Community-Affiliation Graph Model for Overlapping Network Community Detection

Jaewon Yang
Stanford University
crucis@stanford.edu

Jure Leskovec
Stanford University
jure@cs.stanford.edu

***Abstract*—One of the main organizing principles in real-world networks is that of *network communities*, where sets of nodes organize into densely linked clusters. Communities in networks often overlap as nodes can belong to multiple communities at once. Identifying such overlapping communities is crucial for the understanding the structure as well as the function of real-world networks.**

Even though community structure in networks has been widely studied in the past, practically all research makes an implicit assumption that overlaps between communities are less densely connected than the non-overlapping parts themselves. Here we validate this assumption on 6 large scale social, collaboration and information networks where nodes *explicitly* state their community memberships. By examining such ground-truth communities we find that the community overlaps are more densely connected than the non-overlapping parts, which is in sharp contrast to the conventional wisdom that community overlaps are more sparsely connected than the communities themselves.

Practially all existing community detection methods fail to detect communities with dense overlaps. We propose Community-Affiliation Graph Model, a model-based community detection method that builds on bipartite node-community affiliation networks. Our method successfully captures overlapping, non-overlapping as well as hierarchically nested communities, and identifies relevant communities more accurately than the state-of-the-art methods in networks ranging from biological to social and information networks.

I. Introduction

Nodes in networks organize into densely linked groups that are commonly referred to as *network communities*, clusters or modules [8], [27]. There are many reasons why networks organize into communities. For example, in social networks communities emerge since society organizes into groups, families, friendship circles, villages and associations [6], [28]. In the graph of the World Wide Web topically related pages link more densely among themselves and communities naturally emerge [7]. And in biological networks communities emerge since proteins belonging to a common functional module are more likely to interact with each other [9], [13].

Communities in networks are thought of as groups of nodes that share a common functional property or role, and the goal of network community detection is to identify such sets of functionally related nodes from the unlabeled network alone [8]. The understanding and models of network communities has evolved over time [8], [17], [30]. Early works on network community detection were heavily influenced by the research on the strength of weak ties [11]. This lead researchers to think of networks as consisting of dense clusters that are linked by a small number of long-range ties (Figure 1(a)) [10]. Graph partitioning [27], modularity [21] as well as betweenness centrality [10] based methods all assume such view of network communities and thus search for edges that can be cut in order to separate the clusters.

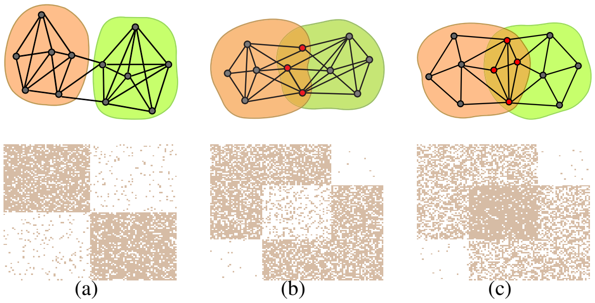


Figure 1. Conventional view of (a) two non-overlapping and (b) two overlapping communities. Notice that the nodes in the overlap are less connected. (c) Our findings suggest densely connected community overlaps. Top: network; Bottom: corresponding adjacency matrix.

Later it was realized that such definition of network communities does not allow for community overlaps. In many networks a node may belong to multiple communities simultaneously which leads to overlapping community structure [1], [2], [23]. The overlapping nature of communities can lead to communities that have more external than internal connections. To deal with this community detection algorithms based on identifying overlapping cliques [23], articulation points, as well as hierarchical clustering of the edges [1] have been proposed. However, practically all present overlapping community detection approaches have a hidden underlying assumption that was left unnoticed. In particular, present overlapping community detection methods assume that community overlaps are *less densely* connected than non-overlapping parts of communities (Figure 1(b)). In other words, this assumption means that the *more* communities a pair of nodes shares, the *less* likely it is they are connected. One possible reason that this assumption went unnoticed and untested could simply be due to the challenges of evaluating community detection — the lack of reliable ground-truth makes the evaluation extremely difficult.

Present work: Empirical observations. Here we validate the above assumption by studying the connectivity structure of *ground-truth communities* [30]. Recently we identified a set of 6 different large social, collaboration, and information networks where we can reliably define the notion of *ground-truth communities* [30]. Networks we study come from a number of different domains and research areas. In all these networks nodes explicitly state their ground-truth community memberships [30]. The availability of reliable ground-truth communities has two important consequences. It allows us to empirically study the structure of true communities and validate present assumptions. Moreover, the ground-truth also allow us to move from qualitative to *quantitative* evaluation of network community detection methods [30].

In this paper we study the overlaps of ground-truth communities and discover that the probability of nodes sharing an edge *increases* as a function of the number of communities they have in common. We find an *increasing* relationship between the number of shared communities of a pair of nodes and the probability of them being connected by an edge. A direct consequence of this is that parts of the network where communities overlap tend to be *more densely* connected than the non-overlapping parts of communities (Figure 1(c)). This observation stands in sharp contrast to present structural definitions of network communities and also means that present methods [1], [2], [23] are *not able* to correctly identify such community overlaps. Present community detection algorithms would either mistakenly identify the overlap as a separate cluster or merge two overlapping communities into a single cluster.

Present work: Model-based Community detection. We then proceed and ask the following question: What underlying process causes community overlaps to be denser than the communities themselves? To answer this question, we build on models of affiliation networks [4], [15] and develop the *Community-Affiliation Graph Model* (AGM) which reliably reproduces the organization of networks into communities and the overlapping community structure [31]. In our model communities arise due to shared group affiliations [4], [28], [6]. The central idea of generating social networks based on the affiliation network is that links among people stem from common group affiliations [4]. We model the probability of an edge between a pair of nodes as a function of the communities that the two nodes share. Community assignments in our model are probabilistic which allows for flexibility in the structure of community overlaps: The AGM can model overlapping, non-overlapping, as well as hierarchically nested communities in networks.

Based on the AGM we then develop a community detection method that successfully detects overlapping, non-overlapping, as well as nested communities in networks. We achieve this by fitting AGM (*i.e.*, discovering the node-community affiliation graph) to an unlabeled undirected network. Using the Markov Chain Monte Carlo method and convex optimization, we develop a fitting algorithm for identifying node community affiliations. We also present a method that automatically determines the number of communities in a given network.

Experiments on social, collaboration, information and biological networks reveal that AGM discovers overlapping as well as non-overlapping community structure more accurately than present state-of-the-art methods [1], [23], [2], [26]. The success of our approach relies on the flexibility of the AGM, which allows for modeling overlapping, non-overlapping as well as hierarchically nested communities in networks.

In summary, our work has three main contributions:

- The observation that community overlaps are more densely connected than the non-overlapping parts.
- Community-Affiliation Graph Model that explains the emergence of dense community overlaps and accurately models network community structure.
- Model-based community detection method that detects overlapping, non-overlapping, as well as nested communities in networks.

II. Networks with Ground-truth Communities

We examine a collection of 6 large social, collaboration and information networks where nodes explicitly state their community memberships [30]. Members of these ground-truth communities share properties or attributes, common purpose or function. We did our best to identify networks in which such ground-truth communities can be reliably defined and identified. Table I gives the dataset statistics.

First we consider 4 online social networks: the LiveJournal blogging community [3], the Friendster online network [19], the Orkut social network [19], and the Youtube social network [19]. In each of these networks 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 interests, hobbies, affiliations, and geographical regions. For example, LiveJournal categorizes communities into the following types: culture, entertainment, expression, fandom, life/style, life/support, gaming, sports, student life and technology. For example, there are over 100 communities with ‘Stanford’ in their name, and they range from communities based around different classes, student ethnic communities, departments, activity and interest based groups, varsity teams, etc.

Figure 2 gives the distribution (Complementary CDF) of ground-truth community sizes and the number community memberships of nodes in LiveJournal. First notice a clear power-law distribution of the community size distribution. The exponent of the cumulative distribution is 1.3, which is slightly higher than what has been reported in the past [5]

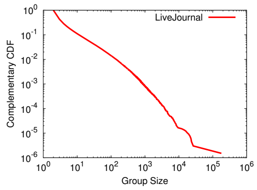


Figure 2. LiveJournal ground-truth communities: (a) Community size distribution, (b) Distribution of the number of communities a node belongs to.

Dataset	N	E	C	S	A
LiveJournal	4.0 M	34.9 M	310 k	40.06	3.09
Friendster	120 M	2,600 M	1.5 M	26.72	0.33
Orkut	3.1 M	120 M	8.5 M	34.86	95.93
Youtube	1.1 M	3.0 M	30 k	9.75	0.26
DBLP	0.43 M	1.3 M	2.5 k	429.79	2.57
Amazon	0.34 M	0.93 M	49 k	99.86	14.83

Table I

DATASET STATISTICS. N: NUMBER OF NODES, E: NUMBER OF EDGES, C: NUMBER OF COMMUNITIES, S: AVERAGE COMMUNITY SIZE, A: COMMUNITY MEMBERSHIPS PER NODE. M DENOTES A MILLION AND k DENOTES ONE THOUSAND.

(based on detected rather than ground-truth communities). On the other hand the distribution of the number of community memberships of a node seems to follow a log-normal distribution of average 3.09. Overall, there are over three hundred thousand explicitly defined communities in the LiveJournal network.

Friendster, Youtube and Orkut online social networks define topic-based communities in the same way as LiveJournal. Users create explicit groups that others then join. Each user can join to zero, one or more such groups. We consider each such group as a ground-truth community. Friendster is the largest network we consider in this study. It contains 120 million nodes, 2.6 billion edges and 1.5 million ground-truth communities.

The second type of network data we consider is the Amazon product co-purchasing network [16], where the notion of community is quite different from that in the social networks. Here 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.

Finally, we also consider the collaboration network of DBLP [3] where nodes represent authors/actors and edges connect nodes that have co-authored a paper. In DBLP we use publication venues as ground-truth communities which serve as proxies for highly overlapping scientific communities around which the network then organizes. In this network communities heavily overlap and tend to be larger than in other networks we consider here (Table I).

The size of the networks we consider here ranges from hundreds of thousands to hundreds of millions of nodes and edges (Table I). The number of ground-truth communities varies from hundreds to millions and there is also a nice range in group sizes and the node membership distribution. Overall, the networks range from those with modular to highly overlapping community structure and represent a wide range of edge densities, numbers of explicit communities, as well as amounts of community overlap (Table I).

We refer the reader to [30] for further discussion on the choice of definition of ground-truth for each network. We were very careful to define ground-truth communities based on common *functions* or *roles* around which networks organize into communities [6], [11]. Note that this is fundamentally different from Ahn et al. [1], who evaluated communities based on attribute similarity of the members. The problem with this approach is that it folds all social dimensions (family, school, interests) around which separate communities form into a single similarity metric. In contrast, we harness explicitly labeled functional groups as labels of ground-truth communities [30]. All the networks we use are complete and publicly available at `http://snap.stanford.edu`.

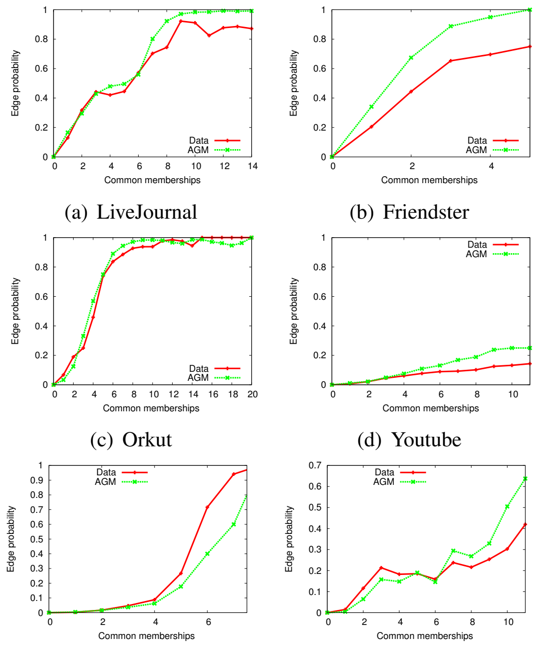
Even though our networks come from very different domains and have very different motivation for formation of communities the results we will present are consistent and robust across all of them.Our work is consistent with the premise that is implicit in all network community literature: members of real communities share some (latent/unobserved) functional property that serves as an organizing principle of the nodes and gives them a distinct structural connectivity pattern in the network. We use these groups around which communities organize to explicitly define ground-truth [30].

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 a particular 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).

III. EMPIRICAL OBSERVATIONS

The availability of reliable ground-truth communities [30] allows us to empirically study the structure of communities and community overlaps. Based on empirical findings, we will then develop a new method for detecting overlapping communities. We study the structure of community overlaps by asking what is the probability that a pair of nodes being connected if they share k common community memberships, *i.e.*, the nodes belong to the overlap of same k communities. Figure 3 plots this probability for all six datasets.

We discover an increasing relationship for all datasets. This means that, the *more* communities a pair of nodes



(a) LiveJournal

(b) Friendster

(c) Orkut

(d) Youtube

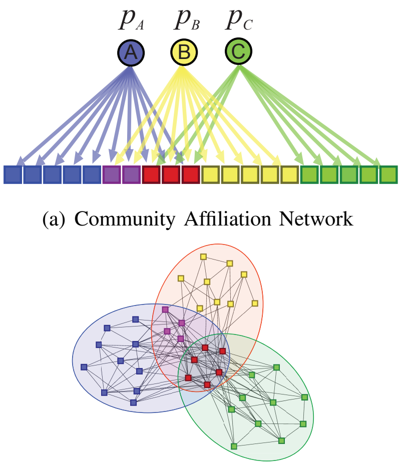
(e) DBLP

(f) Amazon

Figure 3. Edge probability between two nodes given the number of communities that two nodes share. We observe that the edge probability is an increasing function of the number of common communities in all the networks.

has in common, the *higher* the probability of them being connected. In LiveJournal, for example, if a pair of nodes has 8 groups in common, the probability of friendship is nearly 80%. To appreciate how strong the effect of shared communities is on the edge probability, 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 from 10^{-5} to 10^{-1}. That is by 4 orders of magnitude! 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.

Discussion. The above result is very intuitive. While nodes belong to multiple communities (people have friends, families and co-workers), links often exist as a result of one dominant reason (people are in the same family, work together, or share common hobbies and interests). Thus, the more communities people have in common, the more opportunities there are to create links. So, people sharing multiple interests have a higher chance of becoming friends [**?**], researchers with many common interests are more likely to work together [24], and proteins belonging to multiple common functional modules are more likely to interact [9], [13]. Communities thus serve as organizing principles of nodes in social networks and are created based on shared affiliation, role, activity, social circle, interest or function.



(a) Community Affiliation Network

(b) Overlap

Figure 4. (a) Bipartite community affiliation graph. Circles: communities, Squares: nodes in (b) (one square is shown for two squares in (b)). Edges indicate node community memberships. (b) Network generated by AGM.

Our finding suggests communities overlap as illustrated in Figure 1(c). Since the probability of an edge *increases* as a function of the number of shared communities this means that nodes in the overlap of two (or more) communities are more likely to be connected. This view of network formation is consistent with works that predate the "strength of weak ties" literature. In particular, dense community overlaps are consistent with the works of Simmel [28] on the web of affiliations, and Feld [6] on the 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.

We also point the contrast between our finding and the currently predominant view of network communities. Current understanding of network communities is based on two fundamental social network theories: triadic closure and "strength of weak ties" [11] which leads to the picture of network communities as illustrated in Figure 1(a). It also suggests that homophily in networks operates in small pockets where nodes gather in dense non-overlapping clusters (Fig. 1(a)). Moreover, in some networks communities tend to overlap by nodes belonging to multiple communities at once (and thus residing in the "overlap") [23]. Applying the conventional view in this case leads to the structure of community overlaps as illustrated in Figure 1(b): Community overlaps are *less densely* connected than the groups themselves. Our results show the contrary is true.

Last, as a consequence this also means that present overlapping community detection methods [1], [2], [23] are not able to correctly identify such overlaps. They would either mistakenly identify the overlap as a separate cluster or merge two overlapping communities into a single cluster.

IV. Community-Affiliation Graph Model

We proceed by formulating a simple conceptual model of networks that naturally leads to densely overlapping communities. We then design a model fitting procedure that detects communities from a given unlabeled network.

We present the *Community-Affiliation Graph Model* (AGM), a probabilistic generative model for graphs that reliably reproduces the organization of networks into overlapping communities. Our model is based on two main ingredients. The first ingredient is based on Breiger's foundational work [4] which recognized that communities arise due to shared group affiliations [4], [28], [6]. We represent node community memberships with a bipartite affiliation network that links nodes of the social network to communities that they belong to.

The second ingredient of our model is based on the fact that people belong to multiple communities (people have friends, families and co-workers) but the links between them often exist as a result of one dominant reason. We can model this by having each community also carry a single parameter that captures the probability that nodes belonging to that community to share a link. This means every community that a pair of nodes shares gets an independent chance of connecting the nodes. Thus, naturally, the more communities a pair of nodes shares, the higher the probability of linking.

Figure 4(a) 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, the nodes on the top represent communities, and the edges indicate node community memberships. We denote the bipartite affiliation network as $B(V, C, M)$, where V the set of nodes of the underlying network G, C the set of communities, and M the edge set.

Now, given the affiliation network $B(V, C, M)$, we want to generate a social network $G(V, E)$. To achieve this we need to specify the process that generates the edges E of G given the affiliation network 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 are connected by more than one community, the duplicate edges are 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 $(u, c) \in M$ means that node $u \in V$ belongs to community $c \in C$. Let also $\{p_c\}$ be a set of probabilities for all $c \in C$. Given $B(V, C, M)$ and $\{p_c\}$, the Community-Affiliation Graph Model generates a graph $G(V, E)$ by creating edge (u, v) between a pair of nodes $u, v \in V$ 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 ($C_{uv} = \{c|(u, c), (v, c) \in M\}$).

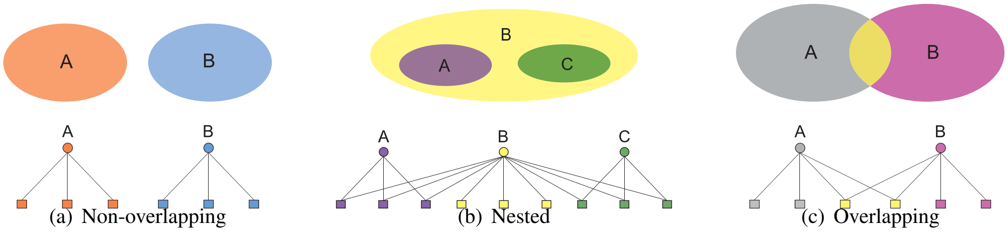
Note that this simple process already ensures that pairs of nodes that belong to multiple common communities are more likely to link. This is due to the fact that nodes that share multiple community memberships receive multiple chances to create a link. For example, pairs of purple nodes in the overlap of communities A and B in Figure 4(a) get two chances to create an edge. First they can be connected with probability p_A (due to their 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 $1 - (1 - p_A)(1 - p_B)$ which is greater than either of p_A or p_B.

We also point out that the Community-Affiliation Graph Model is very similar to the model of Lattanzi and Sivakumar [15]. However, there are two crucial differences. First, [15] posed a model where edge creation probability *decreases* with community size. AGM relaxes this and allows communities to have arbitrary edge probabilities, in order to flexibly model the community structure of real-world networks. Second while [15] focuses on generating synthetic networks with desirable properties, our work aims to *detect* the community structure by developing an efficient fitting algorithm for AGM.

ε-community. In the formulation of Equation 1, AGM does not allow for the edges between the nodes that do not share any common communities. To allow for edges between nodes that do not share any common communities, we assume an additional community, called the ε-community, which connects *any* pair of nodes with a very small probability ε. We find that setting ε to the background probability of a pair of nodes being connected by an edge ($\varepsilon = 2|E|/|V|(|V| - 1)$) works well in practice. In case of our datasets, $\varepsilon \approx 10^{-8}$.

Flexibility of the AGM. Last, we also point out the flexible nature of the Community-Affiliation Graph Model, which allows for modeling a wide range of network community structures. Figure 5 illustrates the structure of affiliation network for three possible community structures. Figure 5(a) shows an affiliation graph of a network with two non-overlapping communities. (Note the presence of ε-community which allows for edges between communities A and B.) Figure 5(b) shows an example of hierarchical community structure where communities A and C are nested inside community B. Finally, Figure 5(c) illustrates an affiliation network corresponding to a pair of overlapping communities. This means that the flexibility of the affiliation network structure allows the AGM to simultaneously model non-overlapping, hierarchically nested as well as overlapping communities in networks.

In [31] we further evaluate the ability of AGM to generate



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

Figure 5. AGM allows for rich modeling of network communities: (a) non-overlapping, (b) nested, (c) overlapping. In (a) we assume that nodes in two communities connect with small prob. ε (refer to the discussion in the main text).

realistic networks with realistic community structure. Our results show that AGM is able to generate networks with heavy-tailed degree distributions, high clustering as well as realistic overlapping, non-overlapping and hierarchical community structures.

V. Community Detection with Community-Affiliation Graph Model

Now that we defined the AGM model, we explain how to detect network communities using the model. Given an unlabeled undirected 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 $L(B, \{p_c\}) = P(G|B, \{p_c\})$ of the underlying graph G:

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

To solve the above optimization problem we employ coordinate ascent strategy where iterate the following two steps. First, we update $\{p_c\}$ by keeping B fixed. Then we update B while keeping $\{p_c\}$ fixed. To start the process we need to initialize B. We achieve this by generating a binary affiliation graph B on K communities ($K = |C|$) by using the configuration model [20].

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

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

with the constraints $0 \leq p_c \leq 1$. Although this problem is non-convex, we can transform it to a convex optimization problem. We maximize the logarithm of the likelihood and change the variables $e^{-x_k} = 1 - p_k$:

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

And the constraints $0 \leq p_c \leq 1$ become $x_c \geq 0$. This problem is a convex optimization of $\{x_c\}$, which means we can find globally optimal solution by using efficient algorithms such as gradient descent or Newton's method.

Updating B**.** To update B, we use the Metropolis-Hastings [22] algorithm where we stochastically update B using a set of 'transitions'. Given the current community affiliation graph $B(V, C, M)$, we consider three kinds of transitions to generate a new community affiliation graph $B'(V, C, M')$ (Figure 6).

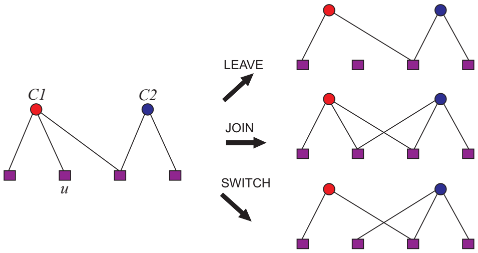


Figure 6. The 3 'transitions' for updating the community affiliation graph B.

- LEAVE simulates node u dropping the membership in community c. We choose a random node-community edge $(u, c) \in M$ and remove it from M (*i.e.*, $M' = M \setminus \{(u, c)\}$).
- JOIN corresponds to node u joining community c. We randomly choose node-community pair $(u, c) \notin M$ and add it to M' (*i.e.*, $M' = M \cup \{(u, c)\}$).
- SWITCH corresponds to node u switching the membership between communities c_1 and c_2. We choose a node-community pair $(u, c_1) \in M$, $(u, c_2) \notin M$ uniformly at random and set $M' = (M \setminus \{(u, c_1)\}) \cup \{(u, c_2)\}$.

Once we have generated new community affiliation B', we accept B' with probability $\max(1, L(B', \{p_c\})/L(B, \{p_c\}))$. In other words, we start the process with some B and then perform a large number of steps, where at each step i we take B_i (we initialize $B_1 = B$) and apply a random 'transition' generating a new affiliation network B'_i. At each step we 'accept' the transition (*i.e.*, we set $B_{i+1} = B'_i$) probabilistically based on the ratio of log-likelihoods. In case the transition is not accepted we do not update B_i (*i.e.*, $B_{i+1} = B_i$).

In our experiments the Markov chain of searching for a good B exhibits relatively quick convergence within $O(|V|^2)$ steps, which makes the complexity of our algorithm effectively quadratic in the number of nodes. Although this is not a rigorous theoretical guarantee, experiments show that our fitting algorithm works well in practice. The algorithm can fit AGM to networks with a few thousand vertices in about an hour.

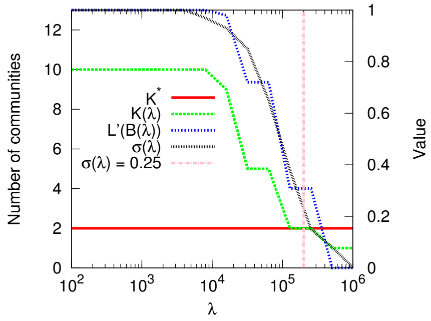


Figure 7. Example of finding the number of communities. Two Y-axes used. Left axis: K^*, $K(\lambda)$. Right axis: $L'(B(\lambda))$, $\sigma(\lambda)$. Refer to the main text for description.

Finding the number of communities. So far we left the question of how to determine the number of communities K unanswered. To address this, we develop a method that automatically estimates the number of communities in a given network where we find the minimal number of communities that still sufficiently explain the edges of the network G.

We begin with fitting a candidate community affiliation graph $B_0(V, C_0, M_0)$ with a very large number of communities ($|C_0| = O(|V|)$). We then aim to force the model to use the minimal number of communities while still accurately modeling the network. We achieve this by placing a l_1 penalty term on parameters $\{p_c\}$ and solve the following modified optimization problem:

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

where λ is a given regularization intensity. The effect of l_1 penalty is that it will force p_c to 0 (*i.e.*, probability of edge between the nodes belonging to that community is 0). This means we can ignore communities in B_0 whose values p_c are 0.

We solve the above problem for various values of λ obtaining $B(\lambda) = B(V, C(\lambda), M(\lambda))$ which consists of the communities with nonzero $\hat{p_c}(\lambda)$. We find that the log-likelihood $L(B(\lambda))$ exhibits a step-like behavior as a function of λ (Figure 7). Let $K(\lambda)$ be the number of communities in $B(\lambda)$, and let λ^* be the value of λ at which $L(B(\lambda))$ experiences the step-like transition. We use $K(\lambda^*)$ as our estimate for the number of communities.

To determine the value of λ^*, we use the following heuristic. We measure $L(B(\lambda))$ for several values of λ (Blue line in Figure 7). Then we fit the sigmoid function $\sigma(\lambda)$ (Black line) to the normalized likelihood, and find λ^* such that $\sigma(\lambda^*) = 0.25$ (Pink vertical line). For the particular case in Figure 7 the method correctly estimates the true number of communities (Red line, K^*). Overall, experiments show that this strategy succeeds in estimating the number of communities more accurately than other methods (Figure 9).

VI. Experiments

We proceed by evaluating the performance of AGM and comparing it to the state-of-the-art community detection methods on a range of networks from a number of different domains and research areas.

Experiments on synthetic networks. Maximum likelihood estimation of AGM is non-convex. To verify that our fitting algorithm does not suffer from poor local optima, we conduct the following experiment on synthetic networks. We generated 100 synthetic networks using AGM. Each network had 200 nodes with randomly chosen B^* as well as $\{p_c^*\}$. Each B^* has 5 communities whose sizes are uniformly sampled from the interval $[40, 80]$ and p_c^* are uniformly sampled from $[0.05, 0.25]$. Then, for each of the 100 networks, we fit AGM from 10 different random initializations to recover $B^*, \{p_c^*\}$. In 97% of cases our fitting algorithm reconstructs B^* with reliable accuracy (F1-score higher than 0.8), and in 50% of cases our algorithm discovers B^* almost perfectly (F1-Score > 0.98). This result suggests that the optimization space of fitting AGM is nicely structured in a sense that the likelihood has several local optima which are almost equivalent to the global optimum.

Experiments on ground-truth communities. We also perform experiments on the 6 networks described in Section II where nodes explicitly state their ground-truth community memberships. Explicitly labeled communities in these networks allow us to measure the 'accuracy' of community detection methods by comparing the level of correspondence between the detected and the explicitly labeled ground-truth communities. Our goal here is very natural. Given an unlabeled undirected network G (with known ground-truth communities C^*) we aim to discover communities $\hat{C}$ such that discovered communities $\hat{C}$ closely match the ground-truth communities C^*.

Experimental setup. We focus the evaluation of community detection methods on their ability to correctly identify overlapping communities. Running community detection algorithms on full networks is not feasible for two reasons. First, *all* the community detection algorithms that we consider here do not scale to networks of millions of nodes. And second, many nodes in our networks do not indicate their ground-truth community memberships, which would complicate the evaluation procedure.

To remedy these problems we use the following evaluation scenario where the goal is to obtain a large set of relatively small subnetworks with overlapping community structure. To obtain one such subnetwork we pick a random node u in the given graph G that belongs to at least two communities. We then take the subnetwork to be the induced subgraph of G consisting of all the nodes that share at least one ground-truth community membership with u. Figure 8 illustrates how a subnetwork (right) is created from $G(V, E)$ (left) based on the red node u. In our experiments we created

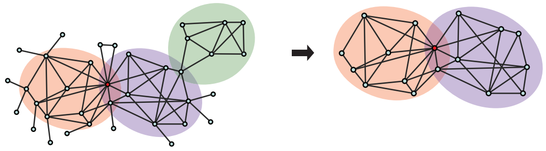


Figure 8. Sampling subnetworks of G.

500 different subnetworks for each of the six datasets.

Baselines. We compare AGM to several state-of-the-art community detection methods. We choose three most prominent community detection methods: Link clustering (LC) [1], Clique Percolation Method (CPM) [23], and the Mixed-Membership Stochastic Block Model (MMSB) [2].

These methods have a number of parameters that need to be set. For CPM, we have to choose the clique size k. We use $k = 5$ since the number of communities discovered by CPM with $k = 5$ best approximates the true number of communities. MMSB also requires the number of communities K as an input parameter. We use the Bayes Information Criterion as described in [2] to choose K. MMSB outputs a stochastic vector for each node representing partial memberships to each of the K communities. To generate "hard" memberships we assign a node to a community if the corresponding stochastic membership is non-zero. For CPM and LC we used the implementation in the Stanford Network Analysis Platform[1], while for MMSB we used publicly-available 'LDA' R package. We note that we also considered Infomap [26], which is the-state-of-the-art non-overlapping community detection method. We omit the results as the performance of the method was not competitive.

Evaluation metrics. Evaluation metrics establish the level of correspondence 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 assess the level of correspondence of $\hat{C}$ to C^*, we use four accuracy metrics:

- **Average F1 score.** To compute the F1 score, we need to determine which $C_i \in C^*$ corresponds to which $\hat{C}_i \in \hat{C}$. We define the F1 score as the average of the F1-score of the best-matching ground-truth community to each detected community, and the F1-score of the best-matching detected community to each ground-truth community: $F1 = \frac{1}{2}(\bar{F}_g + \bar{F}_d)$ where $\bar{F}_g = \frac{1}{|C^*|}\sum_i \max_j F1(C_i, \hat{C}_j)$, $\bar{F}_d = \frac{1}{|\hat{C}|}\sum_i \max_j F1(C_j, \hat{C}_i)$ and $F1(C_i, \hat{C}_j)$ is the harmonic mean of precision and recall of C_i and $\hat{C}_j$.
- **Omega Index [12]** is the accuracy on estimating the number of communities that each pair of nodes shares, *i.e.*, $\frac{1}{|V|^2}\sum_{u,v \in V} \mathbf{1}\{|C_{uv}| = |\hat{C_{uv}}|\}$ where C_{uv} is the set of ground-truth communities that u and v share and $\hat{C_{uv}}$ is the set of detected communities that they share.
- **Normalized Mutual Information** adopts the criterion used in information theory to compare the detected communities and the ground-truth communities. Normalized Mutual Information has been proposed as a performance metric for community detection. Refer to [14] for details.
- **Accuracy in the number of communities** is the relative accuracy between the detected and the true number of communities, $1 - \frac{||C^*|-|\hat{C}||}{|C^*|}$.

Note that for all metrics higher values mean better performance. Maximum value of 1 is obtained when the detected communities exactly correspond to the ground-truth communities.

Results on ground-truth communities. For each community detection method and each dataset we measure the average value of the 4 evaluation metrics over the 500 subnetworks. Then, for each evaluation metric separately we scale the scores of the methods so that the best performing community detection method achieves the score of 1. Finally, we compute the composite performance by summing up the 4 normalized scores. If a method outperforms all the other method in all the scores, then the composite performance of the method is 4.

Figure 9 displays the composite performance of the methods over all six networks. 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 measures in each network. On average, the composite performance of AGM is 3.56, which is 57% higher than that of Link clustering (2.27), 48% higher than that of CPM (2.41), and 10% 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), 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.

EXPERIMENTS ON THE NETWORKS IN AHN ET AL. [1]. Last we also evaluate the performance of AGM by adopting exactly the same data, evaluation metrics and experimental setup as in Ahn et al. [1].

Experimental setup. We use seven different networks.[2] 5 biological networks: 4 protein-protein interaction networks of *Saccharomyces cerevisiae* and the metabolic network of *E. coli* K-12 MG1655 strain (iAF1260); The network of Wikipedia pages of 1,218 famous philosophers; And the Word association network discovered by the University of

[1] http://snap.stanford.edu

[2] We thank Sune Lehmann for generously sharing the data.

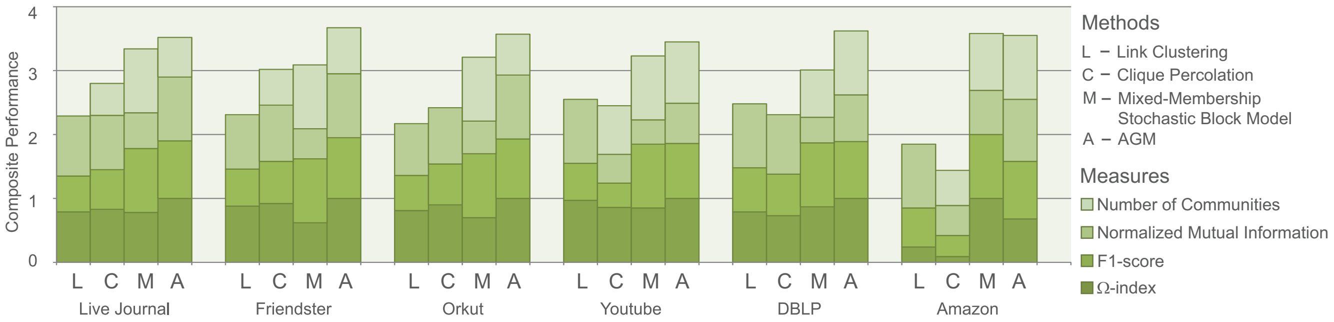


Figure 9. The composite performance of the algorithms on the networks with ground-truth communities.

South Florida and the University of Kansas. For further details about these datasets, refer to [1]. We also adopt the same data-driven measures defined in [1]: Community Coverage, Overlap Coverage, Community Quality, Overlap Quality. We compute these measures using the same metadata as in [1]. The idea behind these scores is that good communities are such where the diversity in the metadata of its members is small. These networks only have about 1,000 nodes each, so we apply the community detection methods to the whole networks. Since these metrics are heavily biased towards methods that find a large number of communities, we fit AGM using the same number of communities as detected by LC.

Results. Following [1] we compute the composite performance by normalizing the scores the same way as we did in the experiments with ground-truth communities. Figure 10 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 (2.74) by 12%, Clique percolation (1.51) by 102%, and MMSB (2.75) by 11%.

VII. Discussion and Conclusion

In this paper we developed a novel community detection method that accurately discovers the overlapping community structure of real-world networks. We identified a set of networks where nodes explicitly state their ground-truth community membership. We then studied the structure of community overlaps in a set of networks with explicitly defined ground-truth communities. We observed that the overlaps of communities are more densely connected than the non-overlapping parts of communities, which is in sharp contrast to assumptions made by present community detection models and methods. Based on this observation, we then developed the *Community-Affiliation Graph Model* (AGM), a conceptual model of network community structure, which reliably captures the overall structure of networks. We then presented an efficient algorithm to fit AGM to a given network whose community structure is unknown. Experiments show that the AGM outperforms the state-of-the-art community detection methods in accurately discovering network communities as well as the overlaps between communities.

We note that the finding that community overlaps are denser than communities themselves nicely extends the notion of homophily in networks [**?**]. The ‘strength of weak ties’ and small-world models [11] lead to the idea that homophily in networks operates in small pockets where inside the pocket nodes link strongly among themselves, and weakly to other pockets. In this respect our work here represents an extension to the understanding of homophily. In a sense we are discovering *pluralistic homophily*[3] where the similarity of one node to another is the number of shared affiliations, not just their similarity along a single dimension. This view of tie formation is consistent with works that predate the “strength of weak ties” literature. In particular, dense community overlaps are consistent with the works of Simmel [28] on the web of affiliations, and Feld [6] on the 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. Thus, network communities should not be thought of as a set of ‘clusters’ but rather as a set of overlapping tiles where the density of the edges increases with the number of tiles that overlap.

Our work has several important implications: First, our analysis sheds light on the organization of complex networks and provides new directions for research on community detection. Second, ground-truth communities offer a reliable way for evaluating community detection methods. And last, the AGM provides a realistic benchmark network on which new community detection algorithms can be developed and evaluated. More generally, a shift in perspective from sparse to dense community overlaps represents a new way of studying networks and provides a unifying framework for

[3] We thank Michael Macy for coining the term.

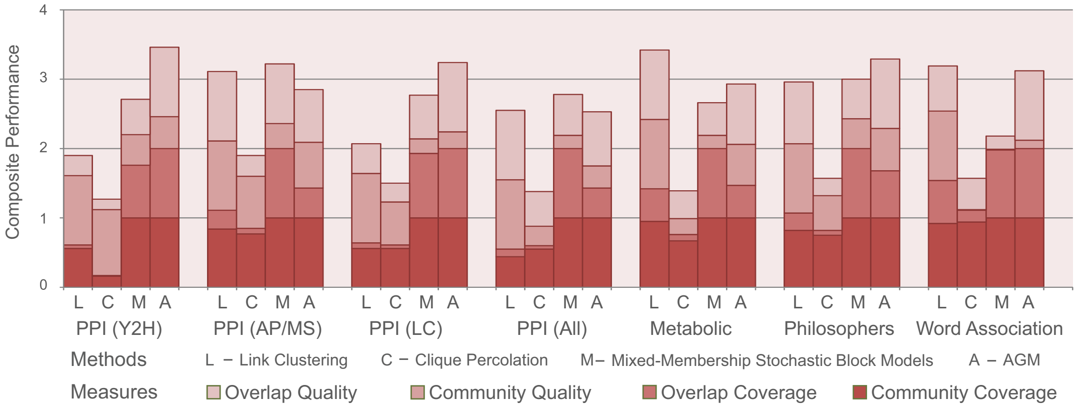


Figure 10. Experiments on the data-driven benchmark by Y.Y. Ahn et al. [1].

network community detection.

Acknowledgements. This research has been supported in part by NSF IIS-1016909, CNS-1010921, CAREER IIS-1149837, IIS-1159679, DARPA XDATA, DARPA GRAPHS, Albert Yu & Mary Bechmann Foundation, Boeing, Allyes, Samsung, Intel, Alfred P. Sloan Fellowship and the Microsoft Faculty Fellowship.

REFERENCES

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

[2] E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing. Mixed membership stochastic blockmodels. *JMLR*, 2007.

[3] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: membership, growth, and evolution. In *KDD '06*, 2006.

[4] R. L. Breiger. The duality of persons and groups. *Social Forces*, 1974.

[5] A. Clauset, M. Newman, and C. Moore. Finding community structure in very large networks. *Phys. Rev. E*, 2004.

[6] S. L. Feld. The focused organization of social ties. *American Journal of Sociology*, 1981.

[7] G. Flake, S. Lawrence, and C. Giles. Efficient identification of web communities. In *KDD '00*, 2000.

[8] S. Fortunato. Community detection in graphs. *Physics Reports*, 2010.

[9] A. Gavin et al. Proteome survey reveals modularity of the yeast cell machinery. *Nature*, 2006.

[10] M. Girvan and M. Newman. Community structure in social and biological networks. *PNAS*, 2002.

[11] M. S. Granovetter. The strength of weak ties. *American Journal of Sociology*, 1973.

[12] S. Gregory. Fuzzy overlapping communities in networks. *J. of Stat. Mech.*, 2011.

[13] N. Krogan et al. Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature*, 2006.

[14] A. Lancichinetti and S. Fortunato. Community detection algorithms: A comparative analysis. *Phys. Rev. E*, 2009.

[15] S. Lattanzi and D. Sivakumar. Affiliation networks. In *STOC '09*, 2009.

[16] J. Leskovec, L. Adamic, and B. Huberman. The dynamics of viral marketing. *ACM TWeb*, 2007.

[17] 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*, 2009.

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

[19] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. In *IMC '07*, 2007.

[20] M. Molloy and B. Reed. A critical point for random graphs with a given degree sequence. *Random Structures and Algorithms*, 1995.

[21] M. Newman. Modularity and community structure in networks. *PNAS*, 2006.

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

[23] G. Palla, I. Derényi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 2005.

[24] 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. *Am. J. of Sociology*, 2005.

[25] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi. Defining and identifying communities in networks. *PNAS*, 2004.

[26] M. Rosvall and C. T. Bergstrom. Maps of random walks on complex networks reveal community structure. *PNAS*, 2008.

[27] S. Schaeffer. Graph clustering. *Computer Science Rev.*, 2007.

[28] G. Simmel. *Conflict and the web of group affiliations*. Simon and Schuster, 1964.

[29] J. Yang and J. Leskovec. Community-Affiliation Graph Model for Overlapping Network Community Detection. Extended version.

[30] J. Yang and J. Leskovec. Defining and evaluating network communities based on ground-truth. In *ICDM '12*, 2012.

[31] J. Yang and J. Leskovec. Structure and Overlaps of Communities in Networks In *SNAKDD '12*, 2012.

[32] E. Zheleva, H. Sharara, and L. Getoor. Co-evolution of social and affiliation networks. In *KDD '09*, 2009.