(d) PPI networks **(e) Foodwebs**

Figure S28: **Largest connected component size on an induced subgraph of nodes belonging to at least l communities.**

Figure S28 displays $c(l)$ for each of the 5 types of networks. As shown in the main text, we observe that the protein-protein interaction networks and the Amazon product network have local cores, while the other types have a global core.

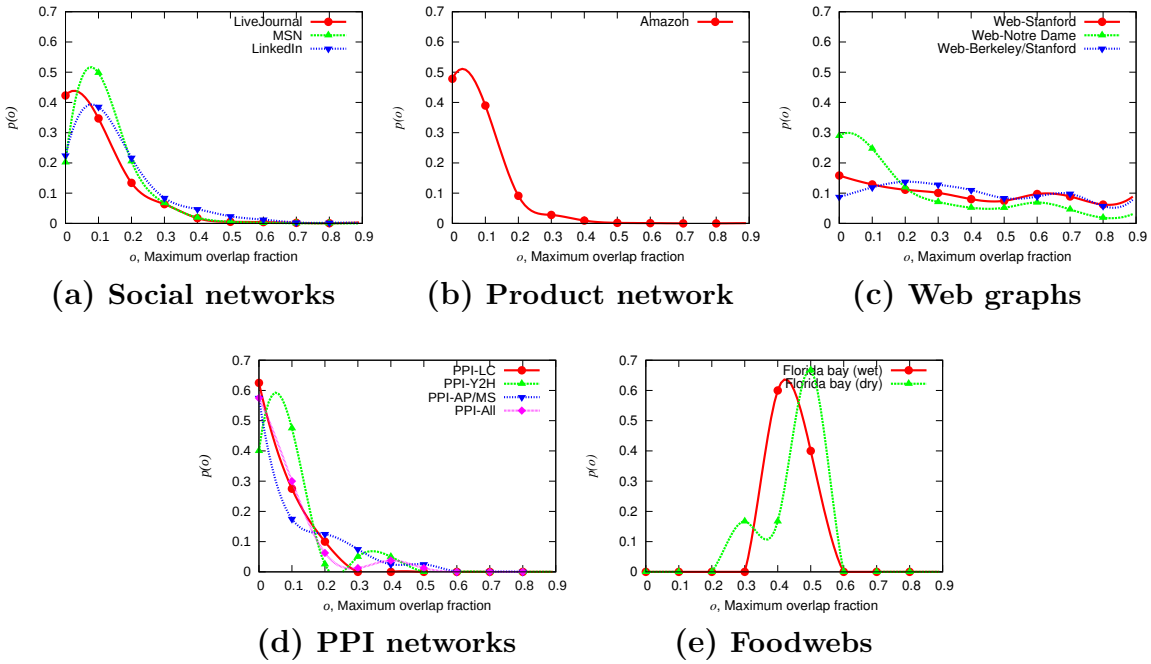
Maximum overlap fraction. Finally, we characterize how much communities overlap with each other in different types of networks. Maximum overlap fraction o_c of a given community c quantifies the fraction of c's members in the largest overlap with any other community.

Figure S29 shows the distribution of o in the 5 types of networks. Communities in the protein-protein interaction networks, social and product co-purchasing network are mainly non-overlapping whereas the communities in the foodweb and the web graph are pervasively overlapping.

S9.2 Comparison to other notions of core-periphery

In order to argue about the core-periphery structure of networks we so far used the fact that communities behave as tiles in the sense that overlap of two communities leads to higher edge density (higher tile thickness). Combining this with the observation that communities overlap most pervasively in the center of the network leads to the conclusion about the global core-periphery structure. However, there are many other methods that identify core-periphery structure in networks and our goal is to quantify the agreement of our methodology and existing methods.

We aim to quantify the agreement between cores we find here and the cores detected by existing methods. In particular, we compare the method invented by Rombach et al. [53].



(a) Social networks **(b) Product network** **(c) Web graphs**

(d) PPI networks **(e) Foodwebs**

Figure S29: **Distribution of maximum overlap fraction.**

Since AGM is proposed for community detection rather than core detection, our goal here is to measure the correspondence between the core determined by [53] and the core (*i.e.*, high membership nodes) detected by the AGM.

Rombach et al. computes a real-valued "core score" $CS(i)$ for each node i which specifies how likely i belongs to a core. In our experiments we used the number of community memberships $m(i)$ of node i to indicate whether i belongs to the core or not. Since $m(i)$ and $CS(i)$ are scores rather than binary indicators, we aim to measure the Pearson correlation coefficient [13] between $m(i)$ and $CS(i)$.

For these experiments we consider two networks that were also considered in Rombach et al. [53]. First, we use the Zachary's karate club network [64]. And second, we also consider the London underground network between the metro stations. Since the London underground network is a weighted network, we build an unweighted network for AGM by connecting two nodes when the edge weight is larger than 2. In Table S3, we observe the correlation coefficient for the Zachary's karate club network is 0.774, and 0.408 for the London underground network. In the second row of the table, we also compute the p-value for the null hypothesis that there is no positive correlation between the two values. We use Student's t-test to achieve this. As p-values are far lower than the standard 0.05, we confirm that the cores that we find by AGM (*i.e.*, the high membership nodes) correspond well to the cores found by the state-of-the-art methods. The level of correlation is lower in the London underground network, which can be explained by the fact that some information is lost when converting the weighted network to an unweighted network.

Dataset	Zachary	London underground
Correlation coefficient	0.774	0.408
p-value	4.01×10^{-8}	6.12×10^{-4}

Table S3: **Comparison with the cores detected by Rombach et al. [53]** The Pearson's correlation coefficient between the core score computed by Rombach et al. [53] and the number of communities the node belongs to as determined by AGM. p-values are also computed for the null hypothesis using the Student's t-test. High correlation coefficient implies that high membership nodes under the AGM are more likely to belong the network core as detected by Rombach et al.

References

[1] Y.-Y. Ahn, J. P. Bagrow, and S. Lehmann. Link communities reveal multi-scale complexity in networks. *Nature*, 466:761–764, Oct. 2010.

[2] E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing. Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9:1981–2014, 2007.

[3] A. Arenas, L. Danon, A. Díaz-Guilera, P. Gleiser, and R. Guimerà. Community analysis in social networks. *The European Physical Journal B*, 38(2):373–380, 2004.

[4] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nature Genetics*, 25:25–29, 2000.

[5] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: membership, growth, and evolution. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 44–54, 2006.

[6] B. Ball, B. Karrer, and M. E. J. Newman. Efficient and principled method for detecting communities in networks. *Physical Review E*, 84:036103, 2011.

[7] G. F. Berriz, J. E. Beaver, C. Cenik, M. Tasan, and F. P. Roth. Next generation software for functional trend analysis. *Bioinformatics*, 25(22):3043–3044, 2009.

[8] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

[9] S. P. Borgatti and M. G. Everett. Models of core/periphery structures. *Social Networks*, 21:375 – 395, 1999.

[10] E. Boyle, S. Weng, J. Gollub, H. Jin, D. Botstein, J. Cherry, and G. Sherlock. GO::TermFinder - open source software for accessing Gene Ontology information and

finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics*, 20(18):3710–3715, 2004.

[11] R. L. Breiger. The duality of persons and groups. *Social Forces*, 53(2):181–190, 1974.

[12] G. Casella and R. L. Berger. *Statistical Inference*. Duxbury Press, 2001.

[13] S. Chatterjee and A. Hadi. *Regression Analysis by Example*. Wiley Series in Probability and Statistics. Wiley, 2006.

[14] A. Clauset, C. Moore, and M. Newman. Hierarchical structure and the prediction of missing links in networks. *Nature*, 453(7191):98–101, 2008.

[15] S. R. Collins, P. Kemmeren, X.-C. Zhao, J. F. Greenblatt, F. Spencer, F. C. P. Holstege, J. S. Weissman, and N. J. Krogan. Toward a comprehensive atlas of the physical interactome of saccharomyces cerevisiae. *Molecular & Cellular Proteomics*, 6(3):439–450, March 2007.

[16] L. Danon, J. Duch, A. Diaz-Guilera, and A. Arenas. Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment*, 29(09):P09008, 2005.

[17] T. S. Evans and R. Lambiotte. Line graphs, link partitions, and overlapping communities. *Physical Review E*, 80:016105, 2009.

[18] A. M. Feist, C. S. Henry, J. L. Reed, M. Krummenacker, A. R. Joyce, P. D. Karp, L. J. Broadbelt, V. Hatzimanikatis, and B. Ø. Palsson. A genome-scale metabolic reconstruction for escherichia coli k-12 mg1655 that accounts for 1260 orfs and thermodynamic information. *Molecular Systems Biology*, 3(121):121, 2007.

[19] S. L. Feld. The focused organization of social ties. *American Journal of Sociology*, 86(5):1015–1035, 1981.

[20] S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75 – 174, 2010.

[21] S. Fortunato and M. Barthélemy. Resolution limit in community detection. *Proceedings of the National Academy of Sciences of the United States of America*, 104(1):36–41, 2007.

[22] A. Gavin, P. Aloy, P. Grandi, R. Krause, M. Boesche, M. Marzioch, C. Rau, L. Jensen, S. Bastuck, B. Dumpelfeld, A. Edelmann, M. Heurtier, V. Hoffman, C. Hoefert, K. Klein, M. Hudak, A. Michon, M. Schelder, M. Schirle, M. Remor, T. Rudi, S. Hooper, A. Bauer, T. Bouwmeester, G. Casari, G. Drewes, G. Neubauer, J. Rick, B. Kuster, P. Bork, R. Russell, and G. Superti-Furga. Proteome survey reveals modularity of the yeast cell machinery. *Nature*, 440(7084):631–636, 2006.

[23] M. Girvan and M. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99(12):7821–7826, 2002.

[24] C. Granell, S. Gómez, and A. Arenas. Hierarchical multiresolution method to overcome the resolution limit in complex networks. *International Journal of Bifurcation and Chaos*, 22(7), 2012.

[25] M. S. Granovetter. The strength of weak ties. *American Journal of Sociology*, 78:1360–1380, 1973.

[26] S. Gregory. Finding overlapping communities in networks by label propagation. *New Journal of Physics*, 12(10):103018, 2010.

[27] S. Gregory. Fuzzy overlapping communities in networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2011(02):P02017, 2011.

[28] P. W. Holland, K. B. Laskey, and S. Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137, 1983.

[29] P. Holme. Core-periphery organization of complex networks. *Physical Review E*, 72:046111, 2005.

[30] P. Holme and G. Ghoshal. Dynamics of networking agents competing for high centrality and low degree. *Phys. Rev. Lett.*, 96:098701, 2006.

[31] B. Karrer and M. Newman. Stochastic blockmodels and community structure in networks. *Physical Review E*, 83:016107, 2010.

[32] N. Krogan, G. Cagney, H. Yu, G. Zhong, X. Guo, A. Ignatchenko, J. Li, S. Pu, N. Datta, A. Tikuisis, T. Punna, J. Peregrín-Alvarez, M. Shales, X. Zhang, M. Davey, M. Robinson, A. Paccanaro, J. Bray, A. Sheung, B. Beattie, D. Richards, V. Canadien, A. Lalev, F. Mena, P. Wong, A. Starostine, M. Canete, J. Vlasblom, S. Wu, C. Orsi, S. Collins, S. Chandran, R. Haw, J. Rilstone, K. Gandi, N. Thompson, G. Musso, P. St Onge, S. Ghanny, M. Lam, G. Butland, A. Altaf-Ui, S. Kanaya, A. Shilatifard, E. O'Shea, J. Weissman, C. Ingles, T. Hughes, J. Parkinson, M. Gerstein, S. Wodak, A. Emili, and J. Greenblatt. Global landscape of protein complexes in the yeast saccharomyces cerevisiae. *Nature*, 440(7084):637–643, 2006.

[33] A. Lancichinetti and S. Fortunato. Community detection algorithms: A comparative analysis. *Physical Review E*, 80(5):056117, 2009.

[34] A. Lancichinetti, F. Radicchi, J. J. Ramasco, and S. Fortunato. Finding statistically significant communities in networks. *PLoS ONE*, 6(4):e18961, 2011.

[35] C. Lee, F. Reid, A. McDaid, and N. Hurley. Detecting highly overlapping community structure by greedy clique expansion. In *Proceedings of the Fourth international workshop on Advances in social network mining and analysis*, 2010.

[36] S. Lehmann, M. Schwartz, and L. K. Hansen. Biclique communities. *Phys. Rev. E*, 78:016108, 2008.

[37] J. Leskovec, L. Adamic, and B. Huberman. The dynamics of viral marketing. *ACM Transactions on the Web*, 1(1), 2007.

[38] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney. Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. *Internet Mathematics*, 6(1):29–123, 2009.

[39] D. Lusseau, K. Schneider, O. Boisseau, P. Haase, E. Slooten, and S. Dawson. The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations. *Behavioral Ecology and Sociobiology*, 54:396–405, 2003.

[40] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[41] M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27:415–444, 2001.

[42] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. In *IMC '07: Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pages 29–42, 2007.

[43] M. Mørup, M. N. Schmidt, and L. K. Hansen. Infinite multiple membership relational modeling for complex networks. *CoRR*, abs/1101.5097, 2011.

[44] D. L. Nelson, C. L. McEvoy, and T. A. Schreiber. The university of south florida word association, rhyme, and word fragment norms, 1998. `http://www.usf.edu/FreeAssociation/`.

[45] M. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences of the United States of America*, 103(23):8577–8582, 2006.

[46] M. Newman and G. Barkema. *Monte Carlo Methods in Statistical Physics*. Oxford University Press, 1999.

[47] G. Palla, I. Derényi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818, 2005.

[48] Y. Park, C. Moore, and J. S. Bader. Dynamic networks from hierarchical bayesian graph clustering. *PLoS ONE*, 5(1):e8118, 2010.

[49] W. W. Powell, D. R. White, K. W. Koput, and J. Owen-Smith. Network dynamics and field evolution: The growth of interorganizational collaboration in the life sciences. *American Journal of Sociology*, 110(4):1132–1205, 2005.

[50] T. Reguly, A. Breitkreutz, L. Boucher, B.-J. Breitkreutz, G. Hon, C. Myers, A. Parsons, H. Friesen, R. Oughtred, A. Tong, C. Stark, Y. Ho, D. Botstein, B. Andrews, C. Boone, O. Troyanskya, T. Ideker, K. Dolinski, N. Batada, and M. Tyers. Comprehensive curation and analysis of global interaction networks in saccharomyces cerevisiae. *Journal of Biology*, 5(4):11, 2006.

[51] J. Reichardt and S. Bornholdt. Detecting fuzzy community structures in complex networks with a potts model. *Physical Review Letter*, 93:218701, Nov 2004.

[52] C. Rivera, R. Vakil, and J. Bader. Nemo: Network module identification in cytoscape. *BMC Bioinformatics*, 11(Suppl 1):S61, 2010.

[53] M. P. Rombach, M. A. Porter, J. H. Fowler, and P. J. Mucha. Core-periphery structure in networks. *SIAM Journal of Applied Mathematics*, 74(1):167–190, 2014.

[54] F. D. Rossa, F. Dercole, and C. Piccardi. Profiling core-periphery network structure by random walkers. *Scientific Reports*, 3, 2013.

[55] M. Rosvall and C. T. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences of the United States of America*, 105:1118–1123, 2008.

[56] H. Shen, X. Cheng, K. Cai, and M.-B. Hu. Detect overlapping and hierarchical community structure in networks. *Physica A: Statistical Mechanics and its Applications*, 388(8):1706 – 1712, 2009.

[57] G. Simmel. *Conflict: the Web of Group Affiliations. Trans. by Kurt H. Wolff and Reinhard Bendix*. Free Press, 1955.

[58] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

[59] R. E. Ulanowicz, C. Bondavalli, and M. S. Egnotovich. Network analysis of trophic dynamics in south florida ecosystem, FY 97: The florida bay ecosystem. *Annual Report to the United States Geological Service Biological Resources Division*, pages 98–123, 1998.

[60] S. Wasserman and K. Faust. *Social Network Analysis*. Cambridge University Press, 1994.

[61] D. Watts and S. Strogatz. Collective dynamics of small-world networks. *Nature*, 393:440–442, 1998.

[62] J. Yang and J. Leskovec. Defining and evaluating network communities based on ground-truth communities. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, pages 745–754, 2012.

[63] H. Yu, P. Braun, M. A. Yldrm, I. Lemmens, K. Venkatesan, J. Sahalie, T. Hirozane-Kishikawa, F. Gebreab, N. Li, N. Simonis, T. Hao, J.-F. Rual, A. Dricot, A. Vazquez, R. R. Murray, C. Simon, L. Tardivo, S. Tam, N. Svrzikapa, C. Fan, A.-S. de Smet, A. Motyl, M. E. Hudson, J. Park, X. Xin, M. E. Cusick, T. Moore, C. Boone, M. Snyder, F. P. Roth, A.-L. Barabasi, J. Tavernier, D. E. Hill, and M. Vidal. High-quality binary protein interaction map of the yeast interactome network. *Science*, 322(5898):104–110, 2008.

[64] W. Zachary. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, 33:452–473, 1977.

[65] E. Zheleva, H. Sharara, and L. Getoor. Co-evolution of social and affiliation networks. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data mining*, pages 1007–1016, 2009.

A Appendix

A.1 Raw performance scores of the experiments with ground-truth communities

Table S4 provides the unnormalized value of evaluation metrics in the experiments in Section S5.

Data	LiveJournal					Friendster					Orkut				
Method	L	C	I	M	A	L	C	I	M	A	L	C	I	M	A
Omega Index	0.39	0.41	0.46	0.39	0.50	0.50	0.52	0.61	0.35	0.56	0.46	0.51	0.55	0.40	0.57
F1 Score	0.35	0.39	0.45	0.63	0.57	0.34	0.39	0.47	0.59	0.56	0.34	0.39	0.46	0.62	0.57
Mutual Information	0.13	0.12	0.17	0.08	0.14	0.17	0.18	0.28	0.10	0.20	0.17	0.18	0.26	0.11	0.21
Number of Communities	0.00	0.27	0.00	0.54	0.34	0.00	0.29	0.00	0.51	0.37	0.00	0.00	0.00	0.54	0.35
Data	Youtube					DBLP					Amazon				
Method	L	C	I	M	A	L	C	I	M	A	L	C	I	M	A
Omega Index	0.48	0.42	0.53	0.42	0.49	0.40	0.37	0.41	0.44	0.50	0.04	0.02	0.08	0.19	0.13
F1 Score	0.33	0.22	0.42	0.57	0.49	0.41	0.38	0.43	0.59	0.53	0.47	0.26	0.61	0.78	0.70
Mutual Information	0.12	0.05	0.17	0.05	0.08	0.24	0.22	0.21	0.10	0.17	0.10	0.05	0.11	0.07	0.10
Number of Communities	0.00	0.37	0.00	0.49	0.46	0.00	0.00	0.00	0.39	0.52	0.00	0.27	0.10	0.44	0.49

Table S4: **Performance of the methods on the networks with ground-truth communities.** Raw scores of the methods in the experiments in Section S5. L: Link clustering, C: Clique percolation, I: Infomap, M: Mixed membership stochastic blockmodels and A: AGM.

A.2 Raw performance scores of the experiments with biological networks

Table S5 gives the unnormalized values of evaluation metrics (*i.e.*p-values) in the experiments in Section S7. Scores are the lower the better.

Data	PPI (Y2H)					PPI (AP/MS)				
Method	L	C	I	M	A	L	C	I	M	A
Cellular Component	0.43	0.39	0.26	1.00	0.02	0.08	0.07	0.07	0.80	0.01
Biological Process	0.24	0.16	0.12	1.00	0.02	0.03	0.02	0.03	0.80	1.5×10^{-6}
Molecular Function	0.18	0.16	0.14	1.00	0.02	0.07	0.05	0.06	0.80	0.01
Data	PPI (LC)					PPI (All)				
Method	L	C	I	M	A	L	C	I	M	A
Cellular Component	0.06	0.10	0.06	1.00	1.9×10^{-6}	0.08	0.24	0.21	1.00	0.01
Biological Process	1.6×10^{-3}	3.6×10^{-3}	0.01	1.00	2.4×10^{-8}	0.06	0.08	0.09	1.00	0.01
Molecular Function	0.04	0.05	0.01	1.00	5.5×10^{-5}	0.07	0.09	0.09	1.00	3.4×10^{-3}

Table S5: **Performance of the methods on the biological networks measured by the GO term finder.** The average p-value of the detected communities computed by the GO term finder in the experiments in Section S7. L: Link clustering, C: Clique percolation, I: Infomap, M: Mixed membership stochastic blockmodels and A: AGM.

A.3 Raw performance scores of the experiments in Ahn et al.

Table S6 shows the unnormalized values of evaluation metrics in the experiments in Section S8.

Data	PPI (Y2H)					PPI (AP/MS)					PPI (LC)				
Method	L	C	I	M	A	L	C	I	M	A	L	C	I	M	A
Community Coverage	0.56	0.16	0.99	1.00	1.00	0.84	0.77	0.99	1.00	0.97	0.56	0.56	0.99	1.00	0.98
Overlap Coverage	0.73	0.18	0.99	11.00	1.62	2.58	0.82	0.99	9.66	3.21	0.93	0.60	0.99	10.55	1.52
Community Quality	2.30	2.18	2.33	1.00	1.75	2.90	2.16	2.76	1.04	2.70	4.71	2.94	3.67	1.00	3.59
Overlap Quality	0.08	0.04	0.00	0.14	0.11	0.28	0.08	0.00	0.24	0.34	0.15	0.09	0.00	0.21	0.14
Data	PPI (All)					Metabolic					Philosophers				
Method	L	C	I	M	A	L	C	I	M	A	L	C	I	M	A
Community Coverage	0.44	0.55	0.99	1.00	0.99	0.95	0.67	1.00	1.00	1.00	0.82	0.75	0.99	1.00	1.00
Overlap Coverage	1.28	0.59	0.99	12.07	2.31	4.66	0.88	1.00	9.82	6.25	2.66	0.77	0.99	10.59	4.64
Community Quality	5.43	1.51	3.99	1.01	3.37	5.33	1.25	4.91	1.00	6.39	2.37	1.17	1.98	1.01	2.43
Overlap Quality	0.15	0.08	0.00	0.09	0.12	0.31	0.12	0.00	0.14	0.38	0.46	0.13	0.00	0.29	0.60
Data	Word associations														
Method	L	C	I	M	A										
Community Coverage	0.92	0.94	1.00	1.00	1.00										
Overlap Coverage	5.12	1.41	1.00	8.16	10.48										
Community Quality	86.20	1.18	12.72	1.01	28.06										
Overlap Quality	0.09	0.06	0.00	0.03	0.20										

Table S6: **Performance of the methods in the experiments of Ahn et al.** The unnormalized scores in the experiments of Ahn et al. [1] in Section S8. L: Link clustering, C: Clique percolation, I: Infomap, M: Mixed membership stochastic blockmodels and A: AGM.