Combining relations and text in scientific network clustering

David Combe, Christine Largeron, Elöd Egyed-Zsigmond, Mathias Géry

Combining relations and text in scientific network clustering

David Combe*, Christine Largeron*, Előd Egyed-Zsigmond†, Mathias Géry*
*Université de Lyon, F-42023, Saint-Étienne, France,
CNRS, UMR 5516, Laboratoire Hubert Curien, F-42000, Saint-Étienne, France
Université de Saint-Étienne, Jean-Monnet, F-42000, Saint-Étienne, France
Email: {david.combe, christine.largeron, mathias.gery}@univ-st-etienne.fr
†Université de Lyon
UMR 5205 CNRS, LIRIS
7 av J. Capelle, F-69100 Villeurbanne, France
Email: elod.egyed-zsigmond@insa-lyon.fr

***Abstract*—In this paper, we present different combined clustering methods and we evaluate their performances and their results on a dataset with ground truth. This dataset, built from several sources, contains a scientific social network in which textual data is associated to each vertex and the classes are known. Indeed, while the clustering task is widely studied both in graph clustering and in non supervised learning, combined clustering which exploits simultaneously the relationships between the vertices and attributes describing them, is quite new. We argue that, depending on the kind of data we have and the type of results we want, the choice of the clustering method is important and we present some concrete examples for underlining this.**

I. Introduction

The goal of graph node clustering, related to community detection within social networks, is to create a partition of the vertices, taking into account the topological structure of the graph, such that the clusters are composed of vertices strongly connected [1], [2], [3], [4]. With the appearance of the social networks on the internet, the number of methods for graph clustering has grown recently. Among the core methods proposed in the literature, we can mention those that optimize a quality function to evaluate the goodness of a given partition, like the modularity, the ratio cut, the min-max cut or the normalized cut [5], [6], [7], [8], [9], hierarchical techniques like divisive algorithms based on the minimum cut [10], spectral methods [11] or Markov Clustering algorithm and its extensions [12].

These graph clustering techniques are very useful for detecting strongly connected groups in a graph but many of them mainly focus on the topological structure, ignoring the properties of the vertices. Nowadays, various data sources like social networks, patent documents or biological data can be seen as graphs where vertices have attributes. With such kind of data, it can be useful to take into account the vertex properties in the clustering process for increasing the accuracy of the partitions. Generally, this is not the case in graph clustering where usually, only the relationships of the network are used. On the other hand, there are also unsupervised methods to group objects according to their textual or numeric attributes, like hierarchical clustering or k-means [13], [14], [15]. More precisely, unsupervised learning affects the objects, represented by attributes, into clusters so that the objects in the same cluster are more similar to each other than to those in other clusters, according to an attribute-based similarity measure.

A new challenge in graph clustering consists in combining structure data corresponding to the network and attribute data describing the vertices. Recently, several works have attempted to tackle this problem of hybrid clustering. We detail the main ones in the next section. However, the combination of several data types rises the problem of the meaning of the clustering. Indeed, the different comparison and distance functions may not be compatible and, consequently, they may lead to contradictory results. Moreover, these results are difficult to evaluate since there is no real benchmark dataset, with structured data and attributed data, suitable for attributed graph clustering evaluation. For this reason, in this work, we have built a dataset with ground truth in order to compare the community of each vertex with its computed cluster. It is a scientist network, mainly based on the publications and the participation in scientific events. It includes textual data (publication titles, abstracts, full text, etc.) and relationship data (co-authorship, co-participation in a same event). For this reason, a clustering method that takes into account several criteria is needed in order to identify in the network, groups of people who are in relation and who share a same research field. In order to detect strongly connected clusters containing persons with similar research interests, we propose different methods to partition the graph using both the structural data and attribute data. Our experiments show that, depending on the weight allowed to each type of data (textual or structural) and the way to combine them during the clustering, the results can be very different. The rest of the article is organized as follows. The next section is dedicated to recently introduced graph clustering techniques that consider attributes and structural information. We define formally the problem in section III while we propose several approaches which consider simultaneously structure data and attribute data in section IV. Our experimental study to evaluate these approaches is detailed in section V and the results in the

section VI. Finally, section VII concludes the article.

II. State of the art

Among the clustering methods, one can distinguish on the one hand the non supervised learning techniques, also called vector-based clustering, which exploit the attributes describing the objects, like hierarchical clustering or k-means and on the other hand those which consider the relationships between the different objects as it is usually the case in graph clustering. Recently, methods which exploit both data types were introduced in order to detect communities in social networks where documents or features are associated to the vertices. For instance, in a pre-processing step, Steinhaeuser and Chawla compute a similarity metric between pairs of vertices, based on the attributes, which is used as a weight for the corresponding edge. Afterwards, any graph clustering method can be applied on the valued graph [16]. This method is similar to the first approach presented in the next section but their metric requires to set a parameter, which is not the case in our work. Zhou *et al.* exploit also the attribute data in order to extend the original graph [17], [18]. They add attribute vertices and edges which connect original vertices sharing the same value. A K-Medoids clustering is then applied on a random walk distance computed on the attribute augmented graph. One limit of this method lies in the fact that it does not suit to continuous attributes. The second method introduced in the next section exploits the same idea. However, the graph is extended in a simpler way, without restriction on the type of attributes considered. In the hierarchical clustering of Li *et al.*, clusters are built under attribute based constraints [19]. In a first step, cores are detected using only the structure data, and afterwards they are merged in function of their attribute similarity. Other works combine the two types of data during the clustering process [20], [21], [22]. Ester *et al.* treat the question in terms of the "Connected k-Center (CkC) problem and they propose NetScan, an extended version of the k-means algorithm with a constraint of internal connectedness [20], [23]. With this condition, two entities in a cluster are connected by an internal path. In NetScan, like in other partitioning algorithms, the number of clusters must be known, but this point has been relaxed in recent works [24]. Recently, other approaches have also been introduced in order to detect dense subgraphs which are also homogeneous for the attributes [25], [26]. Dang *et al.* have extended the Newman's modularity by adding a term to measure the attribute-based similarity between two nodes [22]. In this way, the two types of data are considered simultaneously during the clustering process. However, the clusters may contain unconnected nodes. In a similar way, in the third method proposed in the next section, the two types of data are also considered simultaneously but they are merged into a global distance used during the clustering with the guarantee that the vertices in a same cluster are connected. In section IV, we present several approaches which consider simultaneously structure data and attribute data and which offer the advantage to be easy to carry out while in the next section, the problem of attributed graph clustering is defined more formally.

III. Problem statement

We consider a graph $G = (V, E)$ where $V = \{v_1, \ldots, v_i, \ldots, v_{|V|}\}$ is the set of vertices and $E \subset V \times V$ is the set of unlabeled edges. Graph node clustering consists in grouping the vertices into clusters taking into consideration the edge structure in such a way that there should be many edges within each cluster and relatively few between the clusters [3], [27]. Even if the case of overlapping community detection in which a vertex can be affected to several clusters has been recently studied [28], in this article we consider the general case where the clustering process consists in partitioning the set V of vertices into r disjoint clusters $\mathscr{P} = \{C_1, \ldots, C_r\}$ such that:

- $\bigcup_{k \in \{1,\ldots,r\}} C_k = V$
- $C_k \cap C_l = \emptyset$, $\forall\ 1 \leq k < l \leq r$
- $C_k \neq \emptyset$, $\forall k \in \{1, \ldots, r\}$

Moreover, we suppose that each vertex $v_i \in V$ is associated to a document represented by a vector $d_i = (w_{i1}, \ldots, w_{ij}, \ldots w_{iT})$ where w_{ij} is the weight of the term t_j in the document d_i. These documents can been seen as vertex attributes and G defined as an attributed graph [17].

In an attributed graph clustering problem, the structural links and the attributes are both considered, in such a way that:

- firstly, there should be many edges within each cluster and relatively few between the clusters;
- secondly, two vertices belonging to the same cluster are more similar in terms of attributes, than two vertices belonging to two different clusters.

Thus, the clusters should be well separated and, the vertices belonging to the same cluster should be connected and homogeneous on attribute data.

IV. Attributed graph clustering approaches

We introduce different approaches to partition the graph using both structural and attribute data. The methods differ on the manner in which the relational and attribute data are combined:

- Structure-based clustering: the vertice attributes are used in order to value the edges of the graph, that can be then processed by any weighted graph clustering algorithm (cf. section IV-A);
- Attribute-based clustering: structural information is used, together with vertex attribute similarity to obtain a distance matrix (between each pair of vertices), which can then be processed by any unsupervised clustering algorithm (cf. section IV-B);
- Hybrid clustering: attributes and structure are considered separately in order to compute a distance on each type of data. These distances are then combined into a global distance that can be exploited by any unsupervised clustering algorithm or used to obtain a valued graph, which can be processed by any weighted graph clustering algorithm (cf. section IV-C);

A. Structure-based clustering on attribute weighted graph

In this method, the attributes are used to obtain a weighted graph. We define a textual attribute-based distance dis_T, for instance the euclidean distance or the cosine distance, well suited for textual attributes. The value $dis_T(d_i, d_j)$ is associated to each edge (v_i, v_j) of E. Then, a graph clustering method which is able to handle weighted graphs is used to partition the set of the vertices V, for example, hierarchical algorithms (agglomerative or divisive) or algorithms which optimize quality functions like the Kernighan-Lin algorithm or algorithms based on the modularity [5].

B. Attribute-based clustering on structural distance

In this method, the structural information is used to define a structure based distance $dis_S(v_i, v_j)$ between each pair of vertices (v_i, v_j). In practice, the length of the shortest path between v_i and v_j can be used as $dis_S(v_i, v_j)$, where the shortest path between v_i and v_j is the path that has the smallest number of edges. More sophisticated distances, like the neighborhood Random Walk Distance [17] can also be used. The attributes are also taken into account to associate a value to each edge of E, as explained in the previous section. In this case, the shortest path between v_i and v_j is the smallest sum of the weights of the path edges. Then, any unsupervised learning technique can be applied on the distance matrix.

C. Hybrid clustering

In this method, a global distance $dis_{TS}(v_i, v_j)$ between two vertices v_i and v_j is defined as a linear combination of two distances corresponding respectively to each type of information:

$$dis_{TS}(v_i, v_j) = \alpha\, dis_T(d_i, d_j) + (1 - \alpha)\, dis_S(v_i, v_j) \quad (1)$$

where $dis_T(d_i, d_j)$ is a distance defined on the attributes, $dis_S(v_i, v_j)$ is defined directly on the graph and α is a parameter between 0 and 1.

As previously, the length of a shortest path between v_i and v_j can be used for $dis_S(v_i, v_j)$, and the euclidean distance or the cosine distance for $dis_T(d_i, d_j)$. Then, the partition can be built either with a graph clustering algorithm applied on the graph valued with the global distance or with a non supervised learning technique using the global distance.

V. Experimental study

We performed experiments to evaluate the different methods presented previously. While there are some benchmark datasets suitable for community detection evaluation, based on networks with ground truth, as far as we know, it does not exist such a benchmark dataset, with structured data and attributed data, suitable for attributed graph clustering evaluation. In our context, where the community of each actor is unknown, one can use the measures used either for community evaluation like the modularity or for cluster evaluation like the sum of the square error. However, it is clear that the evaluation measures are linked to the corresponding clustering strategy. To avoid this bias, we have built a dataset with a ground truth in order to compare the community of each vertex with its cluster. We used the accuracy as evaluation measure. This dataset is presented in the following paragraph.

TABLE I
Number of authors per session and conference

Session	Conference	Size (authors)
A Bioinformatics	SAC	24
B Robotics	SAC	16
C Robotics	IJCAI	38
D Constraint	IJCAI	21
Total		99

A. Network data

In order to build an attributed graph with a ground truth, we concentrated on two conferences: SAC 2009 and IJCAI 2009. A co-participation network was generated from the well-known DBLP[1] dataset and the abstracts, titles and research areas were extracted from the websites of the selected conferences[2].

1) Authors and research areas: Three research areas, corresponding to conference sessions, were selected: Robotics, Bioinformatics and Constraint Programming. In both conferences there is a Robotics session, while only SAC 2009 has a session on Bioinformatics and IJCAI 2009 on Constraint Programming. As shown in Table I, there are 24 authors in the first research area (Bioinformatics), $16 + 38 = 54$ in the second one (Robotics) and 21 in the last one (Constraint Programming). Each of these authors corresponds to one vertex of V and its research area membership is used during the evaluation step.

The abstracts and the titles of the articles published by the authors at IJCAI 2009 and SAC 2009 are represented in the vector space model introduced by Salton *et al.* [29]. After a preprocessing of the text with stemming and stopword removal, an attribute vector d_i, in which the components are computed with the tf-idf formula, is attached to each author of V.

2) Social Network: We consider an event e as a journal or a conference referenced in DBLP between 2007 and 2009. A co-participation network is built on the set V, using the DBLP database, as follows.

Let v_i and v_j be two authors belonging to V, if there exists at least one event e such that v_i and v_j are authors for articles published in e (even if they are not co-authors), then (v_i,v_j) $\in E$.

3) Graph: We obtain the attributed graph $G = (V, E)$ having the vertices created with the authors and the edges given by the co-participation relations. Moreover, each vertex (*i.e.* author), is described by textual attributes corresponding to the tf-idf vector associated to his articles and, its true class is the research area (*i.e.* the session A, B, C or D in SAC 2009 or IJCAI 2009) of this author.

[1] http://www.informatik.uni-trier.de/~ley/db/

[2] The dataset is available at http://labh-curien.univ-st-etienne.fr/~combe/datasets/asonam/datasetSessionsRecognition.zip

TABLE II
TEXT-BASED CLUSTERING USING AVERAGE LINKAGE (MODEL T)

(a) Model T evaluated against P_T (3 research areas)

	1	2	3	Total
A SAC 2009 - Bioinformatics	**11**		13	24
B SAC 2009 - Intelligent robotic syst.			**16**	16
C IJCAI 2009 - Robotics and Vision			**38**	38
D IJCAI 2009 - Constraints		**21**		21
Total	11	21	67	99

(b) Model T evaluated against P_{TS} (4 sessions)

	1	2	3	4	Total
A SAC 2009 - Bioinformatics	**11**			13	24
B SAC 2009 - Intelligent robotic syst.			**2**	14	16
C IJCAI 2009 - Robotics and Vision			4	**34**	38
D IJCAI 2009 - Constraints		**21**			21
Total	11	21	6	61	99

TABLE III
STRUCTURE-BASED CLUSTERING (MODEL S)

(a) Model S evaluated against P_S (2 conferences)

	1	2	Total
A SAC 2009 - Bioinformatics	**24**		24
B SAC 2009 - Intelligent robotic syst.	**16**		16
C IJCAI 2009 - Robotics and Vision		**38**	38
D IJCAI 2009 - Constraints		**21**	21
Total	40	59	99

(b) Model S evaluated against P_{TS} (4 sessions)

	1	2	3	4	Total
A SAC 2009 - Bioinformatics	**24**				24
B SAC 2009 - Intelligent robotic syst.	16				16
C IJCAI 2009 - Robotics and Vision		**38**			38
D IJCAI 2009 - Constraints		11			21
Total	40	59	0	0	99

B. Hypotheses

We enumerate here our clustering scenarios and hypothesis and present the foreseen results. We consider 4 vertex subsets, given by the authors publishing in the 4 extracted sessions:

- A: Bioinformatics (SAC),
- B: Robotics (SAC),
- C: Robotics (IJCAI),
- D: Constraint Programming (IJCAI).

1) Text: 3 research areas / 3 clusters (P_T): Considering only textual vertex attributes, the hypothesis underlying our experiments is that this information should permit to retrieve the three research areas: Robotics, Bioinformatics and Constraint Programming, giving the partition into three clusters containing the authors of the three research areas: $P_T = \{A, B \cup C, D\}$.

2) Structure: 2 conferences / 2 clusters (P_S): On the other hand, we suppose that taking into account only structural data should allow to identify two groups corresponding to authors participating to each conference: SAC2009 and IJCAI2009, which define the partition into two clusters $P_S = \{A \cup B, C \cup D\}$.

3) Text and structure: 4 sessions / 4 clusters (P_{TS}): However, if we want to discover each session separately, both textual and structural information have to be used. In this case the partition will be into four clusters $P_{TS} = \{A, B, C, D\}$.

C. Evaluation

The different strategies were evaluated using the accuracy of the obtained clusters, compared to the ground truth considered: research areas (P_T), conferences (P_S) or sessions (P_{TS}).

VI. EVALUATED STRATEGIES AND RESULTS

In order to check these hypotheses, we evaluate several models combining text and structure (models TS_1, TS_2, TS_3), corresponding to the different approaches detailed in Section IV. We compare our models against two baselines: clustering based on text only (model T) and clustering based on structure only (model S).

A. Text-only based clustering: model T

Textual clustering considers only the attribute data *i.e.* the documents $\{d_i, \forall v_i \in V\}$. This text-based categorization (model T) was firstly performed with the euclidean distance as well as with the cosine distance computed on the tf-idf description, and with the bisecting K-means algorithm [30]. Then, the model T was performed with the cosine distance, still computed on the tf-idf description, and with the average linkage algorithm. As the latter strategy gives better results, it is the only one presented here, as a baseline for our experiments. Consequently, we have also used the average linkage algorithm in all the attribute-based models.

Results obtained using only textual information are presented in Table IIa for the partition in three clusters which should be compared to P_T. This is an accuracy matrix, where the columns of this table contain the number of authors classified in each cluster. Here the method clustered 11 authors in the first cluster, 21 in the second and 67 in the third cluster. The rows contain the ground truth $T_T = \{A, B \cup C, D\}$ obtained by merging the second line and the third line. Looking at the results we can remark, that 13 authors were clustered in the third class, mainly containing people publishing in Robotics, while according to our ground truth, they belong to the bioinformatics community. The Table IIb contains the results for four clusters compared to P_{TS}.

As expected, the accuracy is higher for the partition in three clusters P_T ($\frac{(11+16+38+21)}{99} \times 100 = 87\%$) than for the partition in four clusters P_{TS} (69%). This result confirms our hypothesis according to which the textual data allows to identify the different research areas but fails to detect correctly the four sessions.

B. Structure-only based clustering: model S

The algorithm by Blondel *et al.* [31] only exploits structural data (*i.e.* the graph $G = (V, E)$). This extension of the Newman and Girvan's algorithm [32], well known for its capacity to handle large graphs, is a greedy method which optimizes the "modularity" of the partitions built on the network. This algorithm, applied directly on the graph

TABLE IV
STRUCTURE-BASED CLUSTERING ON ATTRIBUTE WEIGHTED GRAPH: MODEL TS_1

	1	2	3	4	5	Total
A SAC - Bioinformatics		13		**11**		24
B SAC - Intelligent robotic syst.		**11**			5	16
C IJCAI - Robotics and Vision			**38**			38
D IJCAI - Constraints	**15**		6			21
Total	15	24	44	11	5	99

TABLE V
ATTRIBUTES-BASED CLUSTERING ON STRUCTURAL DISTANCE: MODEL TS_2 (AVERAGE LINK)

	1	2	3	4	Total
A SAC 2009 - Bioinformatics		**24**			24
B SAC 2009 - Intelligent robotic syst.	**4**	11	1		16
C IJCAI 2009 - Robotics and Vision			1	**37**	38
D IJCAI 2009 - Constraints I			**7**	14	21
Total	4	35	9	51	99

TABLE VI
HYBRID CLUSTERING: MODEL TS_3

	1	2	3	4	Total
A SAC 2009 - Bioinformatics	**11**			13	24
B SAC 2009 - Intelligent robotic syst.			**2**	14	16
C IJCAI 2009 - Robotics and Vision			4	**34**	38
D IJCAI 2009 - Constraints		**21**			21
Total	11	21	6	61	99

TABLE VII
RESULTS SYNTHESIS: MODELS T, S, TS_1, TS_2 AND TS_3

	Accuracy considering:		
Model	P_T	P_S	P_{TS}
T	87%	-	69%
S	-	100%	63%
TS_1	-	-	76%
TS_2	-	-	73%
TS_3	-	-	47-69%

$G = (V, E)$, provides a bipartition which is exactly the ground truth $P_S = \{A \cup B, C \cup D\}$ as shown in Table IIIa. Thus, the identification of the two conferences using structural data is perfectly achieved. However, the accuracy is only equal to 63% if we consider the four sessions as the ground truth (P_{TS}), see Table IIIb.

C. Structure-based clustering on attribute weighted graph: model TS_1

In this strategy, corresponding to the approach presented in Section IV-A, the cosine distance computed on the tf-idf vectors is associated to each edge in order to obtain a weighted graph. Then, this graph is partitioned by the method of Blondel *et al.*.

As we can note on Table IV, taking into account structural and attribute data improves the accuracy which reaches 76% for the partition in four clusters (P_{TS}), when it is only equal to 69% without attribute data. This result confirms our hypothesis according which the two types of information are useful to identify the four sessions (P_{TS}).

D. Attribute-based clustering on structural distance: model TS_2

Like previously, the cosine distance computed on the TF-IDF vectors is associated to each edge in order to obtain a weighted graph. Then, the geodesic distance between two vertices is defined as the smallest sum of the weights of the path edges between these vertices. Finally, a hierarchical agglomerative clustering is applied on the geodesic distance matrix, using usual distance between clusters: single link, complete link, average link and center of gravity.

Table V presents the results obtained. With a classification accuracy of 73% for the partition in four clusters (P_{TS}), the results are similar to those obtained with the modularity based algorithm and higher than those obtained using only one type of information (textual or structural).

Except the average link, we have also experimented the single link, the complete link and the center of gravity. The results (not presented here) are the same for each method.

E. Hybrid clustering: model TS_3

In this approach, a global distance is defined as a linear combination of two distances, each corresponding to a type of data: cosine distance on textual information and geodesic distance on the graph G. Then a hierarchical agglomerative clustering is applied with the global distance matrix.

This strategy corresponds to the hybrid clustering presented in Section IV-C.

Even if this method appears as a simple solution for exploiting simultaneously the two types of data, it is not so easy to use since it requires to set the parameter α in the linear function. Moreover, in our experiments, the accuracy for the partition in four clusters (P_{TS}) varies in function of α between 47% (α set to 0.85, 0.96) and 69% (α set to 1) as shown in Table VI. Thus, the best accuracy corresponds to those obtained with a text-based clustering and it is not so good than those obtained with the previous methods combining structural data and attribute data.

F. Results synthesis

The results obtained by the models T, S, TS_1, TS_2 and TS_3 are synthesized in Table VII.

VII. CONCLUSION AND FUTURE WORK

As it has been presented in the previous sections, we obtain very different results according to the clustering method combination and the data taken into account when partitioning an attributed graph.

In this study we have searched to point the difficulties of choosing the right clustering methods. We have built a dataset from real world data containing enough nodes so that clustering algorithms can give fine results, yet having

precise measurable clusters according to several different ways to create partitions. Having this ground truth, we have been able to evaluate a series of clustering methods and compare their results. In our experiments, textual attribute based clustering enables quite well to retrieve the research interests, structure based clustering taking into account co-participation information gets perfectly the conferences but the structural information and attribute information are useful to retrieve the four sessions corresponding to participants in one conference who share a common interest.

We have carried out three different scenarios to combine these information in a common clustering. In our case, the linear combination, corresponding to the hybrid clustering deteriorates the results. In addition, the linear combination method is difficult to apply. It needs a weight parameter to precise the relative importance of each type of information.

The other two scenarios starting with the structural and the textual data give better results than the linear combination.

We have also showed that good clustering results can be obtained using simple methods, when having a scenario adapted to the data and having precise criteria characterizing a good cluster.

We intend to study more deeply the usage interests and characteristics of high level multi criteria clustering methods in order to provide precise clustering scenario choice criteria. We are also working on more real world examples and datasets that help choosing quickly the best clustering scenario for a given dataset.

ACKNOWLEDGMENT

This work was partially supported by St-Étienne Métropole (http://www.agglo-st-etienne.fr/) and the Région Rhône-Alpes.

REFERENCES

[1] U. Brandes, M. Gaertler, and D. Wagner, "Experiments on Graph Clustering Algorithms," in *In 11th Europ. Symp. Algorithms*. Springer-Verlag, 2003, pp. 568–579.

[2] M. Newman, "Detecting community structure in networks," *The European Physical Journal B-Condensed Matter and Complex Systems*, vol. 38, no. 2, pp. 321–330, 2004.

[3] S. Schaeffer, "Graph clustering," *Computer Science Review*, vol. 1, no. 1, pp. 27–64, 2007.

[4] A. Lancichinetti and S. Fortunato, "Community detection algorithms: a comparative analysis," *Physical review E*, vol. 80, no. 5, p. 056117, 2009.

[5] B. W. Kernighan and S. Lin, "An Efficient Heuristic Procedure for Partitioning Graphs," *Bell System Technical Journal*, vol. 49, no. 2, pp. 291–307, 1970.

[6] P. K. Chan, M. D. F. Schlag, and J. Y. Zien, "Spectral K-way ratio-cut partitioning and clustering," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 13, no. 9, pp. 1088–1096, 1994.

[7] J. Shi and J. Malik, "Normalized cuts and image segmentation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 22, no. 8, pp. 888–905, 2000.

[8] C. Ding, X. He, H. Zha, and M. Gu, "A min-max cut algorithm for graph partitioning and data clustering," in *Proceedings IEEE International Conference on Data Mining*, 2001, pp. 107 – 114.

[9] M. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Physical review E*, vol. 69, no. 2, pp. 1–16, 2004.

[10] G. Flake, R. Tarjan, and K. Tsioutsiouliklis, "Graph clustering and minimum cut trees," *Internet Mathematics*, vol. 1, no. 4, pp. 385–408, 2003.

[11] U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, no. 4, pp. 395–416, 2007.

[12] V. Satuluri and S. Parthasarathy, "Scalable graph clustering using stochastic flows: applications to community discovery," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2009, pp. 737–746.

[13] J. Ward, "Hierarchical grouping to optimize an objective function," *Journal of the American statistical association*, vol. 58, no. 301, pp. 236–244, 1963.

[14] A. K. Jain and R. C. Dubes, "Algorithms for Clustering Data," in *Prentice Hall Advanced Reference Series*, A. K. Jain and R. C. Dubes, Eds. Prentice Hall, Inc., 1988.

[15] A. Gordon, *Classification, 2nd Edition*. Chapman & Hall, 2000.

[16] K. Steinhaeuser and N. Chawla, "Community detection in a large real-world social network," *Social Computing, Behavioral Modeling, and Prediction*, pp. 168—-175, 2008.

[17] Y. Zhou, H. Cheng, and J. Yu, "Graph clustering based on structural/attribute similarities," *Proceedings of the VLDB Endowment*, vol. 2, no. 1, pp. 718–729, 2009.

[18] Y. Zhou, H. Cheng, and J. X. Yu, "Clustering Large Attributed Graphs: An Efficient Incremental Approach," *2010 IEEE International Conference on Data Mining*, pp. 689–698, Dec. 2010.

[19] H. Li, Z. Nie, W.-C. W. Lee, C. L. Giles, and J.-R. Wen, "Scalable Community Discovery on Textual Data with Relations," *Proceedings of the 17th ACM conference on Information and knowledge management*, pp. 1203–1212, 2008.

[20] M. Ester, R. Ge, B. Gao, Z. Hu, B. Ben-Moshe, and B. B.-M. M. E. R. G. B. J. G. Zengjian Hu, "Joint Cluster Analysis of Attribute Data and Relationship Data: the Connected k-Center Problem," in *SIAM International Conference on Data Mining*. ACM Press, 2006, pp. 25–46.

[21] F. Moser, "Data mining for feature vector networks," Ph.D. dissertation, Simon Fraser University, 2009.

[22] T. A. Dang and E. Viennet, "Community Detection based on Structural and Attribute Similarities," in *International Conference on Digital Society (ICDS)*, 2012, pp. 7–12.

[23] R. Ge, M. Ester, B. J. Gao, Z. Hu, B. Bhattacharya, and B. Ben-Moshe, "Joint cluster analysis of attribute data and relationship data," *ACM Transactions on Knowledge Discovery from Data*, vol. 2, no. 2, pp. 1–35, 2008.

[24] F. Moser, R. Ge, and M. Ester, "Joint Cluster Analysis of Attribute and Relationship Data Without A-Priori Specification of the Number of Clusters," in *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, Aug. 2007, p. 510.

[25] S. Günnemann, I. Farber, B. Boden, and T. Seidl, "Subspace Clustering Meets Dense Subgraph Mining: A Synthesis of Two Paradigms," in *Proceedings of the IEEE International Conference on Data Mining*, 2010, pp. 845–850.

[26] S. Günnemann, B. Boden, and T. Seidl, "DB-CSC: a density-based approach for subspace clustering in graphs with feature vectors," *Machine Learning and Knowledge Discovery in Databases*, pp. 565–580, 2011.

[27] M. Girvan and M. Newman, "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences*, vol. 99, no. 12, p. 7821, 2002.

[28] W. Qinna and E. Fleury, "Detecting overlapping communities in graphs," in *European Conference on Complex Systems*, 2009.

[29] G. Salton and M. J. McGill, *Introduction to modern Information Retrieval*. McGraw-Hill, 1983.

[30] M. Steinbach, G. Karypis, and V. Kumar, "A comparison of document clustering techniques," *KDD workshop on text mining*, vol. 400, no. X, pp. 525–526, 2000.

[31] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, 2008.

[32] M. E. J. Newman, "Fast algorithm for detecting community structure in networks," *Physics*, vol. 69, no. 2, pp. 1–5, 2004.