The Complexity of Clustering in Planar Graphs

J. Mark Keil
Timothy B. Brecht

Department of Computational Science
University of Saskatchewan
Saskatoon, Canada S7N 0W0

Abstract. An h-cluster in a graph is a set of h vertices which maximizes the number of edges in the graph induced by these vertices. We show that the connected h-cluster problem is NP-complete on planar graphs.

1. Introduction

The problem of finding the maximum clique in a graph is one of the most fundamental graph problems and has many applications. In a planar graph, although the absence of any large cliques makes the clique problem trivial, there may still be a need for finding a dense subgraph. We thus turn to the idea of a cluster. The decision problem formulation of the h-cluster problem is as follows: given a graph G and positive integers h and j, does there exist an induced subgraph on h vertices such that this subgraph has at least j edges. It may also be of interest to ensure that the cluster is connected. Both the h-cluster problem and the connected h-cluster problem have been shown to be NP-complete even for bipartite graphs or chordal graphs [5]. The h-cluster problem is also NP-complete for bipartite or regular graphs of fixed degree [7]. In constrast to this, polynomial time algorithms have recently been developed for the optimization versions of both the h-cluster problem and the connected h-cluster problem for various subclasses of planar graphs. These classes include k-outerplanar graphs [2], series parallel graphs [10,11], Halin graphs [3], $\Delta - Y$ graphs [6] and $Y - \Delta$ graphs [6]. The algorithms exist as a result of the above mentioned subclasses of planar graphs being partial k-trees for fixed k [3,6,11]. Both the h-cluster problem and the connected h-cluster problem have polynomial time algorithms on partial k-trees for fixed k [1,4,9]. This then motivates the investigation of the complexity of the clustering problems on general planar graphs. In this paper we show that the connected h-cluster problem is NP-complete on general planar graphs.

2. NP-Completeness

To prove the NP-completeness, we use a reduction from the connected vertex cover problem on planar graphs with maximum degree four, which was proven NP-complete in [8].

Theorem. *The connected h-clustering problem on planar graphs is NP-complete.*

Proof: It is not hard to see that the problem is in NP. To prove the problem NP-hard, we show that the connected k-vertex cover problem on an arbitrary planar graph G with maximum degree four may be polynomially reduced to the connected h-clustering problem for a planar graph G' which is constructed from G. Let n be the number of vertices in G and let m be the number of edges in G. To construct G' from G begin by subdividing each edge by introducing 5 new vertices. Then for each original edge introduce a maximal planar graph with $q = 2m + n$ vertices and identify a vertex from the exterior face of this graph with the middle new vertex in each original edge. Note that $|V'| = n + 4m + m(2m+n)$. We call the n vertices in G' corresponding to vertices in G type G vertices, the $m(2m+n)$ vertices contained within the maximal planar graphs associated with the edges type M vertices and the remaining $4m$ vertices which subdivided the edges of G type S vertices. It is clear that G' can be constructed from G in polynomial time. Figure 1 shows an example of the construction of G' from G.

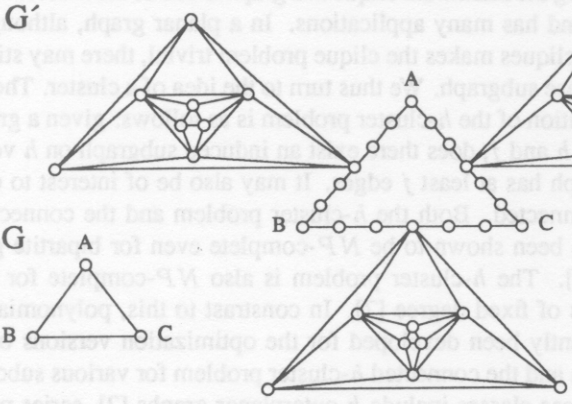

Figure 1

To complete the proof of the theorem it is necessary to verify the following claim.

Claim. *There is a connected vertex cover in G of size k if and only if there is a $qm + k + 2m + 2(k-1)$ connected cluster in G' with $m(3q-6) + 3m + 3(k-1)$ edges.*

Proof of Claim: Given a connected vertex cover in G with k vertices, the required cluster in G' consists of the following sets of vertices: (a) the qm type M vertices of G', (b) the k type G vertices of G' corresponding to the vertex cover of G, (c) the 2 type S vertices of G' for each of the m original edges of G which serve to connect an endpoint type G vertex which is in the vertex cover, to the middle type M vertex in the edge, and (d) the $2(k-1)$ type S vertices of G' which are the two remaining subdividing vertices in each of the original edges in a spanning tree of the subgraph of G induced by the connected vertex cover.

In the cluster, the vertices of (a) are connected to the vertices of (b) via the vertices of (c). The vertices added in (d) ensure that there is a path between each pair of vertices in (b). We thus have that the cluster is connected.

It remains to show that the cluster has the required number of edges. From Euler's formula we know that a maximal planar graph with n vertices has $3n-6$ edges, thus the vertices of (a) contribute $m(3q-6)$ edges to the cluster. There are $3m$ edges in the cluster adjacent to the vertices of (c). The remaining $3(k-1)$ edges are adjacent to vertices of (d).

To complete the proof of the claim we begin with the connected cluster, C, in G' as specified in the claim and show that the required vertex cover exists. First we show that the connected cluster contains vertices associated with each edge of G.

Lemma 1. *The cluster C contains at least three type M vertices associated with each edge of G.*

Proof: Assume to the contrary that there is a "short" maximal planar graph with two or fewer vertices from the cluster C. Then the only vertices available to form the cluster are the at most $2+q(m-1)$ from the maximal planar graphs plus the $n+4m$ type G and type S vertices. Thus a total of at most $n+4m+q(m-1)+2$ vertices are available. But the cluster must be of size $qm+k+2(k-1)+2m$. If $k \geq 2$ then $n+4m+q(m-1)+2 < q(m)+k+2(k-1)+2m$ that is $n+2m+2 < q+3k-2$ replacing q by $2m+n$ we have $n+2m+4 < 2m+n+3k$ or $4 < 3k$. Thus there are not sufficient available vertices to supply the cluster. This is a contradiction to the existence of the cluster and lemma 1 follows. ▮

Since the connected cluster contains vertices associated with each edge of G we know that the type G vertices in the cluster will form a connected vertex cover of some size p. If $p \leq k$ we are done.

It remains to show that the cluster will not have the required number of edges if it includes more than k type G vertices. The following lemma is useful.

Lemma 2. *A portion of the cluster in one of the maximal planar graphs containing $q-x$ vertices can contribute at most $(3q-6)-3x$ edges.*

Proof: A maximal planar graph with n vertices contains $3n-6$ edges. A n vertex maximal planar graph contains the maximum number of edges possible for an n vertex planar graph. Thus the position of the cluster with $q-x$ vertices can contribute at most $3(q-x)-6=(3q-6)-3x$ edges. ▮

The following lemma completes the proof of the claim.

Lemma 3. *Any connected cluster in G' with $b=qm+k+2m+2(k-1)$ vertices with $p > k$ type G vertices will have fewer than $m(3q-6)+3m+3(k-1)$*

edges.

Proof: Consider a connected cluster C in G' with b vertices of which $p > k$ are type G vertices. By lemma 1, C contains vertices in each of the maximal planar graphs associated with the edges of G. For some $x \geq 0$, C will contain $qm - x$ type M vertices. C must also contain $2m$ type S vertices of G' to connect the type M vertices to the type G vertices. Since C is connected, C must also contain $2(p-1)$ type S vertices which are the 2 remaining type S vertices in each of the edges in a spanning tree of the subgraph of G induced by the type G vertices in the cluster. There will also be y additional type S vertices in C where $y = x - 3(p-k)$. By lemma 2 if x vertices are missing from the maximal planar graphs then $3x$ edges will be lost. Thus the $3m - x$ type M vertices in M will contribute at most $m(3q-6) - 3x$ edges. Let all edges in C which lie outside of the maximal planar graphs be charged to the type S vertices in C. These vertices can contribute edges to the cluster at a maximum rate of 3/2. Thus the type G and type S vertices in C contribute at most $3m + 3(p-1) + 3/2y$ edges. Thus the maximum possible number of edges in C is $m(3q-6) - 3x + 3m + 3(p-1) + 3/2(x - 3(p-k))$. The lemma follows if the above quantity is less than $m(3q-6) + 3m + 3(k-1)$ that is if $-3/2x - 3/2p + 3/2k < 0$ or if $0 < x + (p-k)$ which is true since $x \geq 0$ and $p > k$. ∎

3. Conclusions

In this paper we have shown that the connected h-cluster problem on planar graphs is NP-complete. However if the cluster is not required to be connected the complexity of the h-cluster problem on planar graphs remains open.

Acknowledgement

The authors wish to thank the Natural Science and Engineering Research Council of Canada for financial support.

References

1. S. Arnborg, J. Lagergren and D. Seese, *Problems Easy for Tree-Decomposable Graphs*, Proceedings of the 15th International Colloquium on Automata, Languages and Programming (1988), 38–51.
2. B.S. Baker, *Approximation algorithms for NP-complete problems on planar graphs*, Proc. 26th Symp. Foundations of Computer Science (1983), 265–273.
3. H.L. Bodlaender, *Classes of Graphs with Bounded Treewidth*, TR RUU-CS-86-22, Dept. of Computer Science, University of Utrecht (1986).
4. H.L. Bodlaender, *Dynamic Programming on Graphs with Bounded Treewidth*, Proceedings of the 15th International Colloquium on Automata, Languages and Programming (1988), 105–118.

5. D.G. Corneil and Y. Perl, *Clustering and Domination in Perfect Graphs*, Discrete Applied Math. **9** (1984), 27–39.
6. E.S. El-Mallah and C.J. Colbourn, *On Two Classes of Planar Graphs*, Discrete Mathematics **80** (1990), 21–40.
7. H. Everett and A. Gupta, *Clustering in Regular Graphs*, (in preparation).
8. M.R. Garey and D.S. Johnson, *The Rectilinear Steiner Tree Problem is NP-complete*, SIAM J. Appl. Math. **32,4** (1977), 826–834.
9. J.M. Keil and T.B. Brecht, *Clustering in Planar Graphs*, Research Report 88 **5** (1988). Department of Computational Science, University of Saskatchewan.
10. K. Takamizawa, T. Nishizeki and N. Saito, *Linear-time computability of combinatorial problems on series-parallel graphs*, J. ACM **29** (1982), 623–641.
11. J. Wald and C.J. Colbourn, *Steiner trees, partial 2-trees, and minimum IFI networks*, Networks **13** (1983), 159–167.