Community detection in Attributed Network

Issam Falih
Paris 13 University
Villetaneuse, FRANCE
issam.falih@lipn.univ-paris13.fr

Nistor Grozavu
Paris 13 University
Villetaneuse, FRANCE
nistor.grozavu@lipn.univ-paris13.fr

Rushed Kanawati
Paris 13 University
Villetaneuse, FRANCE
rk@lipn.univ-paris13.fr

Younès Bennani
Paris 13 University
Villetaneuse, FRANCE
younes.bennani@lipn.univ-paris13.fr

ABSTRACT

Graph clustering techniques are very useful for detecting densely connected groups in large graphs. Many existing graph clustering methods mainly focus on the topological structure, but ignore the vertex properties. Existing graph clustering methods have been recently extended to deal with nodes attribute. First we motivate the interest in the study of this issue. Then we review the main approaches proposed to deal with this problem. We propose a comparative study of some existing attributed network community detection algorithm on both synthetic data and on real world data.

CCS CONCEPTS

• **Theory of computation → Unsupervised learning and clustering**; **Social networks**;

KEYWORDS

Attributed Network; Social Network Analysis; Community detection; Clustering

ACM Reference Format:
Issam Falih, Nistor Grozavu, Rushed Kanawati, and Younès Bennani. 2018. Community detection in Attributed Network. In *WWW '18 Companion: The 2018 Web Conference Companion, April 23–27, 2018, Lyon, France.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3184558.3191570

1 INTRODUCTION

In many application, real-world graph data are often associated with additional information, i.e. vertices of a graph are associated with a number of attribute that describe the vertex. Indeed, there are two sources of data that can be used to perform the community detection task. The first one is the data about the nodes and their attributes and the second source of data comes from the different kind of connections among vertices. For example, in social networks, edge attributes represents the relationship (friendship, collaboration, family, etc) among people while vertex attribute describe the role or the personality of a person. An other example, is the bibliographical network, a vertex may represent an author and vertex properties describe attributes or features of the author as the area of interest, the number of publications, while the topological structure represent relationships among authors.

Thus, it is important to consider both sources of information simultaneously and consider network communities as sets of nodes that are densely connected, but which also share some common attributes. Node attributes can complement the network structure, leading to more precise detection of communities; additionally, if one source of information is missing or is noisy, the other will be used. However, considering both node attributes and network topology for community detection is also challenging, as the approach have to combine two types of information [33]. Recently, only few recent studies have addressed the problem of clustering in attributed networks [24]; [4]; [21]. The problem is quite challenging because it is based on how to adjust the degree of contributions of topological and attribute information.

We summarize the main contributions of this paper as follows:

- We classify the existing methods that deals with attributed clustering problem into three different main approaches.
- We compare a set of attributed network community detection algorithm on a collection of synthetic data and real data. Experimental results show that algorithms combining both types of information successfully groups vertices into meaningful clusters.

To explore this task, this paper is organized as follow. First, we introduce, in section 2, the attributed network clustering problem. We give a classification of existing state-of-art methods that deals with this problem. Section 3 provides the experimental evaluation of some existing attributed network community detection algorithm. Finally, Section 4 concludes the present article.

2 ATTRIBUTED NETWORK CLUSTERING

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.

Definition. An attributed graph G is defined as a 4-tuple ($\mathcal{V}$, E, $\mathcal{A}$, F), where $\mathcal{V} = \{v_1, v_2, ..., v_n\}$ is a set of n vertices, $E = \{(u, v) : u, v \in |V|, u \neq v\}$ is a set of edges, $\mathcal{A} = \{a_1, a_2, ..., a_T\}$ is a set of T attributes, $F = \{f_1, f_2, ..., f_T\}$ is a set of T attributes functions

WWW '18 Companion, April 23–27, 2018, Lyon, France

ACM ISBN 978-1-4503-5640-4/18/04.
https://doi.org/10.1145/3184558.3191570

and each function $f_t : \mathcal{V} \rightarrow dom(a_t)$ assigns to each vertex in $\mathcal{V}$ 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 $v \in \mathcal{V}$ is essentially associated with an attribute vector of length T, where the element t in the vector is given by the function $f_t(v)$. Table 1 gives the principal notations used in this paper.

Notation	Description
$\mathcal{V}$	Set of vertices $\mathcal{V} = \{v_i : 1 \leq i \leq \lvert\mathcal{V}\rvert\}$
$\mathcal{A}$	Set of attributes $\mathcal{A} = \{a_i : 1 \leq i \leq \lvert\mathcal{A}\rvert\}$
$\mathcal{P}$	Partition where $P = \{C_1, C_2, ..., C_k\}$
(u, v)	Edge between vertex u and vertex v
m	Number of the edges in the network
d_v	Degree of the vertex v
$f_{a_t}(v)$	Function that return the attribute value of the vertex v

Table 1: Attributed networks: Notations

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

(1) $C_i \cap C_j = \emptyset \; \forall i \neq j$ and $\cup_i C_i = |V|$
(2) Vertices within clusters are densely connected, while the vertices in different clusters are sparsely connected.
(3) Nodes in the same clusters are expected to have homogeneous attributes.

Traditional clustering for vector data evaluates clusters w.r.t. all attributes; they do not deal with graph data where the object are nodes connected to each other through edges like K-means. On the other hand, the well-known graph clustering techniques, as clustering based on normalized cut [26], modularity [23] use the relationships of the network to partition the graph into several densely connected components, but do not use the properties of the nodes. The problem is to apply clustering approaches that use graph data and attribute data simultaneously in order to detect clusters that are densely connected in the graph and at the same time to use the similarity in the attribute space. Few recent studies have addressed the problem of clustering in attributed networks.

Community detection in attributed networks identify clusters either in the full space of the network or in multiple sub-spaces [1] [3], [14] [13]. In this paper, we focus on the case of attributed network community detection algorithms in the full space.

Representative approaches of attributed network community detection algorithms can be classified into three main categories based on their methodological principles:

- **Topological-based clustering :** The attribute information are used as additional topological information. Indeed, the attribute information can be used in order to change the initial topology of the input graph.
- **Attributed-based clustering :** Topological information is merged together with vertex attribute into a global similarity/distance which can then be processed by any classical clustering algorithm.
- **Hybrid approach :** attributes and topological information are considered separately. For example, it could be done by computing clusters using only topological methods and using only vectorial clustering approaches. Then the results are merged by an ensemble clustering method.

2.1 Topological-based approach

The basic idea of this class of approaches is to transform the problem of attributed network clustering into a topological clustering problem. Node's attribute are used as *additional* topological information. It will be used to change the initial topology structure of the input graph. Next, we present different ways to consider the node's attribute information.

2.1.1 Edge weighting based approaches. In order to integrate the attribute information in the clustering process, these methods define a similarity measure between node attributes that will be used to weight the existing edges. The similarity between nodes is determined by examining each of T attribute values they have in common. Then any unsupervised clustering algorithm for weighted graphs can be applied. The values of weights will influence the clustering algorithm to privilege the creation of groups in which the nodes are not only well connected but also similar. Algorithm 1 shows the principal outlines of this approach.

Algorithm 1 Edge-weighting based approach

Require: $\mathcal{G}(\mathcal{V}, E, \mathcal{A}, F)$: attributed graph.
$\mathcal{S}$: similarity function.
$clustAlgo_w$: clustering algorithm for weighted graph.
Ensure: Partition of $\mathcal{V}$.
$G_w = (\mathcal{V}, E, w)$; $w : \mathcal{V} \times \mathcal{V} \rightarrow \mathbb{R}$
for $(u, v) \in E(G_w)$ **do**
$w(u, v) = S(f_{1..T}(u), f_{1..T}(v))$
end for
$\mathcal{P} = clustAlgo_w(G_w)$
return $\mathcal{P}$

We report in the following the main works adopting this strategy. In [20], authors propose the *matching coefficient* similarity function that consists on counting, for two connected vertices, the number of attribute values they have in common. Formally, the *matching coefficient* over two vertices (u, v) is given by :

$$\mathcal{S}(u, v) = \begin{cases} \sum_{t=1}^{T} s_{a_t}(u, v) & if (u, v) \in E \quad or \quad (v, u) \in E \\ 0 & otherwise \end{cases} \tag{1}$$

Where

$$s_{a_t}(u, v) = \begin{cases} 1 & if f_t(u) = f_t(v) \\ 0 & otherwise \end{cases}$$

Once the weights are changed, the authors use a classical unsupervised learning algorithm as Karger's Min-Cut [15], spectral clustering [29] on the weighted adjacency matrix. Community detection algorithm that deals with weighted graph as Louvain [2] can also be used. Initially, the *matching coefficient* similarity measure deals only with categorical attribute. It was extended by [27] to handle, at the same time, categorical and continuous attributes. For continuous attributes, first each attribute is normalized to the range

of [0, 1] by adding a normalizing parameter denoted α and then, the arithmetic difference between the pairs of attribute value is used to obtain a similarity score. This similarity metric is expressed as :

$$s'_{a_t}(u,v) = \begin{cases} 1 & if \quad a_t \quad categorical \quad \& \quad f_t(u) = f_t(v) \\ 1 - \alpha_t |f_t(u) = f_t(v)| & if \quad a_t \quad continuous \\ 0 & otherwise \end{cases}$$

Where α_t is a normalizing parameter that corresponds to the attribute a_t. It allows to normalize the attribute value of a_t in the range of [0, 1]. Edge weighting based approaches produces a new edge weights according to node attribute similarity. If the original graph is weighted, the two weight can be merged. This family of techniques are simple to implement but their disadvantage is that they take in consideration only vertices that are directly connected. Vertices that are not directly connected in the graph have a similarity equal to zero regardless to their attribute value.

2.1.2 Augmented graph based approaches. 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*. An *attribute vertex* $v_{a_{ti}}$ represents an attribute-value pair (a_t, a_{ti}) where $a_{ti} \in dom(a_t)$ is a value of the attribute a_t. If a vertex v has the value a_{ti} on the attribute a_t, an *attribute edge* is added between the vertex v and the *attribute vertex* $v_{a_{ti}}$. With such graph augmentation, the attribute similarity is expressed as vertex neighborhood in the augmented graph: two vertices which share the same attribute value are connected by a common *attribute vertex*. Since each vertex v_i has T attribute values, there are totally $|\mathcal{V}| \times T$ attribute edges added to the original graph. 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 or a community detection algorithm can be applied to find out the set of clusters. Next, in algorithm 2, we present in a formal way the principle of attributed network algorithm based on augmented graph.

Algorithm 2 Augmented graph based approach

Require:
$\mathcal{G}(\mathcal{V}, E, \mathcal{A}, F)$: attributed graph.
clust : clustering algorithm
Ensure: A partition of $\mathcal{V}$.
1: $\mathcal{V}' = \mathcal{V} \cup \mathcal{V}_a; \mathcal{V}_a = \{(a_t, a_{ti})\}$ with $t \in \{1..T\}$ and $i \in dom(a_t)$
2: $E' = E \cup E_a$ with $E_a \subseteq \mathcal{V} \times \mathcal{V}_a$
3: $G' = (\mathcal{V}', E')$
4: $\mathcal{P} = clust(G')$
5: **return** $\mathcal{P}$

Authors in [34], [5] propose SA-Cluster algorithm that use the neighborhood random walk distance to compute a unified distance between vertices on the augmented graph. The random walk distance between two vertices is based on the paths consisting of both structure and attribute edges. In this way, it combines the structural closeness and attribute similarity through the random walk distance measure. Then, the random walk distance is used as pairwise similarity measure in the clustering process by K-Medoids clustering approach to partition the graph into k clusters. Approaches based on augmented graph can handle only categorical attribute but it can be easily extended to handle both categorical and continuous attribute. Usually, for continuous attribute, the values are transformed in different intervals value. The disadvantage of these approaches is that they are limited to small networks with few attribute values.

2.1.3 Quality function optimization based approaches. This family of approaches extend the well-know graph based methods to consider both attributes information and topological structure. The existing approaches mainly extend the Louvain algorithm [2] as linear combination of the Newman [22] modularity and new measure that computes the attribute similarity.

Cruz & al. [8] include the entropy optimization as an intermediate step between modularity optimization and community aggregation. This is done to minimize semantic disorder of nodes by moving nodes among the clusters found during the modularity optimization. These steps are iterated until the modularity is not longer improved. In [9], authors propose an extension of the Louvain algorithm [2] with a modification of modularity by including the similarity of the attributes given by:

$$Q^+ = \sum_{C_i \in P} \sum_{v,u \in C_i} \alpha.[\frac{1}{2m}(A_{vu} - \lambda \frac{d_v d_u}{2m})] + (1-\alpha).S(u,v)$$

Where $S(v,u)$ is a similarity function based on type of attributes of v and u and it can be adapted according to how the attributes are represented. $\alpha \in [0,1]$ is a weighting factor which represents the degree of contribution of structural and attribute information.

Another extension of Louvain is proposed by [7] called ILouvain algorithm which uses the inertia based modularity combined with the Newman's modularity.

Modularity optimization approaches make assumption that the best partition of a graph is the one that maximizes the modularity, but [12], [16] have shown that this assumption can not be satisfied if the modularity is not a pertinent measure for some graphs.

2.2 Attribute-based approach

Unlike the topological based approaches which aims to find dense connected subgraph in the network using topological information, attribute-based approach computes a distance matrix or a dissimilarity matrix between all pair of nodes. Next, we present different ways to consider the topological structure information.

2.2.1 Unified distance based approaches . This kind of approach transforms the topological information of the network into a similarity or a distance function between vertices. Generally this distance is defined as a linear combination between a structural distance function and node attribute distance. Once this function is defined, classical distance-based clustering methods can be applied. Formally we have:

$$dis(u,v) = \alpha dis_T(d_u, d_v) + (1-\alpha) dis_S(u,v) \quad (2)$$

Where :

- $dis_T(d_u, d_v)$: represents the topological distance between vertices u and v. Different topological distance can be used as the shortest path, the neighborhood random walk distance, etc.
- $dis_S(d_u, d_v)$: is an attribute distance between vertices u and v.
- $\alpha \in [0,1]$: is a parameter introduced to control the influence of both similarity aspects. The importance of structure and context similarity is variant and depends on the application domain. Therefore, the choice of appropriate value for this parameter is critical. For instance, social networks often exhibit dense regions and follow power-law degree distribution. The higher value for α seems effective for these networks because nodes in dense regions are expected to have similar attributes. However, non-scale free networks, e.g. road networks, need to be treated with a balanced ratio.

As an example, [6], define a unified distance as a linear combination of two distances, each corresponding to a type of data: cosine distance on textual information and geodesic distance on the network structure. Then a hierarchical agglomerative clustering is applied with the unified distance matrix. Another similar unified distance function is proposed by [9] in SAC Algorithm that will be used to build a k-nearest neighbor graph. Communities will be found using the Louvain algorithm.

Authors in [11] proposed an other method of considering the topological structure of the network namely *ANCA*. First, they select a set of qualified nodes called *Seeds* that are landmark in the network that will be used to characterize the set of nodes. The position of each vertex in the graph will be characterized by its relations with the seed nodes. Once *seeds* are selected, topological feature i.e. distance will be used to characterize the relation between *seeds* and all nodes of the network. Then, using a weighting factor, merge topological structure features and attribute information. Encapsulating topological information in attributes enriches the dimension space. To this end, authors use the spectral clustering techniques in order to find communities.

2.3 Hybrid approach

This class of approaches consider the attribute information and topological structure separately. Next, we briefly explain the following methods that deals with this kind of approaches.

2.3.1 Ensemble/selection based approaches. These approaches consist to combine the result of clustering using different methods of clustering. *Ensemble methods* can be used to combine the found partitions. For instance, HyperGraph Partitioning Algorithm (HGPA) [28] where cluster ensemble problem is posed as a partitioning problem of a hypergraph by cutting a minimal number of hyper edges, approximates the maximum mutual information objective with minimum cut objective constrains. Or, Cluster-based Similarity Partitioning Algorithm (CSPA)[28] where binary similarity matrix is used to signify relationship between objects in the same cluster in order to establish a pairwise similarity measure that yield a combined clustering.

Authors in [10] combine the result of a topological clustering algorithm as Louvain [2], Licod [32] with the clustering results on attributes i.e. K-means method.

On the other hand, [18] propose to merge 4 models scheme. They combine a topological clustering algorithm, an attributed clustering algorithm, an attribute based approach and the GAMER algorithm [14].

2.3.2 Probabilistic model based approaches. 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.

Xu & al. [30] developed a Bayesian probabilistic model for attributed graphs denoted BAGC, and then formulate the clustering problem as standard probabilistic inference problem to find the clustering that gives the highest probability. The probabilistic model essentially defines a joint probability distribution over the space of all possible clustering and all possible attributed graphs. For a given attributed graph to be clustered, the model assigns a probability for each possible clustering of the vertices. The cluster label of each vertex is represented as a hidden variable. The model enforces the intra-cluster similarity by asserting that the attribute values and edge connections of a vertex should depend on its cluster label. In particular, attribute values and edge connections for vertices within the same cluster should follow a common distributions that are specific to that cluster. BAGC starts with a random assignment of the vertices into clusters. Then, the parameters of all the distributions are iteratively recalculated. More formally, given an attributed graph G defined by its adjacency matrix M_{adj}, its attributes matrix $M_{\mathcal{A}}$ and vector Z contains the assigning of a node to a cluster. This model produces a conjoint probability $p(M_{adj}, M_A, Z)$ and find a partition Z^* such that:

$$Z^* = argmax_Z \quad p(M_{adj}, M_A, Z) \quad (3)$$

The BAGC model has been extended recently to handle weighted attributed graphs [31]. CESNA [33] defines a model on attributed graphs that also enforces the intra-cluster similarities. CESNA models vertex attributes and connections in the same cluster with Bernoulli distributions. CESNA differs from BAGC by identifying overlapping communities.

3 EXPERIMENTATION

In this section, we performed experiments to compare a set of attributed network community detection algorithm. These state-of-art algorithms are *SA-Cluster* [34], *SAC* [9], *IGC-CSM* [19], *NAS* [27], ILouvain [7], *ANCA* [11]. We have also developed the algorithms cited earlier from scratch using open source graph library called *igraph* in R for experimental analysis.

Others comparative analysis have been carried out, where we have consumed the results of clustering this is the case of Ilouvain algorithm [7][1]. The source code of approaches, evaluation measure

[1] http://bit.ly/ILouvain

and the dataset used for the experiments in the paper are available for download[2] as R library.

3.1 Metrics for Evaluating Algorithm Quality

In this study, we use two groups of metrics to evaluate the performance of each algorithm. The first group includes NMI (Normalized Mutual Information) and ARI (Adjusted Rand Index). These are usually used to evaluate a clustering result when the ground-truth decomposition into clusters are known. High values indicates better algorithm performance. The other group consists of *Modularity*, *Density*, *Conductance* and *Entropy*. *Modularity*, *Density* and *Conductance* is used to measure the quality of communities in a network, and a larger value indicates better partition quality. *Entropy* is used to measure the degree of attribute consistency in a community, and a lower *Entropy* value indicates a greater consistency. These metrics are often used when a algorithm is run on a network without ground truth. Formal definition are provided below:

- **Density :** Strong connection among vertices is analyzed by using the density function which represents the ratio between number of edges presented in the clusters and total number of edges in the whole graph. The ratios get accumulated for all clusters to evaluate the overall impact. Density values lie in the interval of $[0, 1]$.

$$\delta(\{C\}_{i=1}^{k}) = \frac{1}{\| E \|} \sum_{i=1}^{k} \| E(C_i) \|$$

- **Entropy :** One of the key aspects to measure the quality of a clustering results is to determine the relevancy among vertices based upon their attributed nature. For each attribute the entropy, in Eq. 3.1, is calculated against each cluster with associated attribute. When all the vertices inside the same cluster are having similar attributes or contexts associated with them, then overall entropy acquires minimum value.

$$entropy(a_t) = \sum_{i=1}^{k} \frac{\| C_i \|}{\| \mathcal{V} \|} \quad entropy(a_t, C_k) \tag{4}$$

$$entropy(a_t, C_k) = - \sum_{s=1}^{\|dom(a_t)\|} p_{ks}^{t} \log p_{ks}^{t}$$

where p_{ks}^{t} is the fraction of vertices in cluster C_k that take the value s where $s \in dom(a_t)$

- **Modularity :** The modularity is the number of edges falling within clusters minus the expected number in an equivalent network with edges placed at random. The modularity can be either positive or negative, with positive values indicating the possible presence of community structure.

$$Q(\{C\}_{i=1}^{k}) = \frac{1}{2m} \sum_{i=1}^{k} \sum_{u,v \in C_i} \left(\left(A_{uv} - \lambda \frac{d_u d_v}{2m} \right) \right) \tag{5}$$

- **Conductance :** The conductance of a partition measures how well-knit the clusters are.

$$\Phi(\{C\}_{i=1}^{k}) = \frac{\phi(C_i)}{k}$$

[2] lipn.univ-paris13.fr/~falih/packages/ANCL/

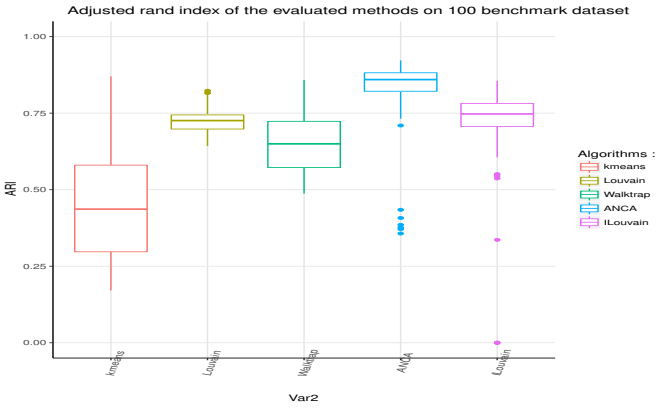

Figure 1: Cluster ARI quality comparison on 100 synthetic data

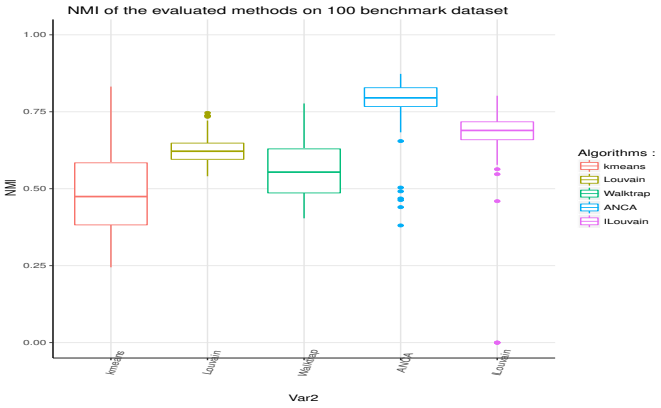

Figure 2: Cluster NMI quality comparison on 100 synthetic data

$$\phi(C_i) = \frac{\sum_{u \in C_i, v \in \bar{C}_i} A_{uv}}{min(\sum_{u \in C_i, v \in \mathcal{V}} A_{uv}, \sum_{u \in \bar{C}_i, v \in \mathcal{V}} A_{uv})}$$

3.2 Experimental Result on Synthetic Networks

In [17], Largeron and al. have provided a generator to generate networks with community structure and numerical nodes attribute. We use this generator to generate 100 attributed network of different shape for which the ground truth decomposition into communities is known. The proposed generator offer attributed network with only numerical nodes attribute value.

The communities were found using *ILouvain* algorithm, *ANCA* algorithm, the *Louvain* algorithm, *k-means* algorithm, since they cope with networks having numerical nodes attributes. Figures 1 and 2 present, respectively the adjusted rand index (ARI) and the normalize mutual information (NMI), result on 100 attributed networks.

The results confirm the interest of using both kind of information, as *ILouvain* and *ANCA* outperforms other approach that takes in

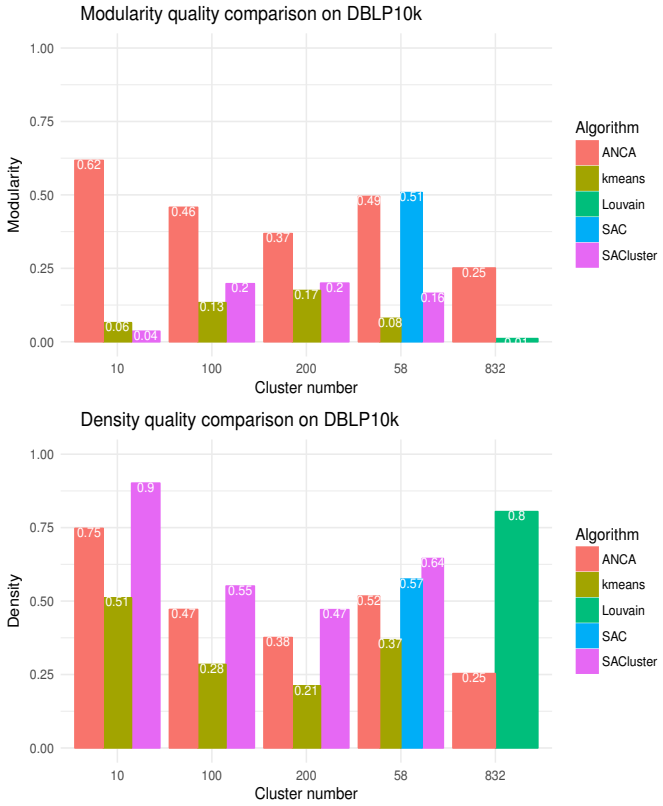

Figure 3: Modularity and Density quality comparison on DBLP10K data

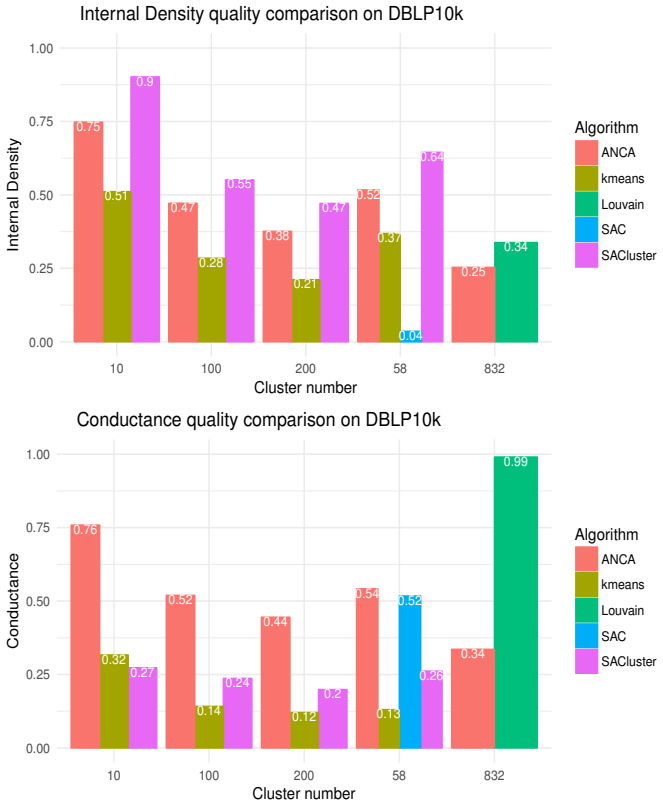

Figure 4: Conductance and internal Density quality comparison on DBLP10K data

consideration only one type of information. Compared to ILouvain, ANCA algorithm produces better results.

3.3 Experimental Result on Real Networks

In addition to our experiments using synthetic networks, we tested the algorithms on two real networks. The nodes were associated with categorical attributes. Detailed information on these data sets is described below.

DBLP10K : is a co-author network extracted from the DBLP Bibliography. Each vertex represents a scholar and each edge represents a co-author relationship between two scholars. The dataset contains 10, 000 scholars who have published in major conferences. Each scholar is associated with two attributes, prolific and primary topic. The attribute "prolific" has three values: "highly prolific" for the scholars with ≥ 20 publications, "prolific" for the scholars with ≥ 10 and < 20 publications, and "low prolific" for the scholars with < 10 publications. The domain of the attribute "primary topic" consists of 100 research topics. Each scholar is then assigned a primary topic out of the 100 topics. This datasets was given by [19].

Emails : due to emails privacy issues, there is no public corpus from a real organization available except for a huge Anonymized Enron email corpus [25]. It contains vast collection of emails covering a time span of 41 months, and also uniquely depicts the ups and downs of the energy giant Enron. It provides an opportunity to determine related mailbox users based on their unique communication and relationship in the email network. We have considered the Enron email data set 5, which contains all the emails of 161 users, managed separately, to infer the community structure from partial information available in terms of personalized emails. Two users, Sally Beck and Louise Kitchen, email network is exploited from this data set for all the experiments in this paper to infer the community structure. The email interactions with the individuals outside the Enron Corporation are explicitly ignored to reflect factual associations.

For real world data, usually the ground truth is not available. However, we can use topological measure as modularity, conductance and attribute measure as entropy to compare the clusters quality. Thus, we analyze next different properties of the clustering results determined by ANCA, Kmeans, Louvain, SAC, SACluster.

SA-Cluster, k-means, ANCA need the number of cluster as input. We set the number of cluster $k = 10, 100, 200$ for DBLP10k dataset and $k = 10, 20, 30, 50$ for Emails dataset. We compare also with methods that are topological only approaches as Louvain. It considers only the topological structure of the network. We added also to k the number of communities found by algorithms that don't need the number of cluster as input in order to compare with it.

The experimental results of the methods on *DBLP10k* and *Emails* are shown in Figure 3, 4, 5 and 6.

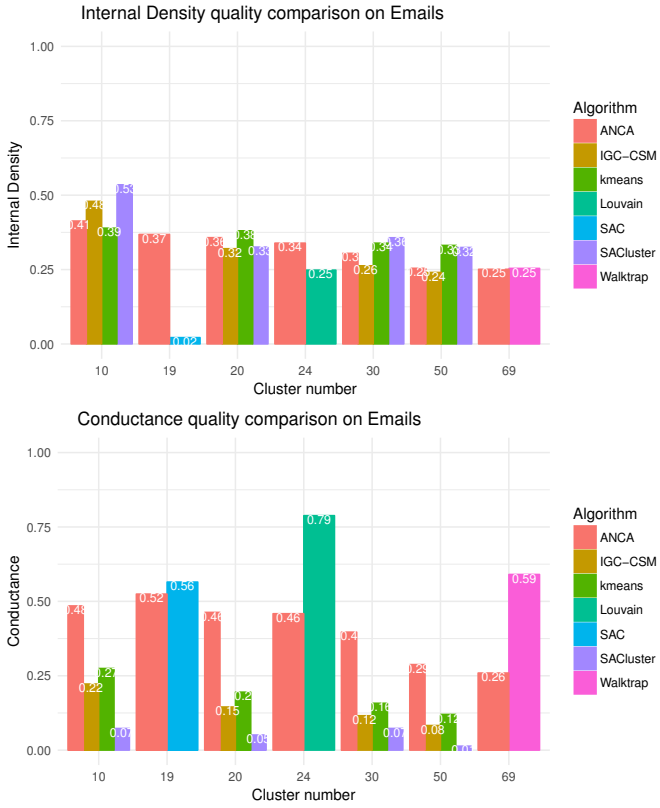

Figure 6: Conductance and internal Density quality comparison on Emails data

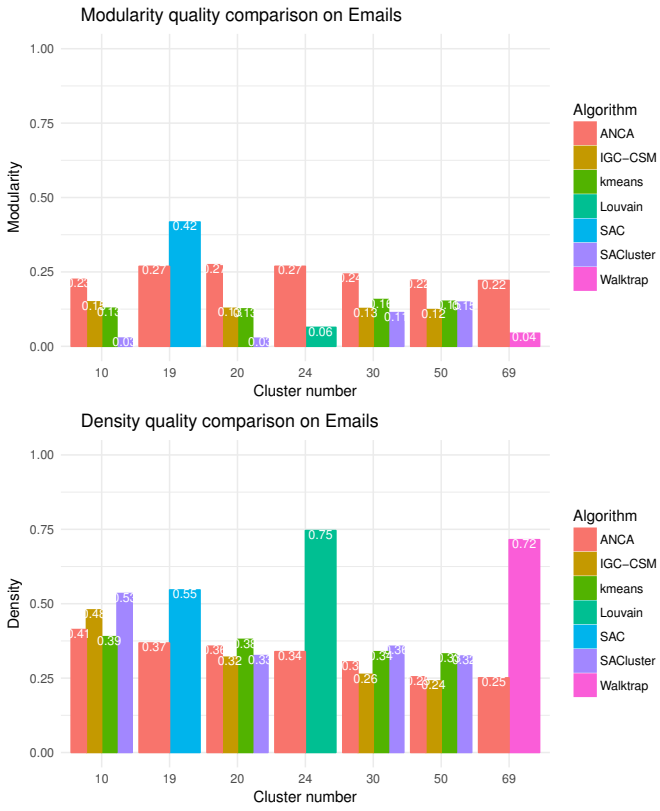

Figure 5: Modularity and Density quality comparison on Emails data

The best methods should preserve the dense connectivity among vertices in the original graph and low entropy values simultaneously. Density, conductance and modularity measures the connectivity around vertices. For ANCA these topological quality measures are correlated and they decrease when k increase. The density values show that the clusters found by SA-Cluster are more dense that then those found by ANCA. However, the modularity and the conductance values by SA-Cluster gives an opposite view. By analyzing at the clusters distribution, SA-Cluster find a large cluster and few vertices in other clusters. From the data in Figure, we have concluded that adding vertices attribute promotes the performance of community detection in most cases. Taking the results of ANCA as an example, most of the result were better than those of the basic Louvain Algorithm on topological structure and k-means on nodes attribute.

4 CONCLUSION

In this paper we provide an overview of the emerging topic of clustering in attributed graph. Only few works have been proposed in the literature and their aim is to partition attributed graphs into dense clusters with vertices having similar attributes. The existing methods can be differentiate into three main approaches based on the manner in which the topological and attribute data are considered. We then compare a set of community detection algorithm in attributed network on both artificial data and real world data.

REFERENCES

[1] Martin Atzmueller. 2015. Subgroup and Community Analytics on Attributed Graphs.. In *SNAFCA@ ICFCA*.
[2] Vincent D Blondel, Jean-loup Guillaume, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* 2008 (2008), P10008. arXiv:arXiv:0803.0476v2
[3] Brigitte Boden. 2014. *Combined clustering of graph and attribute data.* Apprimus Wissenschaftsver.
[4] Cecile Bothorel, Juan David Cruz, Matteo Magnani, and Barbora Micenková. 2015. Clustering attributed graphs: Models, measures and methods. *Network Science* January (2015), 1–37. https://doi.org/10.1017/nws.2015.9 arXiv:1501.0167
[5] Hong Cheng, Yang Zhou, and Jeffrey Xu Yu. 2011. Clustering Large Attributed Graphs: A Balance between Structural and Attribute Similarities. *ACM Trans. Knowl. Discov. Data* 5, 2 (2011), 12:1—-12:33. https://doi.org/10.1145/1921632.1921638
[6] David Combe, Christine Largeron, El\H Od Egyed-Zsigmond, and Mathias Géry. 2012. Combining relations and text in scientific network clustering. *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (2012), 1280–1285. https://doi.org/10.1109/ASONAM.2012.215
[7] David Combe, Christine Largeron, Mathias Géry, and Előd Egyed-Zsigmond. 2015. I-Louvain: An Attributed Graph Clustering Method. In *Advances in Intelligent Data Analysis XIV*. Springer, 181–192.
[8] Juan David Cruz, Cécile Bothorel, and François Poulet. 2011. Entropy based community detection in augmented social networks. In *Computational aspects of social networks (cason), 2011 international conference on*. IEEE, 163–168.
[9] TA Dang and E Viennet. 2012. Community detection based on structural and attribute similarities. In *International Conference on Digital Society (ICDS)*. 7–12.
[10] Haithum Elhadi and Gady Agam. 2013. Structure and attributes community detection: comparative analysis of composite, ensemble and selection methods. *Proceedings of the 7th Workshop on Social Network Mining and Analysis* 13 (2013), 10:1—-10:7. https://doi.org/10.1145/2501025.2501034

[11] Issam Falih, Nistor Grozavu, Rushed Kanawati, and Younès Bennani. 2017. ANCA : Attributed Network Clustering Algorithm. In *Complex Networks & Their Applications VI - Proceedings of Complex Networks 2017 (The Sixth International Conference on Complex Networks and Their Applications), COMPLEX NETWORKS 2017, Lyon, France, November 29 - December 1, 2017. (Studies in Computational Intelligence)*, Chantal Cherifi, Hocine Cherifi, Márton Karsai, and Mirco Musolesi (Eds.), Vol. 689. Springer, 241–252. https://doi.org/10.1007/978-3-319-72150-7_20

[12] B. H. Good, Y.-A. de Montjoye, and A. Clauset. 2010. The performance of modularity maximization in practical contexts. *Physical Review* E, 81 (2010), 046106.

[13] Stephan Günnemann, Brigitte Boden, Ines Färber, and Thomas Seidl. 2013. Efficient mining of combined subspace and subgraph clusters in graphs with feature vectors. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 261–275.

[14] Stephan Günnemann, Ines Farber, Brigitte Boden, and Thomas Seidl. 2010. Subspace clustering meets dense subgraph mining: A synthesis of two paradigms. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*. IEEE, 845–850.

[15] David R. Karger. 1993. Global Min-cuts in RNC, and Other Ramifications of a Simple Min-out Algorithm. In *Proceedings of the Fourth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA '93)*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 21–30. http://dl.acm.org/citation.cfm?id=313559.313605

[16] Andrea Lancichinetti and Santo Fortunato. 2011. Limits of modularity maximization in community detection. *CoRR* abs/1107.1 (2011).

[17] Christine Largeron, Pierre-Nicolas Mougel, Reihaneh Rabbany, and Osmar R. Zaïane. 2015. Generating Attributed Networks with Communities. *Plos One* 10, 4 (2015), e0122777. https://doi.org/10.1371/journal.pone.0122777

[18] Nasif Muslim. 2016. A Combination Approach to Community Detection in Social Networks by Utilizing Structural and Attribute Data. (2016).

[19] Waqas Nawaz, Kifayat-Ullah Khan, Young-Koo Lee, and Sungyoung Lee. 2015. Intra graph clustering using collaborative similarity measure. *Distributed and Parallel Databases* (2015), 583–603. https://doi.org/10.1007/s10619-014-7170-x

[20] Jennifer Neville, Micah Adler, and David Jensen. 2003. 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*. 9–15.

[21] Mark EJ Newman and Aaron Clauset. 2016. Structure and inference in annotated networks. *Nature Communications* 7 (2016), 11863.

[22] M E J Newman. 2003. Mixing patterns in networks. *Physical review. E, Statistical, nonlinear, and soft matter physics* 67, 2 Pt 2 (2003), 026126. https://doi.org/10.1103/PhysRevE.67.026126 arXiv:cond-mat/0209450

[23] M. E. J. Newman and M. Girvan. 2004. Finding and evaluating community structure in networks. *Phys. Rev. E* 69, 2 (Feb. 2004), 026113. https://doi.org/10.1103/PhysRevE.69.026113

[24] Leto Peel, Daniel B. Larremore, and Aaron Clauset. 2017. The ground truth about metadata and community detection in networks. *Science Advances* 3, 5 (2017). https://doi.org/10.1126/sciadv.1602548 arXiv:http://advances.sciencemag.org/content/3/5/e1602548.full.pdf

[25] Jitesh Shetty and Jafar Adibi. 2004. The Enron email dataset database schema and brief statistical report. *Information sciences institute technical report, University of Southern California* 4, 1 (2004), 120–128.

[26] Jianbo Shi and Jitendra Malik. 2000. Normalized cuts and image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 22, 8 (2000), 888–905.

[27] Karsten Steinhaeuser and Nitesh V Chawla. 2008. Community detection in a large real-world social network. In *Social computing, behavioral modeling, and prediction*. Springer, 168–175.

[28] A. Strehl and J. Ghosh. 2003. Cluster ensembles: a knowledge reuse framework for combining multiple partitions. *The Journal of Machine Learning Research* 3 (2003), 583–617.

[29] Sebastian Thrun, Lawrence K Saul, and Bernhard Schölkopf (Eds.). 2004. *Advances in Neural Information Processing Systems 16 [Neural Information Processing Systems, NIPS 2003, December 8-13, 2003, Vancouver and Whistler, British Columbia, Canada]*. MIT Press.

[30] Zhiqiang Xu, Yiping Ke, Yi Wang, Hong Cheng, and James Cheng. 2012. A model-based approach to attributed graph clustering. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*. ACM, 505–516.

[31] Zhiqiang Xu, Yiping Ke, Yi Wang, Hong Cheng, and James Cheng. 2014. GBAGC: A General Bayesian Framework for Attributed Graph Clustering. *ACM Trans. Knowl. Discov. Data* 9, 1 (2014), 1–43. https://doi.org/10.1145/2629616

[32] Zied Yakoubi and Rushed Kanawati. 2014. Licod: Leader-driven approaches for community detection. *Vietnam Journal of Computer Science* 1, 4 (2014), 241–256.

[33] Jaewon Yang, Julian McAuley, and Jure Leskovec. 2013. Community detection in networks with node attributes. *Proceedings - IEEE International Conference on Data Mining, ICDM* (2013), 1151–1156. https://doi.org/10.1109/ICDM.2013.167 arXiv:1401.7267

[34] Yang Zhou, Hong Cheng, and Jeffrey Xu Yu. 2010. Clustering large attributed graphs: An efficient incremental approach. *Proceedings - IEEE International Conference on Data Mining, ICDM* (2010), 689–698. https://doi.org/10.1109/ICDM.2010.41