On the Approximability of Maximum and Minimum Edge Clique Partition Problems

Anders Dessmark[†] **Jesper Jansson**[†] **Andrzej Lingas**[†] **Eva-Marta Lundell**[†]
Mia Persson[⋆]

[†]Department of Computer Science, Lund University, 22100 Lund, Sweden.
Email: andrzej@cs.lth.se

[⋆]School of Technology and Society, Malmö University College, 20506 Malmö, Sweden.
Email: mia@cs.lth.se

Abstract

We consider the following clustering problems: given a general undirected graph, partition its vertices into disjoint clusters such that each cluster forms a clique and the number of edges within the clusters is maximized (*Max-ECP*), or the number of edges between clusters is minimized (*Min-ECP*). These problems arise naturally in the DNA clone classification. We investigate the hardness of finding such partitions and provide approximation algorithms. Further, we show that greedy strategies yield constant factor approximations for graph classes for which maximum cliques can be found efficiently.

Keywords: Approximation algorithms, clique partition

1 Introduction

The *correlation clustering* problem has gained a lot of attention recently (Ailon, Charikar & Newman 2005, Bansal, Blum & Chawla 2004, Charikar, Guruswami & Wirth 2003, Demaine & Immorlica 2003, Emanuel & Fiat 2003, Swamy 2004); given a complete graph with edges labeled "+"(similar) or "−"(dissimilar), find a partition of the vertices into clusters that agrees as much as possible with the edge labels, i.e., that maximizes the *agreements* (the number of "+" edges inside clusters plus the number of "−" edges between cluster) or that minimizes the *disagreements* (the number of "−" edges inside clusters plus the number of "+" edges between clusters).

In this paper, we consider a special variant of the correlation clustering problem in which there are no negative edge labels. Instead, we omit an edge between two vertices of a dissimilar pair. Furthermore, we require an edge between each pair of vertices in a cluster, i.e, every cluster must form a clique. We consider the following two combinatorial optimization problems. The *maximum edge clique partition problem* (*Max-ECP* for short) aims to find a partition of the vertices into cliques such that the total number of edges within all cliques is maximized. The related minimization version of this problem, the *minimum edge clique partition problem* (*Min-ECP* for short), is defined analogously with the exception that the total number of edges between the cliques is minimized.

The *Max-ECP* and *Min-ECP* problems first have been considered in the setting of DNA clone classification (Figueroa, Goldstein, Jiang, Kurowski, Lingas & Persson 2005). In order to characterize cDNA and ribosomal DNA (rDNA) gene libraries, the powerful DNA array based method *oligonucleotide fingerprinting* is commonly used (see, e.g., (Drmanac, Stavropoulos, Labat, Vonau, Hauser, Soares & Drmanac 1996, Herwig, Poustka, Müller, Bull, Lehrach & O'Brien 1999, Valinsky, Della Vedova, Jiang & Borneman 2002, Valinsky, Della Vedova, Scupham, Alvey, Figueroa, Yin, Hartin, Chrobak, Crowley, Jiang & Borneman 2002)). A key step in this method is the cluster analysis, which aims to cluster together similar data, i.e., the *fingerprints*. The problem of clustering binarized fingerprint data such that the number of clusters is minimized was first studied and motivated in (Figueroa, Borneman & Jiang 2004). In (Figueroa, Goldstein, Jiang, Kurowski, Lingas & Persson 2005), Figueroa *et al.* propose new approaches of partitioning binarized fingerprints into disjoint clusters in order to maximize the number of pairs of similar fingerprints lying inside the clusters (equivalently, minimize the number of pairs of similar fingerprints lying in different clusters). These problems can hence be viewed as the *Max-ECP* and *Min-ECP* problems where the vertices are the binarized fingerprints and the edges between them indicate their similarity.

Related results

The well studied correlation clustering problem was first introduced for complete graphs by Bansal *et al.* (Bansal, Blum & Chawla 2004). It has applications in many areas (see, e.g., (Bansal, Blum & Chawla 2004, Demaine & Immorlica 2003)). As noted in (Bansal, Blum & Chawla 2004), the problem of maximizing agreements and minimizing disagreements are equivalent at optimality but differ from the point of view of approximation. In (Bansal, Blum & Chawla 2004), it was established that these problems are NP-hard for complete graphs, and a PTAS was given in the case of maximizing agreements, whereas a constant factor approximation is given in the case of minimizing disagreements. This constant factor approximation was later improved by Charikar *et al.* (Charikar, Guruswami & Wirth 2003) where a factor 4 approximation algorithm is given for complete graphs based on linear programming relaxation. The latter problem was also proved to be APX-hard.

The problems of maximizing agreements and minimizing disagreements were later generalized to include non-necessarily complete graphs with edge weights in (Charikar, Guruswami & Wirth 2003). A factor 0.7664 approximation algorithm based on the rounding of a semidefinite programming relaxation for the problem of maximizing agreements for general

weighted graphs was given in (Charikar, Guruswami & Wirth 2003), but this factor was later improved to 0.7666 by Swamy (Swamy 2004). As for the problem of minimizing disagreements, a factor $O(\log n)$ approximation algorithm for general weighted graphs was proposed (independently) in (Charikar, Guruswami & Wirth 2003),(Demaine & Immorlica 2003), and (Emanuel & Fiat 2003). Recently, Ailon *et al.* (Ailon, Charikar & Newman 2005) have provided a randomized expected 3-approximation algorithm for minimizing disagreements. In the case of weighted complete graphs, which satisfies probability constraints ($w_{ij}^+ + w_{ij}^- = 1$ for edge (i,j)) and triangle inequality constraints ($w_{ik}^- \leq w_{ij}^- + w_{jk}^-$) on the edges, they have provided a factor 2 approximation algorithm.

The APX-hardness of the unweighted version of *Min-ECP* has been established by Shamir *et al.* (Shamir, Sharan & Tsur 2002). They have also presented results in the case when a solution must contain exactly p clusters; *Min-ECP* is solvable in polynomial time for $p = 2$ but NP-complete for $p > 2$.

Our results

In this paper, we investigate the approximability of *Max-ECP* and *Min-ECP*. Specifically, we prove that *Max-ECP* on general, undirected graphs is hard to approximate within a factor of $n^{1-o(1)}$, unless NP $\subseteq$ ZPTIME($2^{(\log n)^{O(1)}}$). On the other hand, we give an n-approximation algorithm running in polynomial time for this problem. In the case of *Min-ECP* we provide a polynomial-time $O(\log n)$-approximation algorithm for this problem on general, undirected graphs with non-negative weights. We also prove that this problem is NP-hard to approximate within $1+\frac{1}{880}-\epsilon$, for any $\epsilon > 0$. We further consider the greedy heuristic and show that it yields a 2-approximation for both *Max-ECP* and *Min-ECP*, under the assumption that the largest clique can be determined in polynomial time. Thus, the greedy method could be applied in practice only to graph classes for which maximum cliques can be found efficiently, for instance chordal graphs, line graphs and circular-arc graphs (cf. (Figueroa, Borneman & Jiang 2004)). We also note that these bounds are actually tight. Figure 1 summarizes our contributions.

Problem	*Lower Bound*	*Upper Bound*
Max-ECP	$n^{1-o(1)}$	n
weightedMin-ECP	$1+\frac{1}{880}-\epsilon$	$O(\log n)$
GreedyMax-ECP	2	2
GreedyMin-ECP	2	2

Figure 1: Summary of results.

Our paper is structured as follows. We give more formal definitions of *Max-ECP* and *Min-ECP* in Section 2. In Section 3, we provide a factor n approximation algorithm for *Max-ECP*. In Section 4, we give a lower bound on approximability of *Max-ECP*. In Section 5, we provide a polynomial-time $O(\log n)$-approximation algorithm for the weighted version of *Min-ECP* and in section 6, we derive a lower bound on approximability of *Min-ECP*. Finally, in Section 7, we consider the greedy algorithm for *Max-ECP* and *Min-ECP* and prove that it yields a 2-approximation.

2 Preliminaries

The formal definition of *Max-ECP* and *Min-ECP* is as follows.

Definition 1 Let $G = (V, E)$ be an undirected graph and let $n = |V|$. The problem of *maximum edge clique partition* (Max-ECP for short) is to find a partition of V into disjoint subsets $V_1, ..., V_k$ such that for each $1 \leq i \leq k$, any two vertices in V_i share an edge and the total number of edges within the subsets $V_1, ..., V_k$ is maximized.

The problem of *minimum edge clique partition* (Min-ECP for short) is defined analogously to *Max-ECP* with the exception that the total number of edges between the subsets $V_1, ..., V_k$ is minimized.

Note that an exact solution to *Max-ECP* is an exact solution to *Min-ECP* and *vice versa*. The example shown in Figure 2 demonstrates two feasible solutions to *Max-ECP* and *Min-ECP*. As depicted in Figure 2(a), the total number of edges inside the clusters is 18, hence the solution to *Max-ECP* has a total cost of 18. On the contrary, the total number of edges outside the clusters in Figure 2(a) is 12, hence the solution to *Min-ECP* has a total cost of 12. The optimal clustering is depicted in Figure 2(b), with the total cost of 24 for *Max-ECP* and the total cost of 6 for *Min-ECP*.

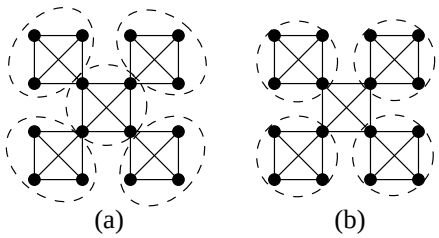

Figure 2: A feasible solution and the optimal solution to *Max-ECP* and *Min-ECP*.

3 A polynomial-time n-approximation algorithm for *Max-ECP*

Max-ECP is NP-hard and even hard to approximate within a factor $n^{1-O(1/(\log n)^\gamma)}$, for some constant γ, as proved in the next section. On the positive side, we prove in this section that *Max-ECP* admits a polynomial-time, factor k approximation algorithm, where k is the number of vertices in the largest clique. The approximation algorithm works as follows: Find a maximum matching in G and output it and the singletons containing the vertices not covered by the matching as a clique partition.

Theorem 1 *Let k be the number of vertices in the largest clique in G. Max-ECP can be approximated within a factor of k in polynomial time.*

Proof: Denote by OPT(G) and APPR(G) the total number of edges within cliques in an optimal solution for *Max-ECP* on G and in the solution returned by the approximation algorithm described above, respectively. Let $(V_1, V_2, \ldots, V_m)$ be an optimal solution for *Max-ECP* on G. There is a matching in G which, for $i = 1, ..., m$, includes at least $\frac{|V_i|-1}{2}$ edges from the clique induced by V_i. Since for $i = 1, ..., m$, $k \geq |V_i|$, such a matching includes at least the $\frac{1}{k}$ fraction of edges from each of the m cliques induced by $V_1, V_2, \ldots, V_m$. Hence, APPR(G) $\geq$ OPT(G) /k holds.

4 A lower bound on the approximability of *Max-ECP*

The maximum clique problem is known to not admit an approximation within $n^{1-O(1/(\log n)^\gamma)}$ for some constant γ unless $\mathcal{NP} \subseteq \text{ZPTIME}(2^{(\log n)^{O(1)}})$ (Khot 2001). It follows that aforementioned lower bound on approximability holds for graphs on n vertices having a clique of size m not less than n^{1-x}, where $x = O(1/(\log n)^\gamma)$. Consider such a graph G. An optimal solution to *Max-ECP* for G has at least $\binom{m}{2}$ edges. Hence, an n^{1-3x} approximation to *Max-ECP* for G has at least $m(m-1)/(2n^{1-3x})$ edges. The size of maximum clique in the approximate solution to *Max-ECP* is minimized if all cliques have equal size h. In this case the total number of edges in the approximate solution is $\binom{h}{2}n/h$ which is less than $nh/2$. Hence, we obtain the inequality $m(m-1)/(2n^{1-3x}) \leq nh/2$ which by our assumptions on G and m yields $h = \Omega(n^x)$. This implies n^{1-x} approximation of the maximum clique problem in G which contradicts (Khot 2001). Thus, we obtain the following theorem.

Theorem 2 *Unless* $\mathcal{NP} \subseteq \text{ZPTIME}(2^{(\log n)^{O(1)}})$, *the Max-ECP problem does not admit an* $n^{1-O(1/(\log n)^\gamma)}$ *approximation, for some constant* γ.

5 A polynomial-time $O(\log n)$-approximation algorithm for weighted *Min-ECP*

Min-ECP can be approximated within a factor of $O(\log n)$ in polynomial time, even for edge-weighted graphs with arbitrary non-negative weights, as follows.

Let $G = (V, E)$ be a given instance of *Min-ECP* in which each edge e has a non-negative weight $w(e)$. Define $W = \max_{e \in E} w(e)$. Construct an edge-weighted, edge-labeled, complete graph $G' = (V, E')$, where each $e \in E'$ is labeled by $'+'$ and assigned weight $w(e)$ if $e \in E$, or labeled by $'-'$ and assigned weight $W \cdot n^2 \log^2 n$ if $e \notin E$. Run any one of the polynomial-time $O(\log n)$-approximation algorithms for Minimum Disagreement Correlation Clustering for weighted graphs (Charikar, Guruswami & Wirth 2003, Demaine & Immorlica 2003, Emanuel & Fiat 2003) on G' to obtain a clustering $\mathcal{C}'$ for V, and return the set $\mathcal{S}$ of subgraphs of G induced by $\mathcal{C}'$.

Lemma 1 *For any two vertices* $u, v \in V$ *which are not joined by an edge in* G, u *and* v *do not belong to the same cluster in* $\mathcal{C}'$.

Proof: Suppose u and v belong to the same cluster in $\mathcal{C}'$. Then the clustering obtained from $\mathcal{C}'$ by placing u in a singleton cluster would have a disagreement score lower than that of $\mathcal{C}'$ by a factor of $\omega(\log n)$, which is a contradiction.

By Lemma 1, the vertices from each cluster in $\mathcal{C}'$ form a clique in G. Since the clusters in $\mathcal{C}'$ are disjoint, $\mathcal{S}$ is a partition of G into cliques, which proves the correctness of the method.

Next, we consider the approximation ratio. For any partition M of G into cliques, denote by $ECP(M)$ the ECP score for M, i.e., the sum of all weights of edges whose two endpoints belong to different cliques in M. Similarly, for any clustering M' of G', let $Disagree(M')$ be the disagreement correlation clustering score for M'. Finally, $MinECP(G)$ and $MinDisagree(G')$ denote the minimum possible scores of ECP for G and $Disagree$ for G', respectively.

Lemma 2 $ECP(\mathcal{S})$ *is at most* $O(\log n)$ *times* $MinECP(G)$.

Proof: Let M be a partition of G into cliques which minimizes ECP, and let M' be the clustering of G' induced by the cliques in M. Then, since only edges labeled by $'+'$ contribute to $Disagree(M')$, we obtain $MinECP(G) = ECP(M) = Disagree(M') \geq MinDisagree(G')$.

Next, observe that $ECP(\mathcal{S})$ is equal to $Disagree(\mathcal{C}')$ because only edges labeled by $'+'$ contribute to $Disagree(\mathcal{C}')$ by Lemma 1. Moreover, $Disagree(\mathcal{C}')$ is at most $O(\log n)$ times $MinDisagree(G')$. It follows that $ECP(\mathcal{S})$ is at most $O(\log n)$ times $MinECP(G)$.

To summarize:

Theorem 3 *Weighted Min-ECP can be approximated within a factor of* $O(\log n)$ *in polynomial time.*

6 A lower bound for *Min-ECP*

Shamir *et al.* have established the APX-hardness of unweighted *Min-ECP* by a reduction from a special variant of set cover in (Shamir, Sharan & Tsur 2002). It folllows by (Shamir, Sharan & Tsur 2002) that the *Min-ECP* problem cannot have a polynomial-time approximation scheme unless P=NP. However, no explicit lower bound on the approximation factor for *Min-ECP* achievable in polynomial time is known in the literature.

In this section, we present a new reduction from the so called three way cut problem to the weighted *Min-ECP* problem which yields an explicit lower bound on the approximation factor.

The problem of three way cut (3WC) is to find a minimum number of edges whose removal disconnects three distinguished vertices.

Let A and B be two optimization problems (maximization or minimization). *A linearly reduces to* B if there are two polynomial time algorithms h and g, and constants $\alpha,\ \beta > 0$ such that

1. For an instance a of A, algorithm h produces an instance $b = h(a)$ of B such that the cost of an optimal solution for b, $opt(b)$, is at most $\alpha \cdot opt(a)$, and
2. For $a,\ b = h(a)$, and any solution y of b, algorithm g produces a solution x of a such that $|cost(x) - opt(a)| \leq \beta |cost(y) - opt(b)|$.

By (Dahlhaus, Johnson, Papadimitriou, Seymour & Yannakakis 1994), if A linearly reduces to B and B has a polynomial-time $1+\epsilon$ approximation algorithm then A has a polynomial-time $(1 + \alpha\beta\epsilon)$ approximation algorithm.

Max-Cut is the problem of finding, for an undirected graph with vertex set V, a partition V_1, V_2 of V such that the number of edges $\{u, v\}$ where $\{u, v\} \cap V_1$ and $\{u, v\} \cap V_2$ are both nonempty is maximized.

In (Dahlhaus, Johnson, Papadimitriou, Seymour & Yannakakis 1994), Dahlhaus *et al.* presented a linear reduction of the Max-Cut problem to 3WC in order to prove that 3WC is APX-hard. Since Max-Cut is APX-hard (Håstad 2001), the APX-hardness of 3WC follows. In the aforementioned reduction $\alpha = 56$ and $\beta = 1$ (Dahlhaus, Johnson, Papadimitriou, Seymour & Yannakakis 1994). In fact, α can be decreased to 55 by the proof of Theorem 5 in (Dahlhaus, Johnson, Papadimitriou, Seymour & Yannakakis 1994)[1].

[1] In the proof of Theorem 5 in (Dahlhaus, Johnson, Papadimitriou, Seymour & Yannakakis 1994), observe that $OPT_{3WC}(f(G)) = 56 \cdot \frac{|E|}{2} - K \leq 56 \cdot OPT_{Max-Cut}(G) - OPT_{Max-Cut}(G) = 55 \cdot OPT_{Max-Cut}(G)$

On the other hand, Håstad has shown that for any $\epsilon > 0$, it is NP-hard to approximate Max-Cut within $1 + \frac{1}{16} - \epsilon$ (Håstad 2001). Hence, we obtain the following lemma.

Lemma 3 *For any $\epsilon > 0$, it is NP-hard to approximate 3WC within $1 + \frac{1}{880} - \epsilon$.*

To reduce 3WC to weighted *Min-ECP*, fix an arbitrary $\delta > 0$, and transform any given instance of 3WC on n vertices to an instance of *Min-ECP* as follows:

1. Assign the weight 1 to each edge in the instance.
2. For each non-adjacent pair u, v of vertices in the instance insert an edge of weight δ/n^2.
3. For each distinguished vertex s_i, $i = 1, 2, 3$, add an auxiliary vertex u_i and make it adjacent with each vertex of the instance. Assign the weight n^2 to each of the three edges (s_i, u_i) and the weight δ/n^2 to the remaining edges incident to the vertices u_i, $i = 1, 2, 3$.

Figure 3 demonstrates how the transformation from an instance of 3WC to an instance of *Min-ECP* works.

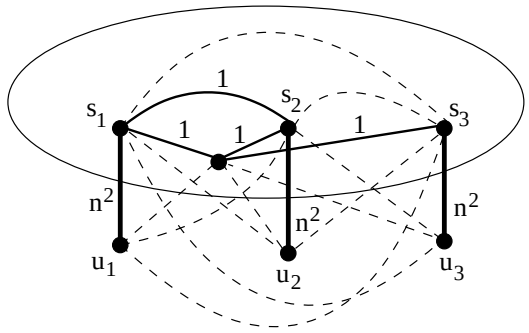

Figure 3: Transformation from 3WC to *Min-ECP*.

In this figure, note that a dashed line between a pair of vertices indicates an edge with weight δ/n^2.

Note that in an optimal *Min-ECP* solution to the transformed instance each of the pairs s_i, u_i, $i = 1, 2, 3$ belongs to a separate clique and the total weight of the edges outside all the cliques in the optimal solution is between *cut* and *cut* + δ where *cut* stands for the value of an optimal solution to the instance of 3WC.

Suppose that for some $\epsilon > 0$, weighted *Min-ECP* could be approximated in polynomial time within a factor of f where $f \leq 1 + \frac{1}{880} - \epsilon$. Then using the set of edges between the three cliques in an approximate solution for weighted *Min-ECP* as an approximate solution for 3WC would yield a three-way cut for the original graph of cardinality at most $f \cdot (cut + \delta) \leq (f + f \cdot \delta) \cdot cut$. By setting $\delta = \frac{\epsilon}{2 \cdot (1 + \frac{1}{880} - \epsilon)}$, we could approximate 3WC in polynomial time within $1 + \frac{1}{880} - \epsilon/2$. We obtain a contradiction with Lemma 3. Hence, we obtain the following theorem.

Theorem 4 *For any $\epsilon > 0$, it is NP-hard to approximate weighted Min-ECP within $1 + \frac{1}{880} - \epsilon$.*

7 Greedy method for *Max-ECP* and *Min-ECP*

The greedy strategy applies naturally to the *Max-ECP* and *Min-ECP* problems: iteratively pick the largest clique until all elements have been partitioned into disjoint clusters. However, the problem of finding a maximum clique is itself known to be extremely hard to approximate (Khot 2001). Thus, the greedy method could be applied in practice only to graph classes for which maximum cliques can be found efficiently (cf. (Figueroa, Borneman & Jiang 2004)).

Theorem 5 *The greedy method yields a 2-approximation for Max-ECP and Min-ECP.*

Proof: Consider an optimal solution to the *Max-ECP* problem (or, the *Min-ECP* problem, respectively) and let us assume that it consists of m cliques. Let C be the largest clique, say on k vertices, picked by the greedy method. Suppose first that the intersection of C with any clique in the optimal partition is a singleton or empty. Thus, in a way, the at most $k(k-1)$ former clique edges are replaced with the $k(k-1)/2$ edges in C (or, the $k(k-1)/2$ edges in C previously outside the cliques with at most $k(k-1)$ new edges outside the cliques, respectively). In the remaining case, if the intersection of C with any of the cliques in the optimal partition contains more than one vertex, less than $k(k-1)$ former clique edges are replaced by the $k(k-1)/2$ edges in C (or, the $k(k-1)/2$ edges in C previously outside the cliques are replaced by less than $k(k-1)$ new edges outside the cliques, respectively). By iterating the argument, we obtain the theorem.

The example shown in Figure 4 demonstrates that our upper bound on the approximation factor of the greedy method for *Max-ECP* is tight. Simply, the greedy method may produce n 2-cliques and $2n$ 1-cliques (singletons) yielding n edges whereas the optimal clique partition consists of $2n$ 2-cliques yielding $2n$ edges.

Figure 4 is also a tight example for greedy *Min-ECP*. Note that the number of edges between cliques will be $2n$ in the approximate solution, whereas the optimum contains n edges between the $2n$ 2-cliques.

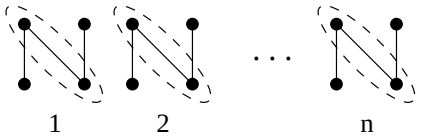

Figure 4: An example illustrating the worst-case performance of the greedy strategy for *Max-ECP* and *Min-ECP*.

8 Final remarks

By using rather maximum weight matching than maximum cardinality matching we can easily generalize our n-approximation method for *Max-ECP* to include edge weights.

It is an interesting open problem whether or not the gap between the upper and lower bounds on approximability of *Min-ECP* could be tightened.

A careful reader might observe that our approximation hardness result for *Max-ECP* does not hold for the graph classes for which our greedy method could be applied practically. The complexity and approximation status of *Max-ECP* and *Min-ECP* for the aforementioned graph classes are interesting open problems.

References

Ailon, N., Charikar, M. & Newman, A. (2005), Aggregating inconsistent information: Ranking and

Clustering, *in* 'Proc. 37th Annual ACM Symposium on Theory of Computing (STOC 2005)', pp. 684–693.

Bansal, N., Blum, A. & Chawla, S. (2004), 'Correlation Clustering', *Machine Learning* **56**(1–3), 89–113.

Charikar, M., Guruswami, V. & Wirth, A. (2003), Clustering with Qualitative Information, *in* 'Proc. 44th Annual Symposium on Foundations of Computer Science (FOCS 2003)', pp. 524–533.

Dahlhaus, E., Johnson, D.S., Papadimitriou, C.H., Seymour, P.D. & Yannakakis, M. (1994), 'The Complexity of Multiterminal Cuts', *SIAM J. Comput.* **23**, 864–894.

Demaine, E. & Immorlica, N. (2003)Correlation Clustering with Partial Information, *in* 'Proc. 6th International Workshop on Approximation Algorithms for Combinatorial Optimization Problems (APPROX 2003)', pp. 1–13.

Drmanac, S., Stavropoulos, N.A., Labat, I., Vonau, J., Hauser, B., Soares, M.B. & Drmanac, R. (1996), 'Gene-representing cDNA clusters defined by hybridization of 57,419 clones from infant brain libraries with short oligonucleotide probes', *Genomics* **37**, 29–40.

Emanuel, D. & Fiat, A. (2003)Correlation Clustering – Minimizing Disagreements on Arbitrary Weighted Graphs, *in* 'Proc. 11th Annual European Symposium on Algorithms (ESA 2003)', pp. 208–220.

Figueroa, A., Borneman, J. & Jiang, T. (2004), 'Clustering binary fingerprint vectors with missing values for DNA array data analysis', *Journal of Computational Biology* **11**(5), 887–901.

Figueroa, A., Goldstein, A., Jiang, T., Kurowski, M., Lingas, A. & Persson, M. (2005)Approximate Clustering of Fingerprint Vectors with Missing Values, *in* 'Proc. 11th Computing: The Australasian Theory Symposium (CATS 2005)', pp. 57–60.

Herwig, R., Poustka, A.J., Müller, C., Bull, C., Lehrach, H. & O'Brien, J. (1999), 'Large-scale clustering of cDNA-fingerprinting data', *Genome research* **9**, 1093–1105.

Håstad, J. (2001), 'Some optimal inapproximability results', *Journal of the ACM* **48**(4), 798–859.

Khot, S. (2001), Improved inapproximability results for MaxClique, chromatic number, and approximate graph coloring ,*in* 'Proc. 42th Annual Symposium on Foundations of Computer Science (FOCS 2001)', pp. 600–609.

Shamir, R., Sharan, R. & Tsur, D. (2002)Cluster Graph Modification Problems, *in* 'Proc. 28th International Workshop on Graph Theoretic Concepts in Computer Science (WG 2002)', pp. 379–390.

Swamy, C. (2004)Correlation Clustering: maximizing agreements via semidefinite programming, *in* 'Proc. 15th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA 2004)', pp. 526–527.

Valinsky, L., Della Vedova, G., Jiang, T. & Borneman, J. (2002), 'Oligonucleotide fingerprinting of ribosomal RNA genes for analysis of fungal community composition', *Applied and Environmental Microbiology* **68**, 5999–6004.

Valinsky, L., Della Vedova, G., Scupham, A., Alvey, S., Figueroa, A., Yin, B., Hartin, R., Chrobak, M., Crowley, D., Jiang, T. & Borneman, J. (2002), 'Analysis of bacterial community composition by oligonucleotide fingerprinting of rRNA genes', *Applied and Environmental Microbiology* **68**, 3243–3250.