CONSENSUS AND MULTIPLEX APPROACH FOR COMMUNITY DETECTION IN ATTRIBUTED NETWORKS

Yuming Huang[⋆]*, Han Wang*[†]

[⋆]Department of Physics
[†]Department of Electrical and Computer Engineering
North Carolina State University
Raleigh, NC

ABSTRACT

An attributed network has nodes with attribute vectors. For community detection on an attributed network, we exploit the attributes to disentangle the potentially mixed topological structures. We describe a multiplex representation scheme for overlapping community detection in attributed networks and find consensus communities across layers of different connection structures. We test the method on Twitter, Facebook and Google+ networks and the results are comparable to state of the art. We show that the use of attribute vectors improves detection accuracy. Experiments on synthesized networks show that this method improves the detectability of communities.

Index Terms— community detection, attributed network, multiplex, consensus, community detectability

1. INTRODUCTION

There may be multiple relations between two nodes in a network. This leads to potentially overlapping communities. Typically in a social network where people have multiple relations, overlapping communities are naturally introduced when one person belongs to several communities [1]. Many overlapping community detection algorithms have been proposed focusing on analyzing the network topology [2][3][4]. As online social networks flourish, much information is available to aid community detection. For example, user profiles can be transformed to node attributes. Recently people incorporate both network topology and node attributes to do community detection, mostly from a matrix factorization perspective [5][6][7]. Here we describe a different scheme, starting with a multiplex network in order to express multiple relations between nodes.

The advantage of using multiplex networks emerges when the underlying network includes overlapping communities. In the context of social networks, two persons can be related due to a myriad of reasons. Therefore, the underlying network may not exhibit clear community structure, thus performing community detection directly on such a network will not be effective. The idea is to separate mixed relations into multiple layers. However the criteria according to which relations are assigned to layers are often debatable. How refined multiple relations are defined has influence on the completeness of each layer. If we look at relations in a coarse scale, i.e., a single layer is the accumulation of several different relations, overlapping communities may be formed in a single layer since one agent can belong to different communities when different relations are considered. On the other hand, if we look at relation networks in the finest level, i.e. we separate two relations when they can at least be slightly distinguished and we put them respectively in two layers, these layers will not suffer from mixture of information, but they may be so sparse that they can only be taken as a fragment of a complete layer. Also, in this way, several layers may present similar pattern or community structures, and this will also introduce the redundancy of multiplex networks. In fact, endeavors [8][9] have been taken to reduce the structure of multiplex networks.

One attempt to tackle this problem is to aggregate similar layers to reduce redundancy and increase community detectability [9][10]. It was found that when several layers of networks are generated from the same stochastic block model, aggregating them into one effectively increases the community detectability. However it is sometimes difficult to determine whether layers are intrinsically the same or different, aggregating them may introduce artifacts for overlapping communities in the combined layer. Also, the same community can exist across multiple layers, and at the same time the layers are different in other parts. In such scenario certain layers imply the existence of the community, but we cannot simply aggregate them, which is related to the so-called emerging clusters [5]. In this paper, we attempt to tackle this difficulty by leveraging similar parts across layers without assuming or requiring that some layers are generated from the same stochastic block model, and hence increase the detection performance of overlapping communities utilizing multiplex network. We test this on node attributed social networks by proposing a way to convert it into a multiplex network, and look for common communities over partial consensus layers.

2. PROBLEM AND METHOD

An attributed network is a network with node or edge attributes. Such additional information may give benefits when we analyze network structures. In this paper we focus on node-attributed networks and our approach can be straightforwardly extended to edge-attributed networks. More specifically, suppose a network consists of N nodes $\{n_1, n_2, n_3...n_N\}$. The edge information is encoded in adjacency matrix A. Suppose each node has L attributes. We define node-attribute incidence matrix X as follows:

$$X_{il} = \begin{cases} 1, \text{ if node } n_i \text{ has attribute } l \\ 0, \text{ if node } n_i \text{ does not has attribute } l \end{cases} \quad (1)$$

where $i \in \{1, ..., N\}$ and $l \in \{1, ..., L\}$. Therefore, each row vector represents a node and each column corresponds to an attribute. We will try to utilize both A and X, to uncover the overlapping community structure.

The method can be expressed in three steps:

(1) Based on each attribute we build a network layer, encoded with adjacency matrix W^l for the l-th attribute, and

$$W^l_{ij} = A_{ij} \cdot X_{il} \cdot X_{jl} \quad (2)$$

therefore, two nodes n_i and n_j are connected in W^l if and only if they both have attribute l and they are connected in the underlying network A. We call them *Source Networks*, denoted as Multiplex(S). The intuition comes from considering topic diffusions. If we treat agents that post certain information as source nodes, it requires that the path of the information flow consists of entirely source nodes, because one has to tweet so the followers can catch and pass the message. In that case attributes are topics. Note that when attributes are users' profiles, as in Facebook networks, this definition assumes two users are connected if and only if they are connected in the network, and share the same attributes (same interest, place, etc).

Other definitions of W might be useful as well. We define $W^l_{ij} = A_{ij} \cdot [X_{il} + X_{jl}]/2$. In this way, we establish a link between two nodes when at least one of them has the attribute, and if both have the attribute, the edge weights heavier. We call them *Audience Networks*, denoted as Multiplex(A). In twitter network, for example, a connected component in Multiplex(A) will include both source nodes and the audiences (followers who do not necessarily pass on the information).

With a proper definition of W, we will derive a multiplex network of L layers, and the corresponding adjacency matrices are denoted as: $W^1, W^2...W^L$, respectively for each layer.

(2) We perform community detection for each layer, W^l, by maximizing the so-called modularity, which is a community quality function [11]: $Q = \frac{1}{2m} \sum_{ij} (W_{ij} - \frac{k_i k_j}{2m}) \delta(c_i, c_j)$, where $m = \frac{1}{2} \sum W_{ij}$, which is a normalization factor; k_i is the weighted degree of node i, and c_i is the community assignment of node i. $\delta(c_i, c_j) = 1$ if $c_i = c_j$, and 0 otherwise. We use the most popular Louvain heuristics, and the code we used is open-sourced [12]. We build a community assignment matrix for each layer, P^l, N-by-K^l, with N nodes being assigned to K^l non-overlapping communities.

$$P^l_{ik} = \begin{cases} 1, \text{ if node } n_i \text{ is in community } k \text{ in layer } l \\ 0, \text{ otherwise} \end{cases} \quad (3)$$

Note that this step does not limit the choice of community detection method. Theoretically any effective community detection method can be used to generate P^l_{ik}.

(3) We concatenate $\{P^1, P^2, P^3...P^L\}$ into a single community assignment matrix, P. P is N-by-K, where $K = \sum_{l=1}^{L} K^l$. Then we cluster columns into K_0 clusters, where K_0 is the real number of communities (maybe known or inferred by data). Here we used k-means clustering over all column vectors in P for simplicity, and we take the union of nodes in the clusters as a community. Nevertheless, as proposed in [13] this step can also be formulated as: $\min \|P - GH\|_1 \; s.t. \; G, H \geq 0$, where $G \in R^{N \times k_0}$ is assignment of N nodes to K_0 communities (a node can be assigned to multiple communities), and $H \in R^{K_0 \times K}$ is the clustering of K communities into K_0 communities. Tepper et.al [13] utilize this clustering approach on a single layer network, to assemble local community detection results into a global one, and they proposed a way to avoid manually setting the number of clusters. Their method can straightforwardly applied in this step to look for consensus communities across multiple layers of different structures.

3. EXPERIMENTS

3.1. Attributed social networks

3.1.1. Data sets

We apply the method to a popular dataset [14] that has been used to test various overlapping community detection algorithms. The data includes social networks of Facebook, Twitter and Google+. The networks encode friendship or follower-followee relationship. Node attributes are provided as well. For Facebook and Google+, node attributes are users' profile data, for instance one's work place, company, education and so on. For Twitter, node attributes are the hashtags(#) and mentions(@) that are used in tweets.

Table 1. F1 score, reference values reprinted from [6]

Methods	Facebook	Google+	Twitter	Average
Infomap	0.1691	n/a	0.2117	0.1904
BigClam	0.4199	0.2475	0.2253	0.2975
CESNA	0.42106	0.2244	**0.2462**	0.29722
3NCD	0.44075	**0.2570**	0.2406	**0.3128**
Multiplex(S)	0.4166	0.2043	0.2421	0.2877
Multiplex(A)	**0.4416**	0.2033	0.2438	0.2962

Table 2. Jaccard similarity, reference values reprinted from [6]

Methods	Facebook	Google+	Twitter	Average
Infomap	0.1063	n/a	0.1461	0.1264
BigClam	0.3016	0.1509	0.1448	0.1991
CESNA	0.3022	0.1428	0.1568	0.2006
3NCD	0.3208	**0.1712**	0.1565	**0.2162**
Multiplex(S)	0.3025	0.1315	**0.1692**	0.2011
Multiplex(A)	**0.3331**	0.1309	0.1625	0.2088

3.1.2. Evaluation metrics

For comparison, we apply the same evaluation metrics as used in [6][4][7]. The evaluation function is defined as:

$$\frac{1}{2\left|C^*\right|}\sum_{C_i^*\in C^*}\max_{C_j\in C}s(C_i^*,C_j)+\frac{1}{2\left|C\right|}\sum_{C_j\in C}\max_{C_i^*\in C^*}s(C_j,C_i^*) \tag{4}$$

where C is a set of detected communities and C^* is a set of ground truth communities, and $s(C_i^*, C_j)$ is a similarity measure between two sets of communities, with a value between 0(poor) and 1(perfect), which makes the evaluation function between 0 and 1 as well. We employ $F1$ score and Jaccard similarity for $s(\cdot)$.

3.1.3. Results

The results are compared with Infomap [2][15], BigClam[4], CESNA[7], 3NCD[6]. The values to compare with are obtained from [6]. We perform exactly the same evaluation, providing the actual number of communities when clustering, as done in the reference.

On average, our method is comparable to these recent algorithms for overlapping community detection. We get decent results on Facebook and Twitter networks, especially for Multiplex(A).

3.1.4. Analysis for combination of attribute and underlying network

In order to justify the way we convert node attributes into layers, we combine the layer generated from attributes and the underlying network:

$$W^* = (1-\alpha)\frac{W}{\bar{W}} + \alpha\frac{A}{\bar{A}} \tag{5}$$

where W^* is the modified adjacency matrix for the layer, $\bar{W}$ and $\bar{A}$ are the mean of non-zero elements in the matrices, and α is a factor that controls how strong the attributes affect the weights on the underlying network. In the following experiment, we want to test whether putting more weights on attributes will improve the results. For a more general comparison across different α values, we skip the clustering in step 3, in order to avoid the effect of manually setting the number of clusters(the actual number of communities). We compare directly the result from step 2 with ground truth communities and calculate the mean F1 score as described before.

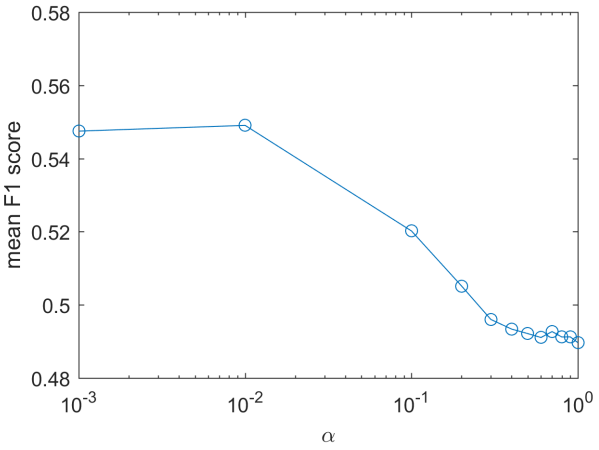


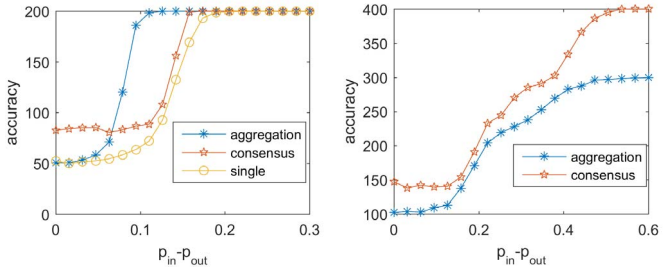
Fig. 1. Effect of combining node attributes with the underlying network. Experiments are performed on Facebook networks. Results are averaged over 5 runs.

In Fig.1, we can see that as we decrease α, i.e. reduce the proportion of underlying adjacency matrix and emphasize the attribute information, the community detection performance(mean F1 score) gets better. This transition happens between around 0.01 to 0.3. It shows that increasing the weight of attribute layer may enhance the community detection performance(mean F1 score) by about 12%. This result supports the idea of converting attributes into layers.

3.2. Detectability transition on synthetic networks

Community detectability undergoes a phase transition as the connection probability difference between intra-community and inter-community changes, which is $P_{in} - P_{out}$ in Fig.2. We examine detectability transition by looking at detection accuracy, i.e., the number of overlapped nodes between results and ground truth. In Fig.2(a) We can see that the circle markers show the transition in a single layer of network, with 200 nodes in total and two communities, 100 nodes each. The asterisk markers show the improvement of detectability after

aggregating three of such layers. Note that these three layers are different realizations from the same stochastic block model. The star markers show the improvement by simply applying our method to find consensus communities across these three layers, without using the fact that they are from the same stochastic block model. The improvement over single layer network is not as strong as aggregating layer. However when layers are different, as in Fig.2(b) especially when overlapping communities are presented, finding consensus communities across layers can give consistently better detection accuracy than layer aggregation method.



(a) identical layers (b) overlapping communities

Fig. 2. (a) For method of aggregation and consensus, 3 identical layers are used. For method of single, only one of the layer is used. In each layer, there are in total 200 nodes, with two separate communities each of 100 nodes. (b) For both methods, a multiplex network of two layers is used. Each layer has two separate communities of 150 nodes and 50 nodes. However the two layers have partial overlap in the 150-node communities, and the overlap is 100 nodes. The two 50-node communities in two layers have no overlap. Results are averaged over 100 runs.

4. CONCLUSIONS

In this paper we demonstrate a new scheme for community detection on attributed networks. We construct a multiplex network considering node attributes. Making use of the popular modularity optimization method as a middle step, we obtain the final results by grouping consensus communities across layers of different structures. Through our experiments, as a proof of concept we show that by incorporating attributes information the accuracy is increased, also, finding consensus communities seems to increase the community detectability than a single layer, without assuming whether or not there are overlaps. To our knowledge, this is the first application/validation of the idea of consensus clustering on multiplex networks. Note that each step we describe here can be modified and optimized for certain data or objectives. The construction of a multiplex network out of a attributed one, is not only posing benefits for overlapping community detection, as a way to decouple the mixed information, this may also help with analysis of network dynamics, such as information diffusion on certain topical layers. Furthermore, finding consensus communities across layer is not limited to the case we described here, and it can be applied to general multiplex networks as well.

5. REFERENCES

[1] Gergely Palla, Imre Derenyi, Illes Farkas, and Tamas Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society - Supplementary," *Nature*, vol. 435, no. 7043, pp. 814–818, 2005.

[2] Martin Rosvall and Carl T Bergstrom, "Maps of random walks on complex networks reveal community structure.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 4, pp. 1118–23, jan 2008.

[3] Ioannis Psorakis, Stephen Roberts, Mark Ebden, and Ben Sheldon, "Overlapping community detection using Bayesian non-negative matrix factorization," *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, vol. 83, no. 6, pp. 1–9, 2011.

[4] Jaewon Yang and Jure Leskovec, "Overlapping community detection at scale: A Nonnegative Matrix Factorization Approach," *Sixth ACM international conference on Web search and data mining*, p. 587, 2013.

[5] Cecile Bothorel, Juan David Cruz, Matteo Magnani, and Barbora Micenková, "Clustering attributed graphs: Models, measures and methods," *Network Science*, , no. January, pp. 1–37, 2015.

[6] Hung T Nguyen and Thang N Dinh, "Unveiling the Structure of Multi-attributed Networks via Joint Nonnegative Matrix Factorization," pp. 1412–1417, 2015.

[7] Jaewon Yang, Julian McAuley, and Jure Leskovec, "Community detection in networks with node attributes," *Proceedings - IEEE International Conference on Data Mining, ICDM*, pp. 1151–1156, 2013.

[8] Manlio De Domenico, Vincenzo Nicosia, Alexandre Arenas, and Vito Latora, "Structural reducibility of multilayer networks," *Nature Communications*, vol. 6, pp. 6864, 2015.

[9] Natalie Stanley, Saray Shai, Dane Taylor, and Peter J. Mucha, "Clustering Network Layers with the Strata Multilayer Stochastic Block Model," *IEEE Transactions on Network Science and Engineering*, vol. 3, no. 2, pp. 95–105, apr 2016.

[10] Dane Taylor, Saray Shai, Natalie Stanley, and Peter J. Mucha, "Enhanced Detectability of Community Structure in Multilayer Networks through Layer Aggregation," *Physical Review Letters*, vol. 116, no. 22, pp. 228301, jun 2016.

[11] M. Newman, "Fast algorithm for detecting community structure in networks," *Physical Review E*, vol. 69, no. 6, pp. 066133, jun 2004.

[12] Inderjit S. Jutla, Lucas G. S Jeub, and Peter J Mucha, "A generalized Louvain method for community detection implemented in MATLAB, http://netwiki.amath.unc.edu/GenLouvain," 2014.

[13] Mariano Tepper and Guillermo Sapiro, "From Local to Global Communities in Large Networks Through Consensus," *Ciarp*, vol. 5197, no. c, pp. 602–609, 2015.

[14] Jure Leskovec and Jj Mcauley, "Learning to discover social circles in ego networks," *Advances in neural information processing ...*, pp. 1–9, 2012.

[15] Alcides Viamontes Esquivel and Martin Rosvall, "Compression of Flow Can Reveal Overlapping-Module Organization in Networks," *Physical Review X*, vol. 1, no. 2, pp. 1–11, 2011.