Jaccard Affiliation Graph (JAG) Model For Explaining Overlapping Community Behaviors

Chen Luo
Rice University
Email: cl67@rice.edu

Anshumali Shrivastava
Rice University
Email: anshumali@rice.edu

Abstract—Understanding overlapping community structures is crucial for network analysis and prediction. AGM (Affiliation Graph Model) is one of the favorite models for explaining the densely overlapped community structures. In this paper, we thoroughly re-investigate the assumptions made by the AGM model on real datasets. We find that the AGM model is not sufficient to explain several empirical behaviors observed in popular real-world networks. To our surprise, all our experimental results can be explained by a parameter-free hypothesis, leading to more straightforward modeling than AGM which has many parameters. Based on these findings, we propose a parameter-free Jaccard-based Affiliation Graph (JAG) model which models the probability of edge as a network specific constant times the Jaccard similarity between community sets associated with the individuals. Our modeling is significantly simpler than AGM, and it eliminates the need of associating a parameter, the probability value, with each community. Furthermore, JAG model naturally explains why (and in fact when) overlapping communities are densely connected. Based on these observations, we propose a new community-driven friendship formation process, which mathematically recovers the JAG model. JAG is the first model that points towards a direct causal relationship between tight connections in the given community with the number of overlapping communities inside it. Thus, *the most effective way to bring a community together is to form more sub-communities within it*. The community detection algorithm based on our modeling demonstrates a significantly simple algorithm with state-of-the-art accuracy on six real-world network datasets compared to the existing link analysis based methods.

I. INTRODUCTION

In a network structure, community means a set of groups of nodes that share common roles or interest, etc. For example, in a protein-protein interaction network, each community in this network shows a common biological property. In social networks, the community can be a family or a friendship circle, etc. In the WWW (World-wide Web), web pages that have the same topic can belong to the same community. Formally, a community is defined to be a set of cohesive nodes that have more connections inside than outside [1], [2], [3].

Community detection algorithms are used for exploring the community structures in networks [2], and are regarded as one of the most important tasks for network analysis [4], [5], [6], [7]. Community detection is of great interest in several real-world applications. For example, the community structure can be used by recommender systems [8]. The community structure of the protein-protein iteration networks help biologist

better understanding the protein structure and properties [9]. In social media analysis, the community structures of users can help us to understand user social behaviors[10].

Prior network community detection methods such as Graph Partition [2], Modularity [11], etc., assume that in a graph, each node belongs to only one community. Fig 1(a) shows the non-overlap community structure in a network. However, in most real-world networks, the communities are often overlapping [12]. For example, in the online social media network Facebook. Each node means a person; each edge represents friendship relationship. The community in a social network can be represented as different friendship circles. It is evident that a person can belong to different friend circles (e.g., College Friends, and High School Friends) at the same time. In a protein-protein network, proteins can have different biology properties, thus can simultaneously belong to several communities [9]. As a result, understanding overlapping community structure of a graph is a crucial task for network analysis.

Because of the significance of the problem, many overlapping community detection methods have been proposed [2], [1]. Early attempts were heavily based on the assumption that the overlapping communities are less densely connected than non-overlap ones, as shown in Fig. 1(b). However, with the availability of large real-data having ground truth labels, it was found that in most real-world social networks, the overlapping communities are more densely connected than non-overlapping ones [13].

Fig. 1(c) shows an example of the overlapping community structure where the overlapped part is densely connected. There were strong empirical observations that probability of edges, in many real networks, keeps increasing with increase in the number of shared community. This observation motivates AGM model: Affiliation Graph Model for overlapping Network Community [13].

AGM assumes that individuals in each community C_i have a *community-specific* probability p_i of forming connections. Furthermore, the probability is independent across different communities. This model naturally provided an explanation for the observation that overlapping communities are densely connected. The more the number of shared communities, the more the probability of the edge formation. Based on AGM model, authors show an alternating sampling and optimization strategy for community detection, in a network. The model has one parameter per community. Thus, the number of parameters

1

(a) Two Non-overlapped Communities   (b) Two overlapped Communities, with sparse link in overlap area   (c) Two overlapped Communities, with dense link in overlap area
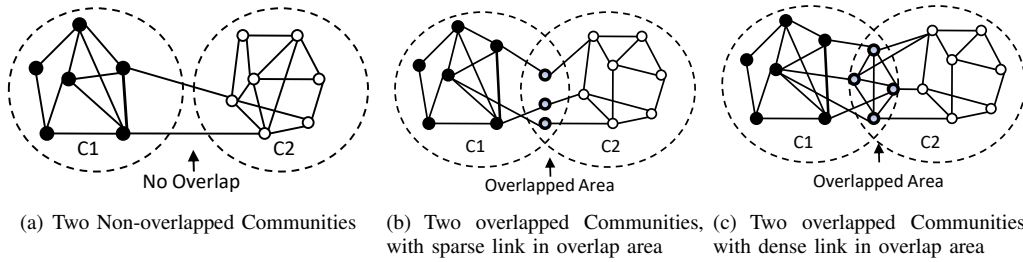
Fig. 1. Network community structures. In this paper, we focus on the community detection problem for densely overlapped structures.

equal to the number of communities present in the network. These parameters are optimized over data in alternating phase for community detection. AGM is currently one of the state-of-the-art algorithms in literature for community detection when only the edge information and no feature information about nodes is available. AGM outperforms several other graph density based parameter free models.

In this paper, we point to a consistent observation (Section III), coming from six real-world networks, that cannot hold true under AGM modeling. Before we go into the actual empirical findings, we illustrate our motivation by a simple example.

A. Motivating Examples:

In a network $G(V, E)$, let v_1. v_2, $v_3 \in G$, be three individuals. v_1 belongs to the community set: $\{C_1, C_2, C_3, C_4, C_5\}$. v_2 belongs to community set: $\{C_1, C_6\}$, and v_3 belongs to the community set $\{C_1, C_7\}$. Intuitively, based on this information, we can expect that v_2 and v_3 are more likely to have an edge than v_1 and v_2 or v_1 and v_3.

However, according to AGM model [13], the edge probability of (v_2, v_3) and (v_1, v_2) (or v_1, v_3) should be exactly the same. This is because they all have only one shared community C_1 between any pair of them. AGM model only considers the identity of shared community between the pair of individuals to decide on edge formation probability. Thus, with AGM, the edge formation probability between individuals has no dependence of what other involvements of individuals are.

Thus, v_1 in our example, even though involved in several other communities, will have the same chance of forming the friendship with v_2 as that of v_3. This behavior does not seem right, as v_2 and v_3 do not have other involvements and therefore are more likely to interact with each other compared to v_3. In this paper, we show that indeed this is the case. We show empiric validation of this phenomenon, on 6 real-world networks, to demonstrate the above argument.

We conduct several validating experiments in section III. These experiments strongly suggest a nearly linear relationship between the edge probability and the Jaccard overlaps of the community sets of two individuals. The same relationship holds at different granularity, and the patterns are consistent across domains. From the empirical findings, it appears that the Jaccard similarity between the community sets of any two individuals is the "sweet" measure, which in itself is sufficient to model the edge probability indicating a parameter-free model.

Strongly motivated by these empirical findings, we propose Parameter-free JAG (Jaccard-based Affiliation Graph) Model for modeling the community structures in social networks. JAG model uses Jaccard similarity of the community sets associated with a pair of individuals as the probability of edge formation between them. This modeling is significantly simpler than AGM, and it eliminates the need of associating a parameter, the probability value, with each community. Furthermore, JAG naturally explains why (and in fact when) overlapping communities are densely connected. Individuals with significant overlap in their communities have high Jaccard Similarity and hence more edge formation probability.

We then show a simple random process of friendship formation based on community structure and random global preferences. Under fairly realistic assumptions, our theory recovers JAG model as the edge formation probability. To the best of our knowledge, this is the first theory that shows how preferences over communities lead to friendship formation. We also validate the consequences of these theoretical claims on all the six real datasets. The new theory and associated math could be of independent interest in itself.

Despite being able to explain more empirical findings, JAG modeling only requires one network specific global parameters compared to AGM which requires one parameter per community. As a consequence, JAG based community detection is significantly cheaper, and accurate, than the one based on AGM. We test JAG model on the task of community detection. JAG model, despite being cheap and parameter-free, consistently outperforms AGM on six different real datasets on three separate performance metrics.

There are a variety of community detection algorithms which use features and other side information for community detection [14]. However, our focus is on understanding communities only from the perspective of connectivity (and links) and not other side information. We will therefore only consider methodologies that exploit network connectivity for community detections, and we do not assume any other information.

II. BACKGROUND AND NOTATIONS

Given an unweighted and undirected network $G(V, E)$, where $V = \{v_1, v_2, ..., v_n\}$ is the set of nodes in G, and $E = \{e_1, e_2, ..., e_m\}$ is the set of edges in G. $n = |V|$ is the number of node, and $m = |E|$ is the number of edges.

2

(a) LiveJournal Network (b) Friendster Network (c) Orkut Network (d) Youtube Network (e) DBLP Network (f) Amazon Network
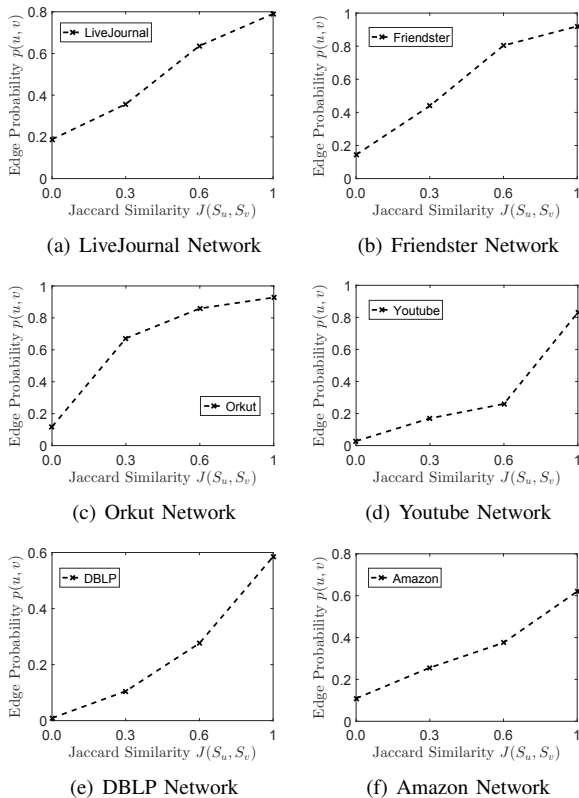
Fig. 2. We fix two large communities in each network dataset, and plot the edge formation probability $p(u,v)$ as a function of the Jaccard Similarity $J(S_u, S_v) = \frac{|S_u \cap S_v|}{|S_u \cup S_v|}$. Our process guarantees that the shared community of all the pairs is same. Thus, if AGM model was to hold true, then the probability of edge formation should be constant in all bins. On the contrary, we observe a roughly linear relationship between the Jaccard similarity and the edge probability irrespective of the dataset used. Thus, we clearly see that AGM model is not good enough to explain this observation.

A. AGM Model

We firstly briefly review the AGM model [13]. AGM model assumes a simple generative process of edge formation based on coin flips. Given any two nodes u and v and let C_{uv} be the set of shared communities between them. For every community $C_k \in C_{uv}$, we flip coins (independently) with community-specific probability p_k. The edge is formed if any of the coin flipped are in favor.

Under AGM modeling, given two nodes u and v, the edge probability between u and v in AGM model is given by:

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

where C_{uv} is the communities common to both u and v. p_k is the edge formation probability of community C_k. The edge formation probability here denotes the probability that two nodes have an edge between them.

Clearly, from the definition, we can see that AGM assumes that each community C_i had a community-specific probability p_i of forming connections. Furthermore, the probability is independent across different communities. The number of parameters required by AGM modeling is therefore equal to the number of communities under consideration.

DATA SET	NODE	EDGE	COMMUNITY
LIVE JOURNAL	4.0 M	34.9M	310 K
FRIENDSTER	120 M	2600 M	1.5 M
ORKUT	3.1 M	120 M	8.5 M
YOUTUBE	1.1 M	3.0 M	30 K
AMAZON	0.34 M	0.93 M	49 K
DBLP	0.43 M	1.3 M	2.5 K

B. Key Notations

Community Set: Given a node u, the community set S_u is defined as the set of communities that node u belongs to. For example, in a Facebook social network, the community set S_u for a person u is the set of his/her friendship circles: High-school Friend Community, Graduate-school Community, Family Community, etc.

Jaccard Similarity: The Jaccard similarity $J(S_u, S_v)$ between two community set S_u and S_v is defined as:

$$J(S_u, S_v) = \frac{|S_u \cap S_v|}{|S_u \cup S_v|},$$

where $|S_u \cap S_v|$ is the number of shared nodes between S_u and S_v, and $|S_u \cup S_v|$ is the number of all the nodes in the union set of S_u and S_v.

III. OBSERVATION EXPERIMENT

We use six publicly available real-world network datasets where we also have ground truth labels for communities of each node. The datasets contain four online social networks: the LiveJournal blogging network, the Friendster online network, the Orkut social network, and the Youtube social network. We also consider an Amazon product co-purchasing network, and the collaboration network: DBLP. The basic statistic of the network datasets is shown in Table I. It is pointed out that all the networks in the data sets are unweighted and undirected static graph. The datasets are publicly available on http://snap.stanford.edu

We first provide the empirical motivation of our proposal. Based on our motivating example, we hypothesize that the edge formation probability $p(u,v)$ between individuals u and v should be some unknown function of the Jaccard similarity between their community set, i.e. $p(u,v) = f(J(S_u, S_v))$ instead of being just dependent on the number of shared communities. In other words, by such modeling, we ensure that the normalization matters. As an individual belonging to many communities will have less attention span on one of his community compared to an individual who only belongs to that community.

A. Validation 1

To validate this hypothesis we performed the following evaluation: Given a network data, we first fix two large communities say C_1 and C_2 in them. Then, we collect all the node pairs which only have these fixed communities in common. Note, the individuals u and v may belong to other

3

(a) Live Journal Network

(b) Friendster Network

(c) Orkut Network

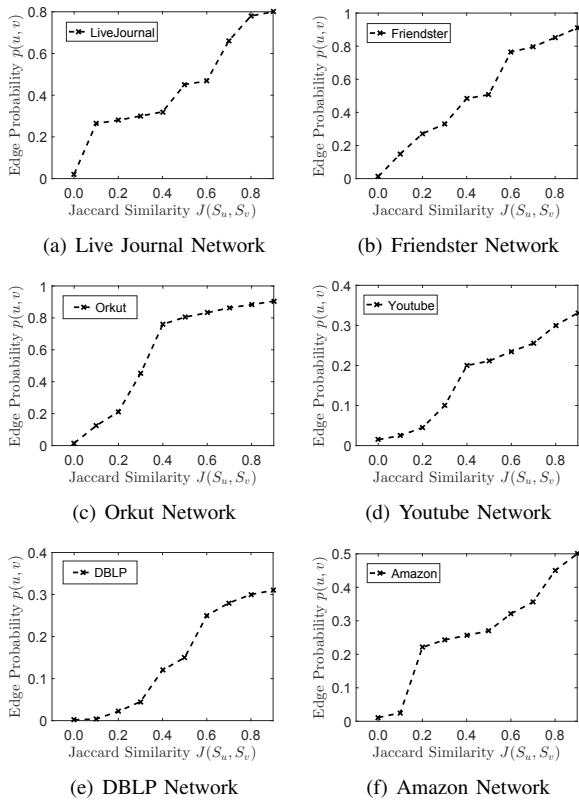(d) Youtube Network

(e) DBLP Network

(f) Amazon Network

Fig. 3. We plot the edge formation probability $p(u,v)$ as a function of the Jaccard Similarity $J(S_u, S_v) = \frac{|S_u \cap S_v|}{|S_u \cup S_v|}$ on full networks. We can clearly see similar trends as shown in Fig.2. With the increasing of the Jaccard similarity, the probability keep increasing. In addition, we can also observe that the relationship between the Jaccard similarity and edge probability can be reasonably approximated by a linear relationship.

communities, but the pair only shares the same *fixed* two communities, i.e. $|S_u \cap S_v| = \{C_1, C_2\}$

We then plot the edge formation probability as a function of the Jaccard Similarity $J(S_u, S_v) = \frac{|S_u \cap S_v|}{|S_u \cup S_v|} = \frac{2}{|S_u \cup S_v|}$. To get such edge probabilities. We sample a pair of nodes satisfying $|S_u \cap S_v| = \{C_1, C_2\}$ by rejection sampling [15]. Then for all those pairs, we bin them into groups based on their Jaccard similarity $J(S_u, S_v)$. In all these bins we compute the probability of an edge, i.e. $p_{edge} = \frac{pairs\ with\ edge}{total\ pairs}$. We ensure that we have sufficient statistics in each bin, by continuing to sample until we get large enough numbers. We also used coarse-grained binning to ensure sufficient statistics.

We plot this probability p_{edge} with Jaccard similarity for all the six datasets (as in Fig. 2). We see a roughly linear relationship between the Jaccard similarity and the edge probability irrespective of the dataset used. This is very fascinating.

First of all, our process guarantees that the shared community of all the pairs is same. Thus, if AGM model was to hold true, then the probability of edge formation should be constant in all bins. On the contrary, we observe a linear trend, i.e. the probability increases with Jaccard similarity of the associated community set. Thus, we clearly see that AGM model cannot explain these observations.

Based on the linear trend, we can hypothesize that our

assumption of $p(u,v) = f(J(S_u, S_v))$ is in fact more like $p(u,v) = \alpha J(S_u, S_v))$, where α is a scalar which is network dependent. Note this modelling just has one parameter.

B. Validation 2

Our previous experiment used constrained pairs, where $|S_u \cap S_v| = \{C_1, C_2\}$. If our assumption was to be true, then we should observe similar behavior even for any pairs. To see that, *we repeated our experiment but this time sampling random pairs instead of forcing $|S_u \cap S_v| = \{C_1, C_2\}$.* We again plot the probability of edge as a function of $J(S_u, S_v)$. This time we have more samples, and so, we can do fine grain binning of $J(S_u, S_v)$ values.

Fig. 3 shows the new plot again for all six data sets. We can clearly see similar trends. With the increasing of the Jaccard similarity, the probability keeps increasing. In addition, we can also observe from the Fig. 3 that, a linear relationship can reasonably approximate the relationship between the Jaccard similarity and edge probability.

IV. JAG: PARAMETER-FREE JACCARD-BASED AFFILIATION GRAPH MODEL

Our observation experiment shows that given two nodes u, and v, and the corresponding community set: S_u and S_v. The probability that node u and v have an edge is roughly linearly correlated with the Jaccard similarity between S_u and S_v. We formalize this by proposing our edge probability as follow:

$$p(u,v) = \alpha J(S_u, S_v). \qquad (2)$$

In above equation, α is a parameter that correspond to the properties of the network. We will discuss α later.

Coming back to our motivating example with v_1, v_2 and v_3 as constructed in the first section. With our model, the probability that v_1 and v_2 are connected is: $p(v_1, v_2) = \alpha/6$, and the probability that v_1 and v_3 are connected is $p(v_1, v_3) = \alpha/3$. Which means v_2 and v_3 are more likely to share an edge. This follows our intuition that v_2 and v_3 are more likely to have common engagments leading to an edge compared to v_1 and v_2 or v_1 and v_3. Moreover, $p(u,v) = \alpha J(S_u, S_v)$ naturally explains why overlapping communities are generally dense.

A. A New Simpler Theory of Friendship via Shared Communities

Although our model is quite simple, it is not clear what process might lead to a linear relationship between community overlap (Jaccard similarity) and edge formation probability. In this section, we show a simple event based friendship formation process which mathematically leads to this linear relationship.

The Friend Formation Process: It is known that community events are a common place for people to meet and form friendships. We assume that activities keep happening. At a given time instance, a random ranking (total order) of preferences are assigned to the communities. In particular, at a given instance nature samples a global ranking of communities in terms of preferences. This preference is uniformly chosen,
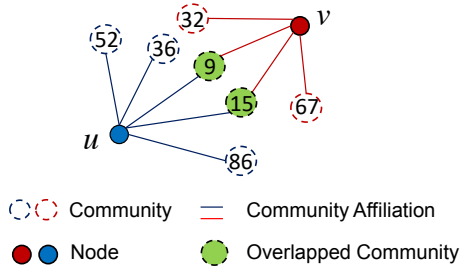
4

Fig. 4. Example of edge formation process. Number inside each community denotes the corresponding preference ranking.

among all possible preferences, at every instance. See Fig. 4 for illustration.

Based on this sampled ranking, every individual u selects the best-ranked community, from the set of the communities he is a member, i.e., S_u. Individual u then chooses to participate in that preferred community event (or is attentive to the community). If two individuals u and v are in the same event, then there is a constant $Const$ chance (usually small) that they form a friendship.

Theorem 1: Under the above friend formation process, the probability of edge between u and v is precisely a constant $\times$ the Jaccard Similarity between their community set $\frac{|S_u \cap S_v|}{|S_u \cup S_v|}$
Proof: For given individuals u and v, consider the set of communities $S_u \cup S_v$ they both can participate. Both individuals will go to the same event, if and only if the most preferred event, out of $S_u \cup S_v$, belongs to the set $S_u \cap S_v$. Note that the ranking is uniform and therefore the probability of u and v ending up in the same event is $\frac{|S_u \cap S_v|}{|S_u \cup S_v|}$.

The final constant α of our model is the $Const$ multiplied by the number of times that the sampling is performed. Note that the probability keeps increasing over time, which is consistent with observed social networks where the edges keep increasing.

Global vs. Personal Preference: It should be noted that our process assumes a global preference over the community events. This assumption is quite realistic on an average. There will be (local) variations, all the time due to personal preferences, but that will have negligible effects so long as a large number of individuals follow the average ranking.

B. How do Connectivity Varies Across Communities?

It is common to have some communities very tightly connected compared to others. AGM model assumes that every community has its own affiliation probability associated with it and edges are formed independently based on this probability. While AGM explains the existence of variations in the connectivity structure, our evidence suggests that it cannot explain the observed findings. Also, AGM requires more parameters, of the order of the number of communities, to specify the connectivity probability.

Under AGM model, there can be two reasons for a given community C to be densely connected: 1) The probability associated with the community is higher and 2) There are several overlapping communities inside C.

On the contrary, based on our model, which does not assume any community specific probability. The likelihood of forming an edge purely depends on the community overlaps. Under this simpler modeling, a community C_1 is tightly connected compared to another community C_2, if and only if, C_1 overlaps with more number of communities compared with C_2. In other words, C_1 has more sub-communities. Thus, our model implies a direct causal relationship between the existence of a larger set of overlapping communities and the density of the connections. This causation is also in line with our edge formation process driven by community preferences. *A direct causal relationship implies that an effective way of promoting friendship is to increase the community affiliation of every individual.*

If JAG model is correct, then different isolated communities, in a given network, should have same edge formation probability (or connectivity density). By isolated communities, we mean a community whose members are not part of any other community, or a community which does not overlap with any other community. We next show this fact is true with surprising precision in all the six network datasets, providing yet another validation of our hypothesis.

C. Validation 3

Consider, *Isolated Communities* which do not have any overlap with any other communities. Then the Jaccard similarity between S_u and S_v of every node u and v in these isolated communities should be equal to $\frac{1}{1} = 1$. Thus the edge formation probability $p(u, v)$ for every node u and v in the isolated community, based on our model, should be $p(u, v) = \alpha J(S_u, S_v) = \alpha$. Note, α is also the slope of the observed plots in Figure 3.

If our model is correct, then we should observe this phenomenon empirically. We again use all the six real networks datasets with ground truth community labels. In all the six network datasets, we extracted the isolated communities. We randomly select 5 communities, where none of its members are part of any other community.

After extracting isolated communities, we calculated the edge probability inside each community. For each dataset, we report the mean and standard deviation of this probability over the extracted five isolated communities. Also, we report the slope of the lines from the Figures 3, which is also the estimated value of α, on each of the six datasets. The results are shown in Table. II.

To our surprise, we can see that the standard deviations are quite small, which clearly indicates that, for a given network, isolated communities have the same edge formation probability. The same holds irrespective of the type of network. Furthermore, this edge formation probability is very close to the estimated slope, or α, as suggested by Equation 2. These observations strongly validate all our claims.

Overall, we have a significantly simpler model, which explains several empirical findings (sections III-A, III-B, and IV-C) consistency on six different real datasets. To the best of our knowledge, there is no other known model which explains all our experimental observations.

5

Data set	α	Edge Probability
Live Journal	0.71	0.70 ± 0.01
Friendster	0.74	0.75 ± 0.02
Orkut	0.85	0.84 ± 0.01
Youtube	0.26	0.26 ± 0.02
Amazon	0.32	0.31 ± 0.03
DBLP	0.41	0.40 ± 0.01

(a) DELETING Operation

(b) ADDING Operation
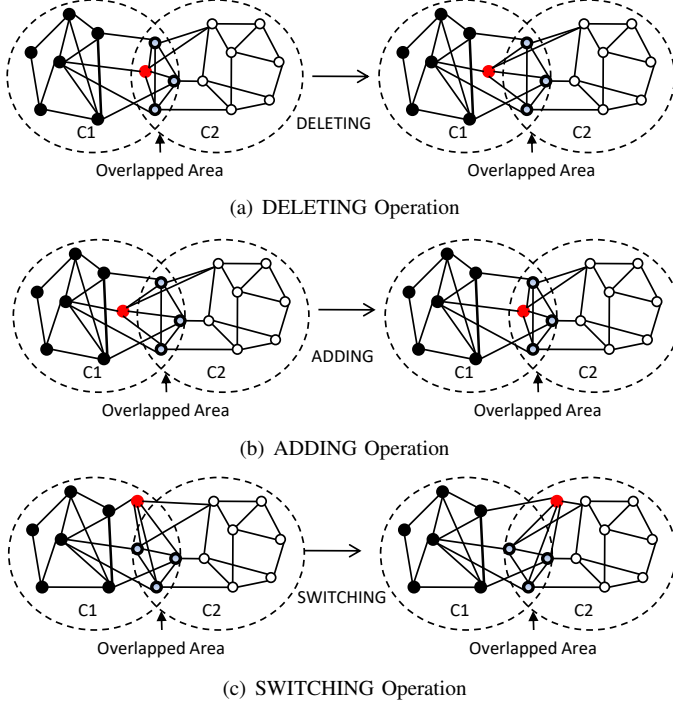

(c) SWITCHING Operation

Fig. 5. Operations for the community assignment: DELETING, ADDING, and SWITCHING

V. FROM MODEL TO COMMUNITY DETECTION ALGORITHM

Having modeled the edge formation probability $p(u,v)$, given any community assignment $A(V,C,M)$, we can evaluate its likelihood. The likelihood of any community assignment can be evaluated by the objective in Equation. 3.

$$
\begin{aligned}
L(A) &= \prod_{(u,v)\in E} p(u,v) \prod_{(u,v)\notin E} (1 - p(u,v)) \\
&= \prod_{(u,v)\in E} \alpha J(S_u, S_v) \prod_{(u,v)\notin E} (1 - \alpha J(S_u, S_v)) \\
&= \prod_{(u,v)\in E} \alpha \frac{|S_u \cap S_v|}{|S_u \cup S_v|} \prod_{(u,v)\notin E} (1 - \alpha \frac{|S_u \cap S_v|}{|S_u \cup S_v|})
\end{aligned}
\tag{3}
$$

In the above, $L(A)$ denotes the likelihood of a community assignment $A(V,C,M)$. S_u and S_v can be obtained from the community assignment $A(V,C,M)$. As a result, we naturally obtain a community detection algorithm. The goal of community detection is to find the community assignment $A(V,C,M)$, and α, which maximizes the likelihood.

Algorithm 1 JAG Algorithm

Input: $G(V,E)$, and $|C|$
Output: $A(V,C,M)$
Initialize a random community assignment A
repeat
    Randomly choose an operation: Deleting, Adding, and Switching.
    Take the Randomly chosen operation on A to generate A'.
    Calculate $\max_\alpha L(A)$ and $\max_\alpha L(A')$ use Equation. 3.
    if $\max_\alpha L(A') > \max_\alpha L(A)$ **then**
        Update $A = A'$
    else
        Update $A = A'$ with probability $\frac{\max_\alpha L(A')}{\max_\alpha L(A)}$.
    end if
until Convergence

A. MCMC Based Assignment Generation

Following [13], we use MCMC (Markov chain Monte Carlo)[16] for updating community assignments. Starting from a random assignment, we keep updating the community assignment $A(V,C,M)$ until we cannot improve the likelihood anymore.

To update the assignment A, we rely on the set of 'OPERATIONS' as defined in [13]. For completeness we reiterate them here. There are three types of operation for generating a new assignment A'. The three operations are showed in Fig. 5. The three operations are DELETING, ADDING, and SWITCHING:

- **DELETING** operation is randomly select an node community pair (v,c) in M, and delete the edge from M. In other words, randomly select a non-zero element in node community relation matrix M', and set this entry as 0; Fig. 5(a) illustrate the deleting process.
- **ADDING** operation is randomly choose an node community pair $(v,c) \notin M$, and add it to M. In other words, randomly select a zero element in node community relation matrix M', and set this entry as 1. 5(b) illustrate the adding process.
- **SWITCHING** operation is randomly select two pairs $(v,c_1) \in M$ and $(v,c_2) \notin M$. Then we delete this the pair (v,c_1), and add the pair $(v,c_2) \notin M'$. 5(c) illustrate the switching process.

After generating a new proposal assignment $A'(V,C,M')$, we then calculate the likelihood value of $\max_\alpha L(A')$. If $\max_\alpha L(A')$ is greater than $\max_\alpha L(A)$, we accept this new community assignment, else, we accept this community assignment with the probability $\frac{\max_\alpha L(A')}{\max_\alpha L(A)}$. This leads to our proposed algorithm: JAG Algorithm.

B. JAG Algorithm

The overall JAG algorithm is summarized as in Algorithm 1. The input of the JAG algorithm contains the unlabeled network structure $G(V,E)$, the number of communities $|C|$. Since there

6

is only one parameter α, we have two options: 1) We can alternate between MCMC and optimizing α, just like [13] or 2) We can simply use a grid search to find the α in $[0, 1]$, that maximizes the likelihood. We use the later, as it is quite easy. Furthermore, there are no concerns about non-convexity of the objective.

AGM, on the other hand, needs to alternate between an MCMC and a costly non-convex vector optimizations which requires another relaxation leading to further approximations to solve it. See [13] for details.

C. Implementation Details

Local Minimum JAG algorithm is a Markov chain searching processing. It is easy to get to a local minimum. So that in practice, we choose a number of initial community assignment and do the Markov chain searching. In addition, at each Markov chain search step, we can usually generate a number of community assignment, and accept the best community assignment. The final result is the community assignment that has the maximum value of Equation .3.

Convergence Although there no rigorous theoretical guarantees for the convergence of our model. But in our experiment, the Markov chain searching process can quickly convergences (A few seconds for a thousand level sampled small graph.). In addition, even for the more complexity model in [13], they also show quickly convergence in practice.

VI. EXPERIMENTAL RESULT

In this section, we show the effectiveness of our modeling assumption by comparing JAG with the state-of-the-art community detection methods on the six real-world datasets that are described before.

A. Dataset Preprocessing

From Table. I, we can see that the six real-world datasets are all very large scale data. For experiments we closely replicate [13], we sample a large enough graph multiple times to get a large number of small sub-networks with overlapping community structures. The steps of preprocessing are shown below: First, we pick a random node u in the given graph G that belongs to at least k communities. In this paper, we use $k = 2$ for preprocessing. We then take the subnetwork to be the induced subgraph of G consisting of all the nodes that share at least one community membership with the randomly sampled node u. Then this subnetwork will be one of the sampled sub-network. In this experiments, we create 500 subnetworks for each network.

B. Baseline Methods

To evaluate the effectiveness of our JAG Model, we choose four baseline algorithms in our experiment: AGM, Link Clustering, Clique Percolation Method, and MMSB.

AGM [13] is the state-of-the-art community detection algorithm which is our main baseline. We further let AGM use this number of communities by providing the fixed value in the code. This modification ensures fairness of comparisons with our algorithm.

LC (Link clustering) [17] is a method that first transfer the adjacent matrix of the graph into a link similarity matrix, and do clustering on the link similarity, then transfer it back to node adjacent metric. For LC, we used the implementation in the Stanford Network Analysis Platform.

CPM (Clique Percolation Method) [18] uses Clique percolation to detect network communities. For CPM, we set the communitie size $k=5$ as suggested in [18]. We use the Stanford Network Analysis Platform for CPM experiment.

MMSB (Mixed-Membership Stochastic Block Model) [19], [20], [21], [22], [23] combines global parameters that instantiate dense patches of connectivity (blockmodel) with local parameters that instantiate node-specific variability in the connections (mixed membership) to do community detection. We use the ground truth community number as the input of MMSB. For MMSB we used publicly available 'LDA' R package.

C. Evaluation metrics

We use three popular evaluation metrics in this paper to evaluate the performance of all the methods: **F1-score** [24], **Omega Index** [24], and **NMI (Normalized Mutual Information)** [25]. Due to the page limits, the details of these three metrics can be found in [24], [25].

All the measures introduced before are in the range of 0 to 1, and a higher value indicates a better community detection result in terms of the ground truth.

D. Result and Discussion

Following [13], we do all the experiment on the sampled 500 sub-networks for all the data sets. We measure the detection result on each sub-network by all the three metrics separately. And then we average the result of all the 500 sampled sub-networks. The result is shown in Table IX.

Compared with AGM methods, we can see that JAG model is accurate on all the dataset. This demonstrates that the Jaccard similarity between the community sets of two individuals is the right "sweet" measure for modeling the edge formation probability. We would like to emphasize that AGM require more community-specific parameter which is hard to estimate, especially for small communities. On the contrary, JAG model only requires one parameter. The better performance of JAG also shows another empirical support of our model.

Compared to all other baseline algorithms, we can see that JAG is consistently achieving the state-of-the-art performance in most of the metric on all the datasets, sometimes significantly outperforming other baselines, including the state-of-the-art AGM. We attribute this to superior and nearly parameter-free modeling which is less prone to over-fitting. Note that JAG with just one parameter can explain several network observation as explained in previews sections.

VII. CONCLUSION

We show that Jaccard similarity of the community sets associated with a pair of individuals seems like the "sweet"

7

Methods	F1-Score	NMI	Omega Index
JAG	**.757**± **.01**	**.612**± **.02**	**.723**± **.03**
AGM	.751± .02	.565± .01	.637± .02
LC	.455± .02	.452± .02	.698± .01
CPM	.557± .01	.511± .02	.658± .02
MMSB	.712± .02	.315± .02	.719± .02

TABLE III
LIVE JOURNAL NETWORK

Methods	F1-Score	NMI	Omega Index
JAG	**.756**± **.01**	**.725**± **.02**	**.716**± **.01**
AGM	.742± .06	.574± .02	.647± .02
LC	.348± .02	.482± .02	.542± .02
CPM	.327± .02	.532± .02	.681± .03
MMSB	.612± .01	.315± .01	.426± .02

TABLE IV
FRIENDSTER NETWORK

Methods	F1-Score	NMI	Omega Index
JAG	.714± .01	**.635**± **.02**	**.759**± **.01**
AGM	.734± .01	.577± .02	.658± .03
LC	.449± .03	.484± .06	.547± .01
CPM	.512± .01	.498± .01	.612± .03
MMSB	**.742**± **.01**	.365± .02	.516± .02

TABLE V
ORKUL NETWORK

Methods	F1-Score	NMI	Omega Index
JAG	.704± .01	**.498**± **.02**	**.709**± **.01**
AGM	.687± .03	.425± .02	.636± .02
LC	.265± .02	.482± .01	.625± .01
CPM	.162± .02	.181± .01	.576± .02
MMSB	**.721**± **.03**	.215± .02	.619± .03

TABLE VI
YOUTUBE NETWORK

Methods	F1-Score	NMI	Omega Index
JAG	**.761**± **.02**	**.528**± **.01**	**.702**± **.01**
AGM	.707± .02	.408± .06	.652± .01
LC	.458± .03	.182± .02	.521± .04
CPM	.428± .02	.539± .02	.646± .01
MMSB	.752± .06	.510± .02	.583± .03

TABLE VII
DBLP NETWORK

Methods	F1-Score	NMI	Omega Index
JAG	**.648**± **.02**	**.575**± **.02**	.572± .01
AGM	.617± .06	.526± .04	.495± .01
LC	.556± .02	.568± .02	.201± .01
CPM	.257± .01	.324± .02	.126± .02
MMSB	.512± .02	.351± .02	**.725**± **.02**

TABLE VIII
AMAZON CO-PURCHASING NETWORK

TABLE IX
THE EXPERIMENT RESULT ON SIX DATASETS. COMPARED TO ALL OTHER BASELINE ALGORITHMS, WE CAN SEE THAT JAG MODEL IS CONSISTENTLY THE BEST IN MOST OF THE METRIC ON ALL THE DATASETS.

measure for modeling edge formation probability. This modeling, unlike popular AGM, does not require to have local parameters for each community, instead, we only need one global hyperparameter to explains most of our observed findings. We call our parameter-free model Jaccard-based Affiliation Graph (JAG) model.

As a result, we obtain a parameter-free community detection algorithm, which on six real-world datasets, achieves state-of-the-art overlap community detection performances compared with other algorithms, including AGM which requires significantly more parameters.

REFERENCES

[1] P. Hoff, "Modeling homophily and stochastic equivalence in symmetric relational data," in *Advances in neural information processing systems*, 2008, pp. 657–664.

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

[3] S. Kumar, J. Cheng, J. Leskovec, and V. Subrahmanian, "An army of me: Sockpuppets in online discussion communities," in *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2017, pp. 857–866.

[4] M. Farajtabar, X. Ye, S. Harati, L. Song, and H. Zha, "Multistage campaigning in social networks," in *Advances in Neural Information Processing Systems*, 2016, pp. 4718–4726.

[5] A. De, I. Valera, N. Ganguly, S. Bhattacharya, and M. G. Rodriguez, "Learning and forecasting opinion dynamics in social networks," in *Advances in Neural Information Processing Systems*, 2016, pp. 397–405.

[6] C. Luo, R. Guan, Z. Wang, and C. Lin, "Hetpathmine: A novel transductive classification algorithm on heterogeneous information networks," in *European Conference on Information Retrieval*. Springer, 2014, pp. 210–221.

[7] C. Luo, W. Pang, and Z. Wang, "Semi-supervised clustering on heterogeneous information networks," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2014, pp. 548–559.

[8] C. Luo, W. Pang, Z. Wang, and C. Lin, "Hete-cf: Social-based collaborative filtering recommendation using heterogeneous relations," in *2014 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2014, pp. 917–922.

[9] J.-F. Rual, K. Venkatesan, T. Hao, T. Hirozane-Kishikawa, A. Dricot, N. Li, G. F. Berriz, F. D. Gibbons, M. Dreze, N. Ayivi-Guedehoussou *et al.*, "Towards a proteome-scale map of the human protein–protein interaction network," *Nature*, vol. 437, no. 7062, pp. 1173–1178, 2005.

[10] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne, "Finding high-quality content in social media," in *Proceedings of the 2008 International Conference on Web Search and Data Mining*. ACM, 2008, pp. 183–194.

[11] M. E. Newman, "Modularity and community structure in networks," *Proceedings of the National Academy of Sciences*, vol. 103, no. 23, pp. 8577–8582, 2006.

[12] J. Xie, S. Kelley, and B. K. Szymanski, "Overlapping community detection in networks: The state-of-the-art and comparative study," *ACM Computing Surveys*, vol. 45, no. 4, p. 43, 2013.

[13] J. Yang and J. Leskovec, "Community-affiliation graph model for overlapping network community detection," in *Data Mining (ICDM), 2012 IEEE 12th International Conference on*. IEEE, 2012, pp. 1170–1175.

[14] J. Yang, J. McAuley, and J. Leskovec, "Community detection in networks with node attributes," in *2013 IEEE 13th International Conference on Data Mining (ICDM)*. IEEE, 2013, pp. 1151–1156.

[15] C. M. Bishop, *Pattern recognition and machine learning*. springer, 2006.

[16] W. R. Gilks, S. Richardson, and D. Spiegelhalter, *Markov chain Monte Carlo in practice*. CRC press, 1995.

[17] Y.-Y. Ahn, J. P. Bagrow, and S. Lehmann, "Link communities reveal multiscale complexity in networks," *Nature*, vol. 466, no. 7307, pp. 761–764, 2010.

[18] I. Derényi, G. Palla, and T. Vicsek, "Clique percolation in random networks," *Physical Review Letters*, vol. 94, no. 16, p. 160202, 2005.

[19] E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing, "Mixed membership stochastic blockmodels," *Journal of Machine Learning Research*, vol. 9, no. Sep, pp. 1981–2014, 2008.

[20] M. Zhou, "Infinite edge partition models for overlapping community detection and link prediction," in *Artificial Intelligence and Statistics*, 2015, pp. 1135–1143.

[21] P. K. Gopalan and D. M. Blei, "Efficient discovery of overlapping communities in massive networks," *Proceedings of the National Academy of Sciences*, vol. 110, no. 36, pp. 14 534–14 539, 2013.

[22] B. Ball, B. Karrer, and M. E. Newman, "Efficient and principled method for detecting communities in networks," *Physical Review E*, vol. 84, no. 3, p. 036103, 2011.

[23] B. Karrer and M. E. Newman, "Stochastic blockmodels and community structure in networks," *Physical Review E*, vol. 83, no. 1, p. 016107, 2011.

[24] C. J. V. Rijsbergen, *Information Retrieval*, 2nd ed. Newton, MA, USA: Butterworth-Heinemann, 1979.

[25] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.