Community Detection In Attributed Graphs

Sheng Guan
School of EECS, Washington State University

Advisor:
Dr. Yinghui Wu
Committee Members:
Dr. Assefaw Gebremedhin (Chair), Dr. Yinghui Wu and Dr. Venera Arnaoudova

---

Abstract

Real-world networks can often be represented as complex networks that have a topology of interconnected nodes, such as social, protein, and chemical network. Most networks of interest display community structure, where vertices are grouped together to serve as a functionality of the system. Community plays an important role when we want to analyze the interactions within the system.

However, conventional community detection often focuses only on single labeled entities and is derived from the network topology, yet ignoring the potential correlation between node attribute values and relationships. Emerging research interests study meaningful community representation in attributed networks. The mined community is jointly characterized by both node attributes and network topology. To integrate attribute values with the existing methods, there are three major challenges as follows. (1) The community structure may change in the semantic context because of attribute values; (2) Topology features and attribute values should be modeled

together; (3) The massive graph size and dirty data issue bring in new challenges.

In this paper, 1) we first introduce the problem of community detection in attributed graphs. We give the summary of one related paper CESNA [1] and present a thorough critical evaluation. Next, 2) we give a detailed survey on community detection in attributed graphs. We classify the existing methods into three different main categories and summarize their strengths and weaknesses. 3) We propose our research idea – attributed backbone structure that tries to deal with the above-mentioned three challenges and can capture the dynamics in the edge cost model by specifying affinitive attributes for each edge.

Contents

1. Introduction

Real-world networks can often be represented as complex networks that have a topology of interconnected nodes [2]. Networks occur in a huge variety of contexts. For instance, the Facebook network is a social network, where more than one billion people are connected via virtual acquaintanceships. Another interesting example is Youtube network, where videos are linked

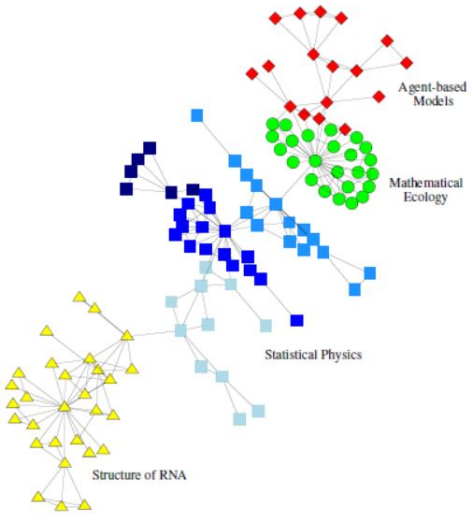

Figure 1: Collaboration network of scientists. Community indicate the research areas of the scientists. Reprinted figure from [4].

through similarity and popularity. Many other examples come from marketing, ecology, biology, telecommunication, physics, economics, engineering, computer science, social and political sciences, etc.

Most networks of interest display community structure, where vertices are organized into groups, often called communities [3]. Fig. 1 shows a collaboration network of scientists. Vertices are scientists. Edges connect coauthors. Different colors represent different communities. Communities here represent scientists working on the same research topic and their collaborations are more explainable. Likewise, communities could represent friendship or common interest in social networks, functional modules in protein-protein interaction networks, chemical reaction modules in the chemical reaction network, and so on.

The community is considered to be a significant property of real-world networks as it often accounts for the functionality of the system. Hence, the community detection problem emerges as an interesting problem to help us understand the interactions within the systems. Numerous techniques have been developed for both efficient and effective community detection [5]. Random walks, spectral clustering, modularity maximization, differential equations, and statistical mechanics have all been used previously [6, 2, 7, 8].

However, the majority of the research work neglects the fact that real-world graph data are often associated with additional information, i.e. vertices of a graph are associated with a number of attributes that describe the vertex. For example, in social networks, edge often represents the relationship (friendship, collaboration, family, etc) among people while the vertex attributes describe the role, the personality or the behavior pattern of a person. Node attributes have the potential to complement the network structure, leading to more precise detection of communities.

The idea that attributes and connections are generated in an interdependent way has led to the development of specialized methods – community detection in attributed graphs.

What is more, while prior work identifies communities (node sets) such that all the nodes in a community share same set of attributes, we would like to consider a structure that allows each edge to have its own affinitive attributes. We formulate attribute-driven backbones to characterize the connectivity patterns among the nodes of interests (Section 4). We develop a bi-criteria cost model to characterize good backbones that incorporates both

node interestingess and edge cost. This measure is determined by the selection of node attributes and topology of the backbones.

We summarize the main contributions of this paper as follows:

(1) We give the problem definition of community detection in an attributed graph and provide a summary of one related paper CESNA [1] that infers community structures by statistically modeling the interaction between the network structure and the node attributes. We present a thorough critical evaluation of this paper and state its limitations.

(2) We classify the existing methods on attributed community detection of graphs into three different main categories. We briefly review the weaknesses of existing methods and indicate that the challenges in the problem of community detection in attributed graphs are still not fully solved.

(3) We formulate attribute-driven backbones to characterize the connectivity patterns among the nodes of interests (Section 4). We develop a bicriteria cost model to characterize good backbones that incorporates both node interestingess and edge cost. This measure is determined by the selection of node attributes and topology of the backbones. To deal with dirty data issue, i.e. erroneous attribute values, we propose to utilize Graph Neural Networks to clean the attributed graph.

To explore this task, this paper is organized as follow. First, we introduce, in section 2, the problem of community detection in attributed graphs and a summary of one related paper CESNA [1]. Section 3 presents a detailed survey on community detection in attributed graphs. In section 4, we propose our own models– attributed backbone as a tree structure representation.

Finally, section 5 concludes the present article.

2. Problem definition and paper summary

In many applications, topological information as well as attribute data are available for the objects. Both types of information can be modeled as a vertex labeled graph such that vertices represent objects, edges represent relations between them, and feature vectors associated to the vertices represent the attributes information for each object.

2.1. Problem Definition

Definition 2.1: An attributed graph G is defined as 4-tuple (V,E,A,X), where $V = \{v_1, v_2, ..., v_n\}$ is a set of n vertices, $E = \{(u, v) : u, v \in |V|, u \neq v\}$ is a set of edges, $A = \{a_1, a_2, ..., a_T\}$ is a set of T attributes, $X = \{f_1, f_2, ..., f_T\}$ is a set of T attributes functions and each function $f_t : V \rightarrow dom(a_t)$ assigns to each vertex in V an attribute value in the domain $dom(a_t)$ of the attribute a_t (for $t : 1 \leq t \leq T$). In the attributed graph G, a vertex $v \in V$ is essential associated with an attribute vector of length T, where the element t in the vector is given by the function $f_{a_t}(v)$. □

Then, we give our problem statement as follows:

Problem 1. *Given an attributed graph G(V,E,A,X) and the number of clusters k, the detection problem is to partition the vertex set V of G into k subsets $P = \{C_1, C_2, ..., C_k\}$, such that:*

(1)If $C_i \cap C_j = \emptyset$ $(\forall i \neq j, \cup_i C_i = |V|)$ satisfies, we call this type of community as non-overlapping community. Otherwise, when $C_i \cap C_j \neq \emptyset$, we call this type of community as overlapping community.

(2)Vertices within the same community are densely connected, while the vertices in different communities are sparsely connected.

(3)Nodes in the same community are expected to have homogeneous attributes.

2.2. Paper Summary

In this section, we give a summary of one related paper [1]. Main contributions, their findings and evaluation approaches/results are covered. Paper [1] statistically modeled the interaction between the network structure and the node attributes, which leads to more accurate community detection as well as improved robustness in the presence of noise in the network structure.

2.2.1. Community Detection in Networks with Node Attributes

This paper developed Communities from Edge Structure and Node Attributes (CESNA), an accurate and scalable algorithm for detecting overlapping communities in networks with node attributes. As we discussed before, traditional community detection methods only consider one of two sources (nodes with attributes, the edge connections) independently. However, these two types of information should be considered interdependently. For instance, attributes can tell us to which community a node belongs to when this node only has few links connected to others. Conversely, the network topology feature can tell us to which community a node belongs to even when this node has no attribute information. In short, node attributes information can complement the network topology information, which leads to a more accurate detection of communities.

What is more, in contrast to a line of previous work, which assumed the communities and attributes are marginally independent, this paper proposed a new view that community memberships generate both the graph and attributes. This point of view is supported by [9] and shows a way to model how the network structure and attribute information depend on node community memberships.

Traditional probability distribution-based community detection methods will assign a probability value to indicate the node's membership. However, since probabilities have to sum to one (we often call this as soft-membership models) [10], the more communities a node belongs to, the less it belongs to each individual community. This is counter-intuitive since one node can be strongly associated with multiple communities at the same time. The CESNA model associates an independent variable for each node and community pair and, thus, do not suffer from this issue made by soft-membership models.

The major contribution of this paper can be summarized as:

(1)CESNA answers the question that how we can model the interaction between the network topology and the attribute values.

(2)CESNA allows for rich modeling of network communities: (a) non-overlapping, (b) overlapping, (c) nested.

(3)The experimental results show both effectiveness and efficiency of the CESNA, which outperforms the baselines.

Next, we give a more detailed explanation of the generative model used in CESNA.

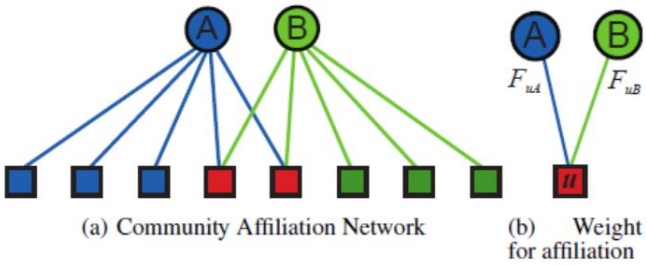

Figure 2: Bipartite community affiliation graph. Circles: Communities, Squares: Nodes of the G. Edges indicate node community memberships. Edge weights F_{uc} indicate strength. Reprinted figure from [11].

2.2.2. Generative model analysis

The CESNA model is a generative model that takes the input and generates the output as follows.

- **Input**: Attributed graph $G(V, E, A, X)$, ;
- **Output**: A partition P and community membership weight F for each node-community pair.

To construct the attributed graph G, node community affiliation influence the likelihood that a pair of nodes is connected. The CESNA starts with a bipartite graph where the nodes at the bottom represent the nodes of the network G, the nodes on the top represent communities C, and the edges M indicate node community affiliations. We denote the bipartite affiliation network as $B(V, C, M)$, as shown in Fig. 2.

The advantage of bipartite graph representation is that CESNA is able to model communities that have overlappings.

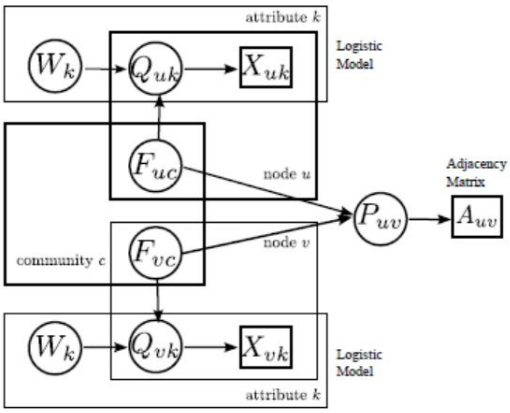

Figure 3: F_{vc}, F_{uc}: membership strength of node u, v to community c, $\rightarrow P_{uv}$: probability that edge A_{uv} exsits. F_{uc}, $W_k \rightarrow Q_{uk}$: probability that attribute value X_{uk} equals to 1. Reprinted figure from [1].

Similarly, for each attribute A_{u1},...,A_{uk}, they modeled the probability of logistic weight vector W_k for attribute k with a computable function $f_2(F_{uc}, W_{kc})$. For each attribute, they used a separate logistic model. The overall model representation of CESNA is shown in Fig. 3.

Then their objective function is a maximum likelihood objective $\log P(G, X|F, W)$ of the observed G and A. They can infer the values of latent variables F and W. To solve the objective function, they adopt a block gradient ascent approach. To be specific, when updating the membership F_u of an individual node u, all other parameters, such as F_v of all other nodes and logistic model parameters W is fixed. Similarly, when updating parameters W of the logistic model, the F is fixed. Hence, an iteratively updating approach is adopted here and the stopping criteria is the difference of the likelihood is below certain pre-defined value.

2.2.3. Experiments

(1)*Datasets*: To evaluate the CESNA, the authors considered 5 real-world datasets with ground truth labels.

(2)*Baselines*: They also consider three different types of baselines. The first baselines consider only topology [12, 11]. The second baselines consider only attributes [13]. The third baselines consider both topology and node attributes [10, 14, 15, 16].

(3)*Evaluation metrics*: The idea is to quantify the agreement between communities C^* with ground truth and communities C generated from CESNA. The larger the value is, the better recovery of communities with ground truth.

(4)*Experiment results*: CESNA yields the best performance in 8 out 10 testing cases of recovery of communities with ground truth.

3. Survey on community detection in attributed networks

Attributed network community detection algorithms can be classified into three main categories based on their methodological principles:

(1)Similarity / Distance measurement-based approach: Topological information and vertex attribute are merged together into a global similarity / distance measurement. Then, any classical clustering algorithm can be further utilized.

(2)Augmented graph-based approach: The attribute information are served as additional topological information. Indeed, attribute information

can change the initial topology of the underlying input graph.

(3)Interaction model learning-based approach: The model-based approach formulates a joint modeling of the interplay between edge connections and vertex attributes and makes use of this model to compute the clustering. CESNA [1] falls into this category.

3.1. Similarity / Distance measurement-based approach

In order to integrate the attribute information in the clustering process, these methods define a similarity measure between node attributes and generate weights on existing edges. The similarity between nodes is determined by examining each of T attribute values they have in common. If original graph is weighted, then two weights can be merged. Then any unsupervised clustering algorithm for weighted graphs can be applied. In this way, the communities returned represent clusters of nodes that are not only well connected but also similar.

We will report in the following the main works adopting this strategy. In [17], the authors discussed three types of edge weighting methods: (1) clustering coefficient similarity, (2) common neighbor similarity, and (3) node attribute similarity. From these methods, (1) and (2) are considered as topological metrics and (3) is considered as attributes metric. The authors in [18] applied matching coefficient to measure the similarity based on attributes by counting the number of attribute values they have in common. Objects that are not directly related by an edge in the graph is treated with similarity of 0 regardless of their attribute values.

Pros: This category of methods usually is easy to understand. What is more, we already have many well-established clustering methods to support the second procedure in this class.

Cons: (1) This category of methods takes into consideration only vertices that are directly connected. Vertices that are not directly connected in the graph have similarity 0 regardless of their attribute values [18]. Hence, this type of method is expected to perform poorly when the graph is expected to have certain missing links or erroneous links. (2) In order to unify two types of distance methods, the aggregation function needs to be hand-designed. One aggregation function usually is only applicable to certain applications.

3.2. Augmented graph-based approach

This kind of approaches seeks to combine the topological structure and the attribute information through an augmented graph. The initial topological structure of the original graph is augmented by new vertices called attribute vertices and new edges called attribute edges. In the augmented graph, two vertices are close if they are connected through many other original vertices, or if they share many common attribute vertices as neighbors. Once the augmented graph is created, distance measure which estimate the pairwise vertex closeness is defined to unify the distance computation on the augmented graph. Then conventional clustering methods can be applied to detect communities.

Authors in [19, 20] proposed SA-Cluster algorithm that use the neighborhood random walk distance to compute a unified distance between vertices

on the augmented graph. Random walk distance here serves as a measurement of similarity. Next, the k-Medoids clustering method is applied for community detection.

Pros: This class of methods can iteratively refine the clusters.

Cons: This class of methods usually are limited to small networks with few attribute values. The scalability is poor when the graph or the attribute size is large.

3.3. Interaction model learning-based approach

The model-based approach formulates joint modeling of the interplay between the edge connections and vertex attributes and makes use of this model to compute the clustering.

In [21], authors developed a Bayesian probabilistic model for attributed graphs denoted BAGC, and then formulated the clustering problem as standard probabilistic inference problem to find the clustering that gives the highest probability.

Pros: This class of methods does not need to hand design the similarity functions.

Cons: A learning objective needs to be carefully designed and usually applicable to specific applications. The learning process can take considerable time and make it not suitable for online-manner tasks.

4. Attribute-driven Backbone Discovery

In this section, we study a new problem called *attributed-driven backbone discovery* (ABD), characterized as follows.

- **Input**: Attributed graph G, a set of interested nodes V_I in G, an interestingness function I, and an edge cost model C that assigns a weight to an edge given its affinitive attributes;
- **Output**: an attributed backbone T that connects the nodes in V_I with maximized interestingness in terms of I, and minimized edge cost in terms of C.

We propose two strategies to solve ABD problems:

(1) When cost model is given, we build the connection between ABD problem and the Prize-collecting Steiner Tree problem [22] to give an approximation algorithm.

(2) When cost model is not given, we utilize the interaction modeling idea similar to CESNA [1] problem to given an EM-learning based heuristic algorithm.

(3) We further propose to apply Graph Neural Networks methods to detect the error in the given attributed graph and clean the data.

5. Conclusion

In this paper,

(1) we give the summary of one related paper CESNA [1] that infers community structures by statistically modeling the interaction between the

network structure and the node attributes. From the experimental results, we show the necessity to consider both community structure and node attributes. We identify one major limitation of CESNA, that in this kind of representation, it is hard to explain the connection between node pairs if there exists. Also, extensive experiments are conducted to show that CESNA cannot scale well when the graph size is large. For instance, in Youtube dataset with 2,509,862 edges, 129,907 nodes, and 8 attributes. CESNA cannot finish the generative model learning after 2 hours passed.

(2) We formulate attribute-driven backbones to characterize the connectivity patterns among the nodes of interests. We develop a bi-criteria cost model to characterize good backbones that incorporates both node interestingness and edge cost. This measure is determined by the selection of node attributes and topology of the backbones. In this case, the interested nodes are maximally included in the attributed backbones and the connection is explainable.

References

[1] J. Yang, J. McAuley, J. Leskovec, Community detection in networks with node attributes, in: Data Mining (ICDM), 2013 IEEE 13th international conference on, IEEE, pp. 1151–1156.

[2] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks, Journal of statistical mechanics: theory and experiment 2008 (2008) P10008.

[3] S. Fortunato, D. Hric, Community detection in networks: A user guide, Physics Reports 659 (2016) 1–44.

[4] M. Girvan, M. E. Newman, Community structure in social and biological networks, Proceedings of the national academy of sciences 99 (2002) 7821–7826.

[5] J. Xie, S. Kelley, B. K. Szymanski, Overlapping community detection in networks: The state-of-the-art and comparative study, Acm computing surveys (csur) 45 (2013) 43.

[6] M. E. Newman, M. Girvan, Finding and evaluating community structure in networks, Physical review E 69 (2004) 026113.

[7] A. Clauset, M. E. Newman, C. Moore, Finding community structure in very large networks, Physical review E 70 (2004) 066111.

[8] P. Pons, M. Latapy, Computing communities in large networks using random walks, in: International symposium on computer and information sciences, Springer, pp. 284–293.

[9] T. La Fond, J. Neville, Randomization tests for distinguishing social influence and homophily effects, in: Proceedings of the 19th international conference on World wide web, ACM, pp. 601–610.

[10] R. Balasubramanyan, W. W. Cohen, Block-lda: Jointly modeling entity-annotated text and entity-entity links, in: Proceedings of the 2011 SIAM International Conference on Data Mining, SIAM, pp. 450–461.

[11] J. Yang, J. Leskovec, Overlapping community detection at scale: a nonnegative matrix factorization approach, in: Proceedings of the sixth ACM international conference on Web search and data mining, ACM, pp. 587–596.

[12] M. Coscia, G. Rossetti, F. Giannotti, D. Pedreschi, Demon: a local-first discovery method for overlapping communities, in: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, pp. 615–623.

[13] A. P. Streich, M. Frank, D. Basin, J. M. Buhmann, Multi-assignment clustering for boolean data, in: Proceedings of the 26th annual international conference on machine learning, ACM, pp. 969–976.

[14] Y. Ruan, D. Fuhry, S. Parthasarathy, Efficient community detection in large networks using content and links, in: Proceedings of the 22nd international conference on World Wide Web, ACM, pp. 1089–1098.

[15] S. Günnemann, B. Boden, I. Färber, T. Seidl, Efficient mining of combined subspace and subgraph clusters in graphs with feature vectors, in: Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer, pp. 261–275.

[16] J. Leskovec, J. J. Mcauley, Learning to discover social circles in ego networks, in: Advances in neural information processing systems, pp. 539–547.

[17] K. Steinhaeuser, N. V. Chawla, Community detection in a large real-world social network (2008) 168–175.

[18] J. Neville, M. Adler, D. Jensen, Clustering relational data using attribute and link information, in: Proceedings of the text mining and link analysis workshop, 18th international joint conference on artificial intelligence, San Francisco, CA: Morgan Kaufmann Publishers, pp. 9–15.

[19] Y. Zhou, H. Cheng, J. X. Yu, Clustering large attributed graphs: An efficient incremental approach, in: Data Mining (ICDM), 2010 IEEE 10th International Conference on, IEEE, pp. 689–698.

[20] H. Cheng, Y. Zhou, J. X. Yu, Clustering large attributed graphs: A balance between structural and attribute similarities, ACM Transactions on Knowledge Discovery from Data (TKDD) 5 (2011) 12.

[21] Z. Xu, Y. Ke, Y. Wang, H. Cheng, J. Cheng, A model-based approach to attributed graph clustering, in: Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data, pp. 505–516.

[22] M. X. Goemans, D. P. Williamson, A general approximation technique for constrained forest problems, SIAM Journal on Computing 24 (1995) 296–317.