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

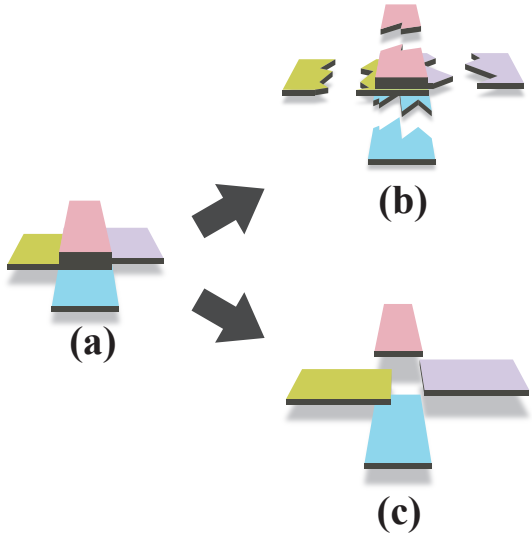
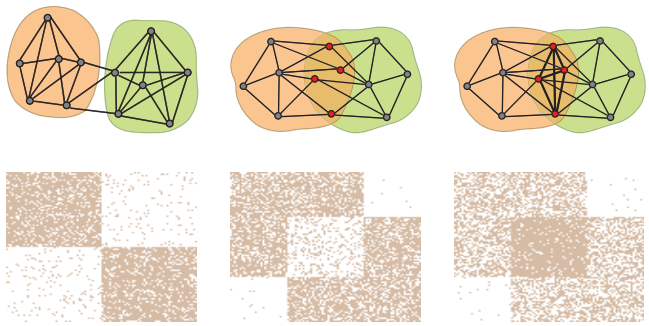


Fig. 1: **Communities as tiles. (a)** Communities in networks behave as overlapping tiles. **(b)** Many methods view communities as clusters with a homogeneous edge density and thus they may break the tiles. **(c)** Our AGM methodology successfully decomposes the network into different tiles (communities).

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)

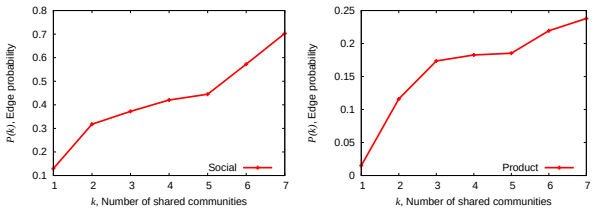
Fig. 2: **Three structural definitions of network communities.** Networks (top) and corresponding adjacency matrices (bottom), where rows/columns denote nodes and dots denote edges: **(a)** two non-overlapping communities; **(b)** two overlapping communities where the overlap is less connected than the 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.

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



(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.

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 we define ground-truth communities based on common *functions* or *roles* around which networks organize into communities [18]. Ground-truth communities are *not* defined based on some observed attribute or property that the nodes share (for example, age, religion, or eye color in a case of a social network) [6]. 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

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

that have one ground-truth community in common and keeps increasing all the way to 0.7 as nodes 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.

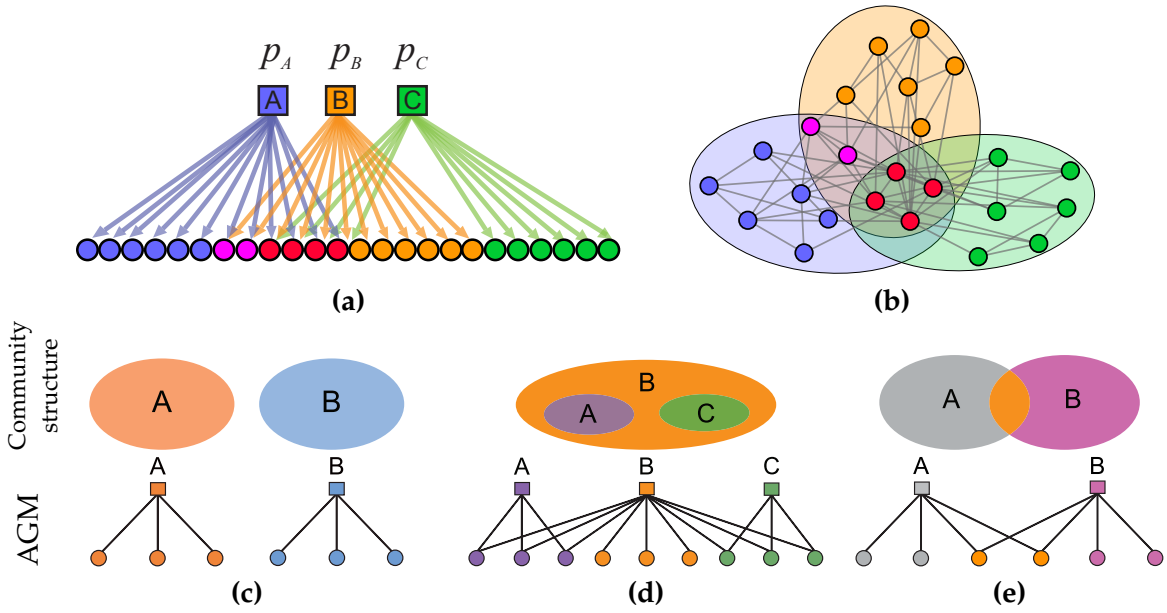


Fig. 4: **Community-Affiliation Graph Model (AGM) [19]. (a)** Squares represent communities and circles represent the nodes of a network. Edges represent node community memberships. For each community A that two nodes share they create a link independently with probability p_A. The probability that a 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

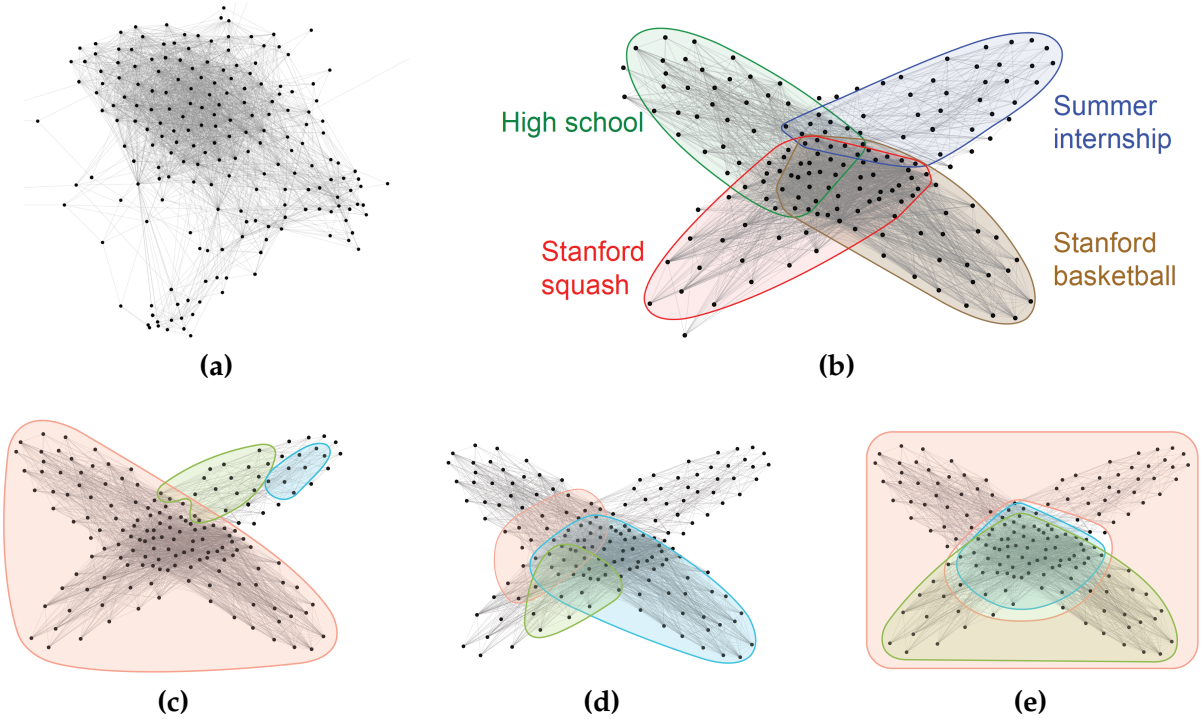


Fig. 5: **An example on a Facebook friendship network of a particular user. (a)** Facebook friendship network of a single user. **(b)** The same network but with communities explicitly labeled by the user: high-school friends, colleagues at the workplace, and university friends with whom the user plays 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.

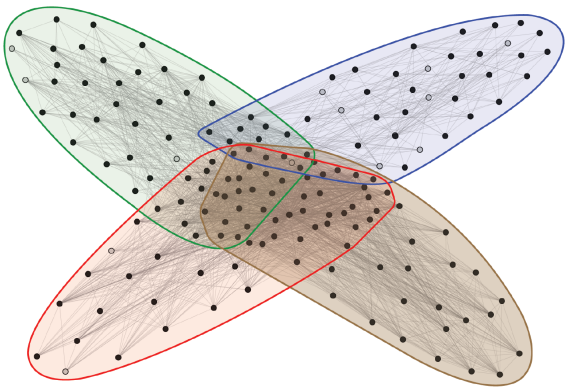
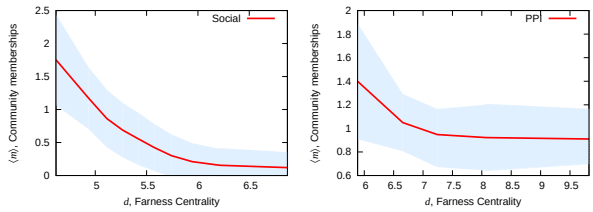


Fig. 6: **AGM on the Facebook network from Figure 5.** AGM successfully decomposes the network into different tiles (communities) and correctly determines community overlaps as well as community memberships for 94% of the nodes.

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 10th 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 $k = 4$.

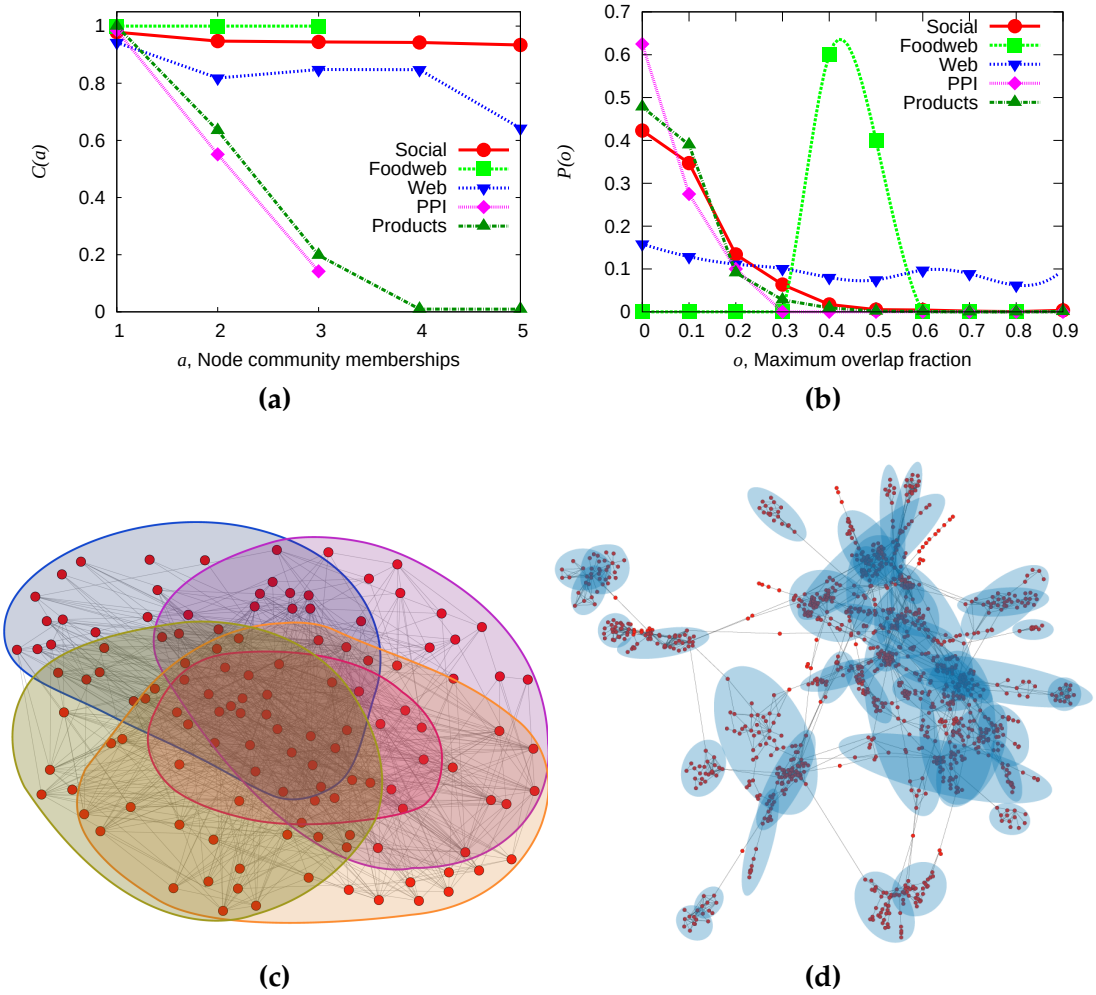
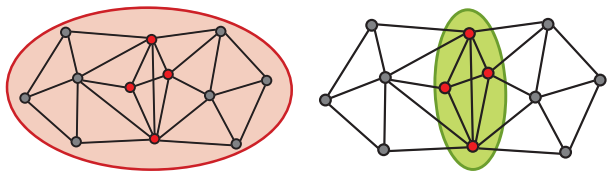


Fig. 8: **Primary and secondary cores in networks. (a)** The fraction of nodes $C(a)$ in the largest connected component of the induced subgraph on the nodes who belong to at least a communities. By thinking of a network as a valley where peaks correspond to cores and peripheries to lowlands, our methodology 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.

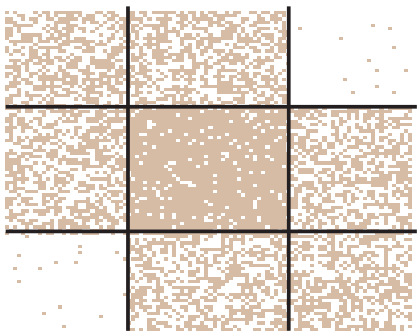


Fig. 10: **The result of a Stochastic Block Model and the Mixed-Membership Stochastic Block model on a network of two communities with dense overlap.** The adjacency matrix of the network in Figure 2c is shown and the bold lines denote the three partitions discovered by the stochastic block models, where 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)*, pp. 745–754, 2012.

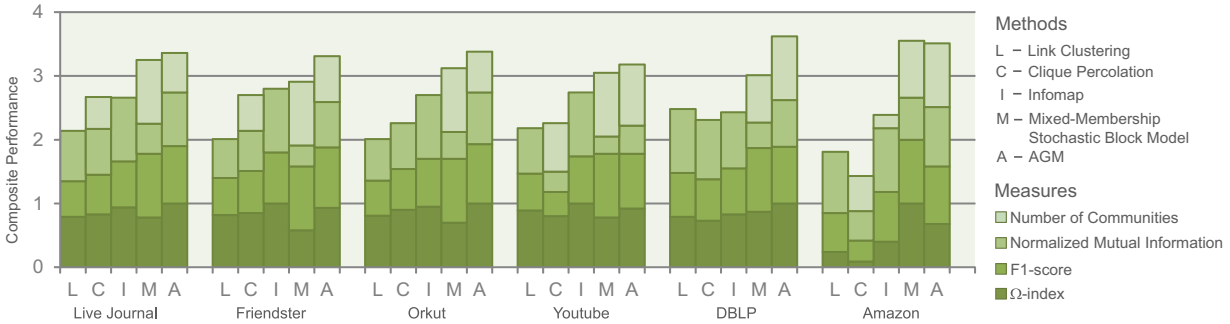
[19] ——, "Community-affiliation graph model for overlapping community detection," in *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, pp. 1170–1175, 2012.

[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.

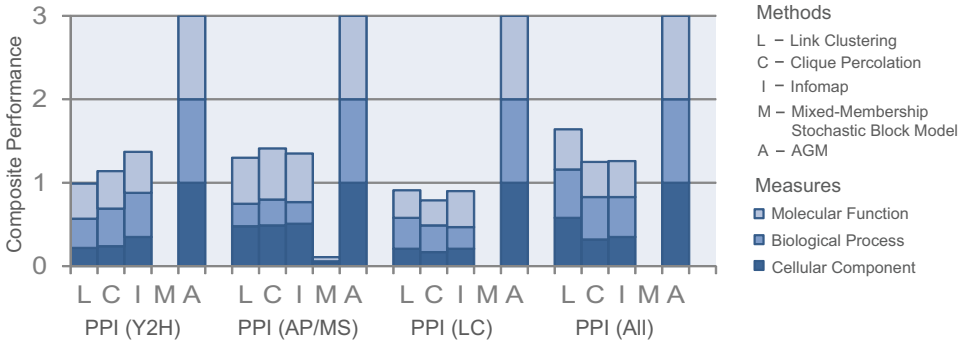
[21] S. P. Borgatti and M. G. Everett, "Models of core/periphery structures," *Social Networks*, vol. 21, pp. 375 – 395, 1999.

[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.



(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.

[24] 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*, 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 *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, pp. 1151–1156, 2013.

[34] J. J. Yang, McAuley and J. Leskovec, "Detecting cohesive and 2-mode communities in directed and undirected networks," in *Proceedings of the ACM International Conference on Web Search and Data Minig (WSDM)*, pp. 323-332, 2014.

[35] J. Yang and J. Leskovec, "Overlapping Communities Explain Core-Periphery Organization of 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 *Proceedings of the ACM International Conference on Web Search and Data Minig (WSDM)*, pp. 587–596, 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.