On the NP-Completeness of Some Graph Cluster Measures

Jiří Šíma[1]* and Satu Elisa Schaeffer[2]**

[1] Institute of Computer Science, Academy of Sciences of the Czech Republic, P. O. Box 5, 18207 Prague 8, Czech Republic, `sima@cs.cas.cz`

[2] Laboratory for Theoretical Computer Science, Helsinki University of Technology, P. O. Box 5400, FI-02015 TKK, Finland, `elisa.schaeffer@tkk.fi`

Abstract. Graph clustering is the problem of identifying sparsely connected dense subgraphs (clusters) in a given graph. Identifying clusters can be achieved by optimizing a fitness function that measures the quality of a cluster within the graph. Examples of such cluster measures include the conductance, the local and relative densities, and single cluster editing. We prove that the decision problems associated with the optimization tasks of finding clusters that are optimal with respect to these fitness measures are NP-complete.

1 Introduction

Clustering is an important issue in the analysis and exploration of data. There is a wide area of applications in data mining, VLSI design, parallel computing, web searching, relevance queries, software engineering, computer graphics, pattern recognition, gene analysis, etc. See also [13] for an overview. Intuitively clustering consists in discovering natural groups (clusters) of similar elements in a data set. The input data sets of current interesting application areas are very large, which motivates research on the complexity of finding and evaluating complete clusterings or single clusters, as such results will help to determine whether certain methods will be scalable for processing large data sets.

An important variant of data clustering is graph clustering where the similarity relation is expressed by a graph. The graph may either be a weighted graph with the similarity values captured by the edge weights, or an unweighted one, where the similarities have been thresholded or otherwise coded such that an edge is only placed between two vertices if the vertices are considered similar; the absence of an edge implies dissimilarity. In this paper, we restrict ourselves to unweighted, undirected graphs with no self-loops.

We first recall some basic definitions from graph theory. Let $G = (V, E)$ be an *undirected* graph and denote by $E(S) = \{\{u, v\} \in E \mid u, v \in S\}$ the set of edges in a *subgraph* $G(S) = (S, E(S))$ *induced* by a subset of vertices

* Research partially supported by project 1M0021620808 of The Ministry of Education of the Czech Republic.

** Research supported by the Academy of Finland, grant 126235, and the Nokia Fndn.

$S \subseteq V$. We say that $S \subseteq V$ creates a *clique* of *size* $|S|$ if edges in $E(S) = \{\{u,v\} \mid u,v \in S, u \neq v\}$ join every pair of different vertices in S. Further denote by $d_G(v) = |\{u \in V \mid \{u,v\} \in E\}|$ the *degree* of vertex $v \in V$ in G. We say that graph G is a *cubic* graph if $d_G(v) = 3$ for every $v \in V$. Moreover, any subset of vertices $S \subseteq V$ creates a *cut* of G, that is a partition of V into disjoint sets S and $V \setminus S$. The *size* of cut S is defined as

$$c_G(S) = |\{\{u,v\} \in E \mid u \in S\,, v \in V \setminus S\}|\,, \tag{1}$$

and

$$d_G(S) = \sum_{v \in S} d_G(v) \tag{2}$$

denotes the sum of degrees in cut $S \subseteq V$.

A canonical definition of a graph cluster does not exist, but it is commonly agreed that a cluster should be a connected subgraph induced by a vertex set S with many internal edges $E(S)$ and few edges to outside vertices in $V \setminus S$ [5, 15]. In this paper we consider several locally computable fitness functions that are used for measuring the quality of a cluster within the graph. The prominent position among graph cluster measures is occupied by the *conductance* [4, 8, 9, 11, 15] which is defined for any cut $\emptyset \neq S \subset V$ in graph G as follows

$$\Phi_G(S) = \frac{c_G(S)}{\min(d_G(S), d_G(V \setminus S))}\,. \tag{3}$$

Furthermore, the *local density* $\delta_G(S)$ [23] (cf. the *average degree* [12]) of a subset $\emptyset \neq S \subseteq V$ in graph G is the ratio of the number of edges in subgraph $G(S)$ induced by S over the number of edges in a clique of size $|S|$ vertices, that is

$$\delta_G(S) = \frac{|E(S)|}{\binom{|S|}{2}} = \frac{2 \cdot |E(S)|}{|S| \cdot (|S| - 1)} \tag{4}$$

for S containing at least two vertices whereas we define $\delta_G(S) = 0$ for $|S| = 1$. Similarly, we define the *relative density* [19] of cut $\emptyset \neq S \subseteq V$ as follows

$$\varrho_G(S) = \frac{|E(S)|}{|E(S)| + c_G(S)}\,. \tag{5}$$

Yet another graph cluster measure which we call *single cluster editing* (cf. [21]) of a subset $S \subseteq V$ counts the number of edge operations (both additions and deletions) needed to transform S into an isolated clique:

$$\varepsilon_G(S) = \binom{|S|}{2} - |E(S)| + c_G(S)\,. \tag{6}$$

Proposed clustering algorithms [4, 13, 19, 20] usually search for clusters that are optimal with respect to the above-mentioned fitness measures. Therefore the underlying optimization problems of finding the clusters that minimize the conductance or maximize the densities or that need a small single cluster editing are of special interest. In this paper we will formally prove that the associated decision problems for the conductance (Sect. 2), local and relative densities (Sect. 3), and single cluster editing (Sect. 4) are NP-complete. These complexity results appear to be well-known or at least intuitively credible, but not properly documented in the literature. However, such results are useful in choosing a fitness measure in the design or application of a clustering algorithm, and especially to justify the use of approximate methods as the amount of computation needed to identify the global optima cannot be expected to scale well for large problem instances.

2 Conductance

Finding a subset of vertices that has the minimum conductance in a given graph has been often stated to be an NP-complete problem in the literature [2, 4, 7, 9, 15, 17, 18]. However, we could not find an explicit proof anywhere. For example, the NP-completeness proof due to Papadimitrou [22] for the problem of finding the minimum *normalized cut* which is in fact the conductance of a weighted graph does not imply the hardness in the unweighted case. Thus we provide the proof in this section. The decision version for the conductance problem is formulated as follows:

Minimum Conductance (CONDUCTANCE)
Instance: An undirected graph $G = (V, E)$ and a rational number $0 \leq \phi \leq 1$.
Question: Is there a cut $S \subset V$ such that $\Phi_G(S) \leq \phi$?

Theorem 1. CONDUCTANCE *is NP-complete.*

Proof. Clearly, CONDUCTANCE belongs to NP since a nondeterministic algorithm can guess a cut $S \subset V$ and verify $\Phi_G(S) \leq \phi$ in polynomial time. For the NP-hardness proof the following maximum cut problem on cubic graphs will be reduced to CONDUCTANCE in polynomial time.

Maximum Cut for Cubic Graphs (MAX CUT–3)
Instance: A cubic graph $G = (V, E)$ and a positive integer a.
Question: Is there a cut $A \subseteq V$ such that $c_G(A) \geq a$?

The MAX CUT–3 problem was first stated to be NP-complete in [24] which became a widely used reference [10] although an explicit proof cannot be found there and we were unable to reconstruct the argument from the sketch. Nevertheless, the NP-completeness of MAX CUT–3 follows from its APX-completeness presented in [1]. The following reduction to CONDUCTANCE is adapted from that used for the minimum edge expansion problem [14].

Given a MAX CUT–3 instance, i.e. a cubic graph $G = (V, E)$ with $n = |V|$ vertices, and a positive integer a, a corresponding undirected graph $G' = (V', E')$ for CONDUCTANCE is composed of two fully connected copies of the complement of G, that is $V' = V_1 \cup V_2$ where $V_i = \{v^i \mid v \in V\}$ for $i = 1, 2$, and $E' =$

$E_1 \cup E_2 \cup E_3$ where $E_i = \{\{u^i, v^i\} \mid u, v \in V, u \neq v, \{u, v\} \notin E\}$ for $i = 1, 2$, and $E_3 = \{\{u^1, v^2\} \mid u, v \in V\}$. In addition, define the required conductance bound

$$\phi = \frac{1}{2n-4}\left(n - \frac{2a}{n}\right) . \tag{7}$$

The number of vertices in G' is $|V'| = 2n$ and the number of edges $|E'| = (2n-4)n$ since

$$d_{G'}(v) = 2n - 4 \quad \text{for every } v \in V' \tag{8}$$

due to G being a cubic graph. It follows that G' can be constructed in polynomial time.

For a cut $\emptyset \neq S \subset V'$ in G' with $k = |S| < 2n$ vertices, denote by

$$S_i = \{v \in V \mid v^i \in S\} \quad \text{for } i = 1, 2 \tag{9}$$

the cuts in G that are projections of S to V_1 and V_2, respectively. Since $c_{G'}(S) = c_{G'}(V' \setminus S)$ it holds that $\Phi_{G'}(S) = \Phi_{G'}(V' \setminus S)$ according to definition (3). Hence, $k \leq n$ can be assumed without loss of generality when computing the conductance in G'. Thus,

$$\Phi_{G'}(S) = \frac{|S| \cdot |V' \setminus S| - c_G(S_1) - c_G(S_2)}{(2n-4) \cdot |S|} \tag{10}$$

follows from condition (8) and the fact that G' is composed of two fully connected complements of G, which can be rewritten as

$$\Phi_{G'}(S) = \frac{1}{2n-4}\left(2n - k - \frac{c_G(S_1) + c_G(S_2)}{k}\right) . \tag{11}$$

Now we verify the correctness of the reduction by proving that the MAX CUT–3 instance has a solution if and only if the corresponding CONDUCTANCE instance is solvable. First assume that a cut $A \subseteq V$ exists in G whose size satisfies

$$c_G(A) \geq a . \tag{12}$$

Denote by

$$S^A = \{v^1 \in V_1 \mid v \in A\} \cup \{v^2 \in V_2 \mid v \in V \setminus A\} \subseteq V' \tag{13}$$

the cut in G' whose projections (9) to V_1 and V_2 are $S_1^A = A$ and $S_2^A = V \setminus A$, respectively. Since $|S^A| = n$ and $c_G(A) = c_G(V \setminus A)$ the conductance of S^A can be upper bounded as

$$\Phi_{G'}\left(S^A\right) = \frac{1}{2n-4}\left(n - \frac{2c_G(A)}{n}\right) \leq \frac{1}{2n-4}\left(n - \frac{2a}{n}\right) = \phi \tag{14}$$

according to equations (11), (12), and (7), which shows that S^A is a solution of the CONDUCTANCE instance.

For the converse, assume that the conductance of cut $\emptyset \neq S \subset V'$ in G' meets

$$\Phi_{G'}(S) \leq \phi\,. \tag{15}$$

Let $A \subseteq V$ be a maximum cut in G. For cut S^A defined according to (13) we prove that

$$\Phi_{G'}\left(S^A\right) \leq \Phi_{G'}(S) \tag{16}$$

which is rewritten to

$$\frac{1}{2n-4}\left(n - \frac{2c_G(A)}{n}\right) \leq \frac{1}{2n-4}\left(2n - k - \frac{c_G(S_1)+c_G(S_2)}{k}\right) \tag{17}$$

according to (14) and (11) where $k = |S| \leq n$ and S_1, S_2 are defined in (9). Since $2c_G(A) \geq c_G(S_1) + c_G(S_2)$ due to A being a maximum cut in G, it suffices to show

$$n - k + \left(\frac{1}{n} - \frac{1}{k}\right)\left(c_G(S_1) + c_G(S_2)\right) \geq 0 \tag{18}$$

which follows from $\frac{1}{n} - \frac{1}{k} \leq 0$ and $c_G(S_1) + c_G(S_2) \leq |S_1| \cdot n + |S_2| \cdot n = kn$. Thus,

$$\frac{1}{2n-4}\left(n - \frac{2c_G(A)}{n}\right) = \Phi_{G'}\left(S^A\right) \leq \Phi_{G'}(S) \leq \phi = \frac{1}{2n-4}\left(n - \frac{2a}{n}\right) \tag{19}$$

holds according to (14), (16), (15), and (7), which implies $c_G(A) \geq a$. Hence, A solves the MAX CUT-3 instance. □

3 Local and Relative Density

The decision version of the maximum density problem is formulated as follows:

Maximum Density (DENSITY)
Instance: An undirected graph $G = (V, E)$, a positive integer $k \leq |V|$, and a rational number $0 \leq r \leq 1$.
Question: Is there a subset $S \subseteq V$ such that $|S| = k$ and the density of S in G is at least r?

We distinguish between LOCAL DENSITY and RELATIVE DENSITY problems according to the particular density measure used which is the local density (4) and the relative density (5), respectively. Clearly, LOCAL DENSITY is NP-complete since this problem for $r = 1$ coincides with the NP-complete CLIQUE problem [16]. Also the NP-completeness of RELATIVE DENSITY can easily be achieved:

Theorem 2. Relative Density *is NP-complete.*

Proof. Obviously, Relative Density belongs to NP since a nondeterministic algorithm can guess a cut $S \subseteq V$ of cardinality $|S| = k$ and verify $\varrho_G(S) \geq r$ in polynomial time. For the NP-hardness proof the following minimum bisection problem on cubic graphs which is known to be NP-complete [6] will be reduced to Relative Density in polynomial time.

Minimum Bisection for Cubic Graphs (Min Bisection–3)
Instance: A cubic graph $G = (V, E)$ with $n = |V|$ vertices and a positive integer a.
Question: Is there a cut $S \subseteq V$ such that $|S| = \frac{n}{2}$ and $c_G(S) \leq a$?

Given a Min Bisection–3 instance, i.e. a cubic graph $G = (V, E)$ with $n = |V|$ vertices, and a positive integer a, a corresponding Relative Density instance consists of the same graph G, parameters $k = \frac{n}{2}$ and

$$r = \frac{3n - 2a}{3n + 2a} \,. \tag{20}$$

Now for any subset $S \subseteq V$ such that $|S| = k = \frac{n}{2}$ it holds that

$$|E(S)| = \frac{3|S| - c_G(S)}{2} = \frac{3n - 2c_G(S)}{4} \tag{21}$$

due to G being a cubic graph, which gives

$$\varrho_G(S) = \frac{3n - 2c_G(S)}{3n + 2c_G(S)} \tag{22}$$

according to (5). It follows from (20) and (22) that $\varrho_G(S) \geq r$ iff $c_G(S) \leq a$. $\square$

4 Single Cluster Editing

The problem of deciding whether a given graph can be transformed into a collection of cliques using at most m edge operations (both additions and deletions) which is called Cluster Editing is known to be NP-complete [3, 21]. When the desired solution must contain exactly p cliques, the so called p–Cluster Editing problem remains NP-complete for every $p \geq 2$. Here we study the issue of whether a given graph contains a subset S of exactly k vertices such that at most m edge additions and deletions suffice altogether to turn S into an isolated clique:

Minimum Single Cluster Editing (1–Cluster Editing)
Instance: An undirected graph $G = (V, E)$, positive integers $k \leq |V|$ and m.
Question: Is there a subset $S \subseteq V$ such that $|S| = k$ and $\varepsilon_G(S) \leq m$?

Theorem 3. 1–Cluster Editing *is NP-complete.*

Proof. Obviously, 1–Cluster Editing belongs to NP since a nondeterministic algorithm can guess a subset $S \subseteq V$ of cardinality $|S| = k$ and verify $\varepsilon_G(S) \leq m$ in polynomial time. For the NP-hardness proof the Min Bisection–3 problem is used again (cf. the proof of Theorem 2) which will be reduced to 1–Cluster Editing in polynomial time.

Given a Min Bisection–3 instance, i.e. a cubic graph $G = (V, E)$ with $n = |V|$ vertices, and a positive integer a, a corresponding 1–Cluster Editing instance consists of the same graph G, parameters $k = \frac{n}{2}$ and

$$m = \frac{12a + n(n-8)}{8}\,. \tag{23}$$

Now for any subset $S \subseteq V$ such that $|S| = k = \frac{n}{2}$ it holds that

$$\varepsilon_G(S) = \frac{|S| \cdot (|S| - 1)}{2} - \frac{3|S| - c_G(S)}{2} + c_G(S) = \frac{12c_G(S) + n(n-8)}{8} \tag{24}$$

according to (6) and (21). It follows from (23) and (24) that $\varepsilon_G(S) \leq m$ iff $c_G(S) \leq a$. □

5 Conclusion

In this paper we have presented the explicit NP-completeness proofs for the decision problems associated with the optimization of four possible graph cluster measures; namely the conductance, the local and relative densities, and single cluster editing. In addition, the results for relative density and single cluster editing are also valid for cubic graphs (by construction). In clustering algorithms, combinations of fitness measures are often preferred as only optimizing one may result in anomalies such as selecting small cliques or connected components as clusters. An open problem is the complexity of minimizing the *product* of the local and relative densities [20] (e.g. their sum is closely related to the edge operation count for the single cluster editing problem). Another important area for further research is the complexity of finding related approximation solutions [2].

References

1. P. Alimonti, V. Kann: Some APX-completeness results for cubic graphs. *Theoretical Computer Science*, **237**(1-2):123–134, 2000.
2. S. Arora, S. Rao, and U. Vazirani: Expander flows, geometric embeddings and graph partitioning. *Proceedings of the STOC'04 Thirty-Sixth Annual ACM Symposium on Theory of Computing*, pp. 222–231, New York: ACM Press, 2004.
3. N. Bansal, A. Blum, and S. Chawla: Correlation clustering. *Machine Learning*, **56**(1-3):89–113, 2004.
4. U. Brandes, M. Gaertler, and D. Wagner: Experiments on graph clustering algorithms. *Proceedings of the ESA 2003 Eleventh European Symposium on Algorithms*, pp. 568–579, LNCS 2832, Berlin: Springer-Verlag, 2003.

5. A. Broder, S.R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener: Graph structure in the web. *Computer Networks*, **33**(1-6):309–320, 2000.
6. T.N. Bui, S. Chaudhuri, F.T. Leighton, and M. Sipser: Graph bisection algorithms with good average case behavior. *Combinatorica*, **7**(2):171–191, 1987.
7. J.J. Carrasco, D.C. Fain, K.J. Lang, and L. Zhukov: Clustering of bipartite advertiser-keyword graph. *The ICDM 2003 Third IEEE International Conference on Data Mining, Workshop on Clustering Large Data Sets*, Melbourne, Florida, 2003.
8. D. Cheng, R. Kannan, S. Vempala, and G. Wang: A divide-and-merge methodology for Clustering. *Proceedings of the PODS 2005 Twenty-Fourth ACM Symposium on Principles of Database Systems*, Baltimore, June 2005.
9. G.W. Flake, K. Tsioutsiouliklis, and R.E. Tarjan: Graph clustering techniques based on minimum-cut trees. Technical report 2002-06, NEC, Princeton, NJ, 2002.
10. M.R. Garey and D.S. Johnson: *Computers and Intractability: A Guide to the Theory of NP-completeness*. San Francisco: W. H. Freeman & Co., 1979.
11. C. Gkantsidis, M. Mihail, and A. Saberi: Conductance and congestion in power law graphs. *Proceedings of the SIGMETRICS 2003 ACM International Conference on Measurement and Modeling of Computer Systems*, pp. 148–159, New York: ACM Press, 2003.
12. K. Holzapfel, S. Kosub, M.G. Maaß, and H. Täubig: The complexity of detecting fixed-density clusters. *Proceedings of the CIAC 2003 Fifth Italian Conference on Algorithms and Complexity*, pp. 201–212, LNCS 2653, Berlin: Springer-Verlag, 2003.
13. A.K. Jain, M.N. Murty, and P.J. Flynn: Data clustering: A review. *ACM Computing Surveys*, **31**(3):264–323, 1999.
14. V. Kaibel: On the expansion of graphs of 0/1-polytopes. Technical report arXiv:math.CO/0112146, 2001.
15. R. Kannan, S. Vempala, and A. Vetta: On clusterings: Good, bad and spectral. *Proceedings of the FOCS'00 Forty-First Annual Symposium on the Foundation of Computer Science*, pp. 367–377, New York: IEEE Computer Society Press, 2000.
16. R.M. Karp: Reducibility among combinatorial problems. In R.E. Miller and J.W. Thatcher, editors, *Complexity of Computer Computations*, pp. 85–103, New York: Plenum Press, 1972.
17. T. Leighton and S. Rao: Multicommodity max-flow min-cut theorems and their use in designing approximation algorithms. *Journal of the ACM*, **46**(6):787–832, 1999.
18. L. Lovász: Random walks on graphs: A survey. *Bolyai Society Mathematical Studies, 2, Combinatorics, Paul Erdös is Eighty*, Vol. 2, pp. 353–397, Budapest: Bolyai Mathematical Society, 1996.
19. M. Mihail, C. Gkantsidis, A. Saberi, and E. Zegura: On the semantics of internet topologies. Technical report GIT-CC-02-07, College of Computing, Georgia Institute of Technology, Atlanta, GA, 2002.
20. S.E. Schaeffer: Stochastic local clustering for massive graphs. *Proceedings of the PAKDD 2005 Ninth Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 354–360, LNCS 3518, Berlin: Springer-Verlag, 2005.
21. R. Shamir, R. Sharan, and D. Tsur: Cluster graph modification problems. *Proceedings of the WG 2002 Twenty-eighth International Workshop on Graph-Theoretic Concepts in Computer Science*, pp. 379–390, LNCS 2573, Berlin: Springer-Verlag, 2002.

22. J. Shi and J. Malik: Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22**(8):888–905, 2000.
23. S.E. Virtanen: Properties of nonuniform random graph models. Technical report HUT-TCS-A77, Laboratory for Theoretical Computer Science, Helsinki University of Technology, Espoo, Finland, 2003.
24. M. Yannakakis: Node- and edge-deletion NP-complete problems. *Proceedings of the STOC'78 Tenth Annual ACM Symposium on Theory of Computing*, pp. 253–264, New York: ACM Press, 1978.