Hardness of edge-modification problems

Noga Alon * Uri Stav †

September 20, 2008

Abstract

For a graph property $\mathcal{P}$ consider the following computational problem. Given an input graph G, what is the minimum number of edge modifications (additions and/or deletions) that one has to apply to G in order to turn it into a graph that satisfies $\mathcal{P}$? Namely, what is the edit distance $\Delta(G,\mathcal{P})$ of a graph G from satisfying $\mathcal{P}$. Clearly, the computational complexity of such a problem strongly depends on $\mathcal{P}$. For over 30 years this family of computational problems has been studied in several contexts and various algorithms, as well as hardness results, were obtained for specific graph properties.

Alon, Shapira and Sudakov studied in [3] the approximability of the computational problem for the family of *monotone* graph properties, namely properties that are closed under removal of edges and vertices. They describe an efficient algorithm that achieves an $o(n^2)$ additive approximation to $\Delta(G,\mathcal{P})$ for any monotone property $\mathcal{P}$, where G is an n-vertex input graph, and show that the problem of achieving an $O(n^{2-\varepsilon})$ additive approximation is NP-hard for most monotone proeprties. The methods in [3] also provide a polynomial time approximation algorithm which computes $\Delta(G,\mathcal{P}) \pm o(n^2)$ for the broader family of *hereditary* graph properties (which are closed under removal of vertices). In this work we introduce two approaches for showing that improving upon the additive approximation achieved by this algorithm is NP-hard for several sub-families of hereditary properties. In addition, we state a conjecture on the hardness of computing the edit distance from being induced H-free for any forbidden graph H.

1 Introduction

In a graph modification problem one is asked what is the minimum number of modifications that need to be applied to an input graph in order to attain some property. The problems vary in the desired property and the type of modifications that are allowed. In a vertex-modification problem,

*Schools of Mathematics and Computer Science, Raymond and Beverly Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv 69978, Israel. Email: nogaa@tau.ac.il. Research supported in part by the Israel Science Foundation, by the Hermann Minkowski Minerva Center for Geometry at Tel Aviv University and by the Von Neumann Fund.

†School of Computer Science, Raymond and Beverly Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv 69978, Israel. Email: uri.stav@gmail.com .

one is allowed to add/remove vertices from the input graph. In an edge-modification problem the edge set of the graph is modified, i.e. edges can be added or removed. More restricted problems are edge-deletion (completion) in which one is only allowed to remove (add) edges from (to) the input graph. In their collection of NP problems, Garey and Johnson [19] mentioned 18 different types of vertex and edge-modification problems. In this work we focus on edge-modification problems. For a graph property $\mathcal{P}$, the edit distance of a graph G from the property $\mathcal{P}$ is denoted by $\Delta(G, \mathcal{P})$, which is exactly the desired output of the edge-modification problem for $\mathcal{P}$.

The study of edge-modification problems is motivated by applications in several fields and hence they were extensively studied in the past 30 years. Some of the important applications of computing $\Delta(G, \mathcal{P})$ for appropriately defined properties $\mathcal{P}$ include Numerical Algebra [26], Molecular Biology (see [14], [20] and [22]), Circuit Design [16], Machine Learning [11] and other combinatorial optimization problems. Nevertheless, it can be observed that in almost all the applications the focus is on graph properties that are hereditary, namely closed under removal of vertices (equivalently, closed under taking induced subgraphs). Let us briefly demonstrate some results which were obtained along the years:

- It has been shown that various edge-deletion problems for specific monotone graph properties are NP-hard, see e.g. Yannakakis [29], Asano and Hirata [10] and Asano [9].
- Natanzon, Shamir and Sharan [24] studied the hardness of the edge-modification / deletion / completion problems for a few natural hereditary graph properties, such as being Chordal, Perfect, and Comparability. They showed that some of the modification problems are NP-hard to approximate within some constant multiplicative factor, and yet give a polynomial approximation algorithm within a constant multiplicative factor for any hereditary property defined by a finite set of forbidden induced subgraphs.
- Clustering problems motivated the study of computing the edit distance of a graph from being a disjoint collection of cliques. This is equivalent to being induced $K_{1,2}$-free, and sometimes referred to as Cluster graphs. This problem was also shown to be NP-hard to compute ([27], see also [11]).

However, the arguments in most of the afore mentioned proofs, as well as others not discussed here, are ad-hoc.

A *monotone* graph property is closed under removal of edges and vertices, hence for a monotone graph property the edge-deletion and edge-modification problems coincide. In [3], Alon, Shapira and Sudakov provide the first result which holds for a large family of graph properties, namely all monotone graph properties. They describe, for every monotone graph property $\mathcal{P}$ and $\varepsilon > 0$, a polynomial time algorithm for approximating the edit distance of a given input graph on n vertices from $\mathcal{P}$. The algorithm obtains an additive approximation within εn^2 of the correct edit distance (a formal and more general statement of this result will be given shortly). The authors of [3] also characterize the monotone properties for which this algorithm achieves essentially the best possible additive approximation (see Section 4 for a comprehensive discussion of this result).

A slightly modified version of the algorithm in [3] provides an approximation algorithm for edge-modification problems for the broader family of hereditary graph properties. A variant of the following result also follows implicitly from the connection between testing and estimating of graph properties which is established by Fischer and Newman in [17][1].

Theorem 1.1. **([3])** *For any fixed $\epsilon > 0$ and any hereditary property $\mathcal{P}$ there is a deterministic algorithm that given a graph G on n vertices computes in time $O(n^2)$ an integer Δ which satisfies*

$$|\Delta - \Delta(G, \mathcal{P})| \leq \epsilon n^2 \ .$$

It is asked in [3] for which hereditary properties this algorithm achieves essentially the best possible additive approximation. In other words, whether one can extend the characterization given for monotone properties in [3] to all *hereditary* graph properties. In this work we give a partial positive answer, focusing on some hereditary properties of the type being induced H-free, denoted by $\mathcal{P}^*_H$. We prove that for some families of graphs H, any essential improvement upon the additive approximation attained by Theorem 1.1 is NP-hard. To this end we explore two possible approaches.

Our main approach is presented in the next section, where we consider the case of H being a cycle. The proof consists of two stages: first proving hardness of the exact computation of $\Delta(G, \mathcal{P}^*_H)$, then amplifying the result to show the hardness of the additive approximation. In Section 3 we generalize the amplification method for a much broader family of graphs. Section 4 consists of the discussion of the second approach which is based on the reduction methods of [3]. This approach, however, seems to be more limited than the other one. In Section 5 we present some conjectures on the hardness of approximating $\Delta(G, \mathcal{P}^*_H)$ for any graph H.

2 Hardness of approximating $\Delta(G, \mathcal{P}^*_{C_\ell})$

The main result of this section is formulated as follows.

Theorem 2.1. *It is NP-hard to approximate the distance $\Delta(G, \mathcal{P}^*_{C_\ell})$ within an additive error of $O(n^{2-\eta})$ for any positive $\eta > 0$ and $\ell \geq 4$.*

Clearly the property $\mathcal{P}^*_{C_\ell}$ is hereditary, hence together with Theorem 1.1 this gives a nearly tight bound on the additive approximation which can be achieved in polynomial time for these

[1] Fischer and Newman [17] show that if a graph property is testable, then one can also estimate a graph's distance from satisfying the property using a constant size sample of the graph. Together with the result of Alon and Shapira [2] this shows that every hereditary property is estimable, and the sampling algorithm in fact provides a randomized algorithm for estimating the distance from satisfying the property.

properties (assuming $P \neq NP$). The proof of the theorem consists of two main steps. In the first one we show that computing the exact edit distance $\Delta(G, \mathcal{P}^*_{C_\ell})$ is NP-hard. The second step is an amplification of the result, that is, reducing the exact computation of $\Delta(G, \mathcal{P}^*_{C_\ell})$ to the problem of obtaining an additive approximation to it.

Recall that for a graph $G = (V, E)$, a set of vertices $X \subseteq V$ is a vertex cover of G is for every edge $xy \in E$ of the graph at least one of its endpoints belongs to the vertex cover: $\{x, y\} \cap X \neq \emptyset$. Note that the complement of a vertex cover is an independent set in the graph. It is well known that it is NP-Hard to find the size of the minimum vertex cover of an input graph G. Thus the starting point of our reductions in this section is the problem of determining the size of a minimum vertex cover in a graph.

Lemma 2.2. *For any fixed $\ell \geq 4$, it is NP-Hard to compute for a given input graph G the distance $\Delta(G, \mathcal{P}^*_{C_\ell})$.*

Proof. We establish a (mapping) reduction from minimum vertex cover as follows. Given a graph $G = (V, E)$ on $|V| = n$ vertices and $|E| = m$ edges, let k denote the size of a minimum vertex cover for G. Our goal is to construct a graph G' such that $\Delta(G', \mathcal{P}^*_{C_\ell}) = k$. We obtain G' from G as follows.

We add a new vertex x which is connected to all the vertices in G, and replace any edge of G by a path of length $\ell - 2$ which connects its endpoints in G'. Therefore, in G' there are $n' = 1 + n + (\ell - 3)m$ vertices and $m' = n + (\ell - 2)m$ edges. We refer to the vertex x as the *special vertex*, the vertices that correspond to the vertices of G as the *old vertices* and the inner vertices of the paths (which replace the edges of G) as the *new vertices* in G'.

The first observation about G' is that *any* copy of C_ℓ in G' must contain x. Hence, given a vertex cover of G, suppose we remove edges in G' from x to all the members of the vertex cover. In what is left, x is connected to a set of vertices which correspond to an independent set of G. It now follows that none of the copies of C_ℓ in G' survives, implying that $\Delta(G', \mathcal{P}^*_{C_\ell}) \leq k$.

On the other hand, consider a minimum set A of vertex pairs in G', of size $a = |A|$, whose modification destroys all the induced copies of C_ℓ. We claim that we can find a set B of at most a edges that are incident to x such that its removal destroys all the copies of C_ℓ in G'. We construct B from A as follows:

- If a pair in A consists of two new vertices which come from two different paths, then we do not add it to B.
- If a pair in A consists of two new vertices which come from the same path, then we add xv to B where v is one of the endpoints of that path.
- If a pair in A consists of a new vertex y and an old vertex v, then we add xv to B.

- If a pair vu that consists of two old vertices belongs to A, then we add xv to B (arbitrarily choosing either v or u).
- If a pair in A consists of the special vertex x and some new vertex y, then we add an edge xv to B, where v is an old vertex which is one of the endpoints of the path to which y belongs.
- If a pair in A consists of x and an old vertex v, we simply add it to B.

This defines the treatment of all the possible pairs in A. Clearly, $|B| \leq |A|$ (note that several different edges in A may correspond to the same edge in B). However, in any induced copy of C_ℓ in G', at least one of its vertex pairs appears in A. Hence, by our choice of B, that pair corresponds to an edge in B which is contained in that C_ℓ. Therefore, by removing all the edges of B from G' indeed all the C_ℓs are destroyed. In other words, removing at most a edges which are incident to x indeed turns G' into an induced C_ℓ-free graph. However, by the construction of G', the set of old vertices which are incident to those edges must establish a vertex cover in G (otherwise, an uncovered edge in G would correspond to an induced copy of C_ℓ in G' after the removal of B). Thus, $k \leq a$ which proves the validity of the reduction.

■

Remark 2.3. *It is not difficult to show that Lemma 2.2 holds also in the case $\ell = 3$. This is possible due to the result of Poljak [25] (see also Theorem 2.1 in [21]) who showed that computing the minimum vertex cover is NP-hard also for triangle-free graphs. Thus, when starting the reduction with a graph G which is guaranteed to be triangle-free, then all the triangles in G' must contain the special vertex x. Hence, again, a minimum set of edges whose modification destroys all the triangles in G' corresponds to a minimum vertex cover in the original graph G. However, the hardness result for $H = C_3$ also follows from [3] since triangle freeness is a monotone property.*

Remark 2.4. *Dinur and Safra [15] proved that approximating the size of a vertex cover within a multiplicative factor of ≈ 1.36 is NP-hard. Their result implies by Lemma 2.2 that approximating the edit distance of a graph from $\mathcal{P}^*_{C_\ell}$ within a multiplicative factor of ≈ 1.36 is NP-Hard. Note that Theorem 1.1 provides an approximation algorithm with a multiplicative factor of $(1+\varepsilon)$ for any positive ε, however this ratio only holds for dense graphs (namely graphs with $\Omega(n^2)$ edges).*

It should be noted that for the reduction of Lemma 2.2 the graph which is constructed is rather sparse, having a linear number of edges. Thus, approximating its edit distance within an additive factor of $n^{2-\eta}$ is trivial for $0 < \eta < 1$. Our next step towards the proof of Theorem 2.1 is an amplification of the above result. By appropriately blowing up the input graph, we are able to enlarge the edge density of the graph together with the gap in the distance from $\mathcal{P}^*_{C_\ell}$. To this end we use a clique blow-up: each vertex in the graph is replaced by a cluster of vertices that spans a clique, and every edge is replaced by a complete bipartite graph between the clusters which correspond to its endpoints. The intuition behind the choice of this type of blow-up is provided in Section 3.

Proof of Theorem 2.1:
For any positive constant $0 < \eta < \frac{1}{4}$, fix $b = b(n, \eta) = n^{\frac{1}{2\eta^2}}$. Our proof in this case is by a reduction from computing the exact distance, which is NP-hard by Lemma 2.2. Given an input graph G, assume that $\Delta(G, \mathcal{P}^*_{C_\ell}) = k$.

We construct a graph G_b which is the b- clique blow-up of G, that is a clique blow-up in which each vertex in G is replaced by a cluster of b vertices in G_b.

Claim 2.5. $\Delta(G_b, \mathcal{P}^*_{C_\ell}) = kb^2$.

Proof. First, suppose we modify in G_b the kb^2 edges of the bipartite graphs that correspond to a set of k edges in G whose modification makes G induced C_ℓ-free. Call this new graph G'_b. Note that also in G'_b a pair of vertices in the same cluster have exactly the same set of neighbors (excluding themselves). Hence, assume towards a contradiction, that G'_b contains an induced C_ℓ. In this case, each of its ℓ vertices comes from a different cluster. Therefore, such an induced copy of C_ℓ in G'_b implies that by taking the vertices in G that correspond to those ℓ clusters, it must also span a C_ℓ which is an induced subgraph of the modified G. This leads to a contradiction, thus proving the first direction, namely $\Delta(G_b, \mathcal{P}^*_{C_\ell}) \leq kb^2$.

On the other hand, assume $k' = \Delta(G_b, \mathcal{P}^*_{C_\ell})$ and that G'_b is the closest induced C_ℓ-free graph to G_b. Pick uniformly independently one vertex from each cluster in G'_b. This is an induced subgraph of G'_b, and hence induced C_ℓ-free. The expected number of modifications (between G_b and G'_b) that are spanned by these vertices is at most k'/b^2. By applying these modifications to G, one obtains an induced C_ℓ-free graph. Thus some choice of such representatives shows that $k = \Delta(G, \mathcal{P}^*_{C_\ell}) \leq k'/b^2$ which completes the proof of the claim. ∎

In G_b there are bn vertices, and hence the output of an algorithm achieving an additive approximation of $n^{2-\eta}$ would give us $d = \Delta(G_b, \mathcal{P}^*_{C_\ell}) \pm (bn)^{2-\eta} = b^2k \pm (bn)^{2-\eta}$. However, for our choice of b and a sufficiently large n, we get that the error term is at most

$$(bn)^{2-\eta} = n^{(1+\frac{1}{2\eta^2})(2-\eta)} = n^{\frac{1}{\eta^2}} n^{2-\eta-\frac{1}{2\eta}} = b^2 o(1) < \frac{1}{2}b^2 \ .$$

Therefore, in this case, kb^2 can be determined from d, and hence also k. We conclude that a polynomial time algorithm for finding such d would imply a polynomial algorithm for finding the exact edit distance $\Delta(G, \mathcal{P}^*_{C_\ell})$. ∎

3 Robustness of the amplification method

As the proof for the case $H = C_\ell$ demonstrates, the hardness of approximation of $\Delta(G, \mathcal{P}^*_H)$ may be proved in two steps. The first one showing that it is NP-hard to exactly compute the distance $\Delta(G, \mathcal{P}^*_H)$, where the second step is showing that an additive approximation can be utilized to deduce the exact value of $\Delta(G, \mathcal{P}^*_H)$. In this section we analyze the robustness of the amplification method which was used for the case $H = C_\ell$. We show that it holds for a large family of (forbidden) graphs H. This, in turn, implies possible extension of the hardness of additive approximation as detailed in Corollary 3.3.

The following definition will be very useful throughout our discussion. For any graph H define a binary relation R_H on its vertex set:

$$xR_Hy \iff N(x) \cup \{x\} = N(y) \cup \{y\} \ ,$$

where $N(x)$ denotes the neighbors of x in H. Note that for xR_Hy it is necessary that xy is an edge of H. It is not difficult to verify that this is an equivalence relation. Let $\hat{H}$ be an induced subgraph of H obtained by picking one vertex from each equivalence class of R_H. Note that all such choices result in graphs isomorphic to $\hat{H}$. We use the following simple claim regarding $\hat{H}$. Recall that a clique blow-up of a graph G is obtained by replacing the vertices of the graph by cliques, and replacing an edge by a complete bipartite graph between the cliques which correspond to the endpoints of that edge.

Claim 3.1. *If a graph $G \in \mathcal{P}^*_{\hat{H}}$ is induced $\hat{H}$-free, and F is a clique blow-up of G, then $F \in \mathcal{P}^*_H$ is induced H-free.*

Proof. Suppose $V(F) = V_1 \cup \ldots \cup V_n$ where each V_i is a clique blow-up of vertex i in G. Note that in F, two vertices in the same V_i are connected and have the same set of neighbors. Assume towards a contradiction that F contains an induced copy of H. Any pair of vertices from different equivalence classes in H cannot appear in the same V_i. Now, if we take the induced copy of H in F and pick one vertex from each V_i (or none if there are no vertices of H), we get a graph that contains at least one member from each equivalence class of H, and thus contains $\hat{H}$ as an induced subgraph. Since there is at most one member from each V_i, this is also an induced subgraph of G, contradiction.

■

Note that Claim 3.1 does not necessarily assume that each vertex of G is replaced (i.e. blown-up) by a clique of the same size. However, the cases which justify the discussion are the graphs H that satisfy $H = \hat{H}$. A special case of those graphs are the cycles C_ℓ of length $\ell \geq 4$ which were discussed in the previous section. The generalized amplification result follows from the next claim.

Claim 3.2. *Let H and G be arbitrary graphs, and assume $H = \hat{H}$. Further assume $k = \Delta(G, \mathcal{P}^*_{\hat{H}})$, and that F is a b-clique-blow-up of G. Then $\Delta(F, \mathcal{P}^*_H) = kb^2$.*

Proof. First, suppose we modify in F the kb^2 edges that correspond to a set of k edges in G which makes G induced $\hat{H}$ free. The resulting graph is in fact a clique blow-up of some member in $\mathcal{P}^*_{\hat{H}}$, and by Claim 3.1 it is in $\mathcal{P}^*_H$. Thus $\Delta(F, \mathcal{P}^*_H) \leq kb^2$.

On the other hand, suppose we modify $s = \Delta(F, \mathcal{P}^*_H)$ edges in F thus turning it into an induced H-free graph denoted F'. Pick uniformly independently one vertex from each blown-up set in F'. This is an induced subgraph of F', and thus it is induced $H = \hat{H}$-free. The expected number of modifications that fall into this induced subgraph is at most s/b^2. Hence for some choice of representatives, viewing them as the vertices of G shows that $k = \Delta(G, \mathcal{P}^*_{\hat{H}}) \leq s/b^2$ which completes the proof.

■

Claim 3.2 and the last part of the proof of Theorem 2.1 imply the following corollary.

Corollary 3.3. *Assume a graph H satisfies*

1. *All the equivalence classes of R_H are of size 1, namely $H = \hat{H}$.*
2. *It is NP-hard to exactly compute $\Delta(G, \mathcal{P}^*_H)$.*

Then for any positive $\eta > 0$, it is NP-hard to approximate $\Delta(G, \mathcal{P}^*_H)$ within an additive error of $n^{2-\eta}$.

Remark 3.4. *At this point it should be noted that we could also define an equivalence relation in which two vertices are related if they have the same set of neighbors and they are* not *connected. This yields a similar amplification result on graphs in which no two vertices are related. Here however this will be achieved by an independent set blow-up, that is a blow-up in which each vertex is replaced by an independent set. Nevertheless, equivalently, this can be achieved by taking the complement of H, and then using the relation R_H as defined before.*

4 Generalizing the hardness proofs from monotone to hereditary properties

As noted before, Alon, Shapira and Sudakov [3] obtain a characterization of the monotone properties for which the algorithm of Theorem 1.1 achieves essentially the best possible approximation as follows:

Theorem 4.1. (Theorem 1.3 in [3]) *Let $\mathcal{P}$ be a monotone graph property. Then,*

1. *If there is a bipartite graph that does not satisfy $\mathcal{P}$, then there is a $\delta > 0$ for which it is possible to approximate $\Delta(G, \mathcal{P})$ to within an additive error of $n^{2-\delta}$ in linear time.*
2. *On the other hand, if all bipartite graphs satisfy $\mathcal{P}$, then for any $\delta > 0$ it is NP-hard to approximate $\Delta(G, \mathcal{P})$ to within an additive error of $n^{2-\delta}$.*

In this section we discuss the possibility of extending the proof of Theorem 4.1 to obtain a similar characterization for hereditary properties. Our emphasis here is on the reduction method which was used in [3]. We thus start by sketching this method. We then outline a proof suitable for hereditary graph properties, which basically follows the approach of [3]. Our final remarks here concern the limitations of this method.

4.1 The reduction for monotone properties

The proof of the first part of Theorem 4.1 follows from the results on the Turán numbers of bipartite graphs. Let us elaborate on the method for proving part 2 in Theorem 4.1. Following the basic approach of [1], a blow-up of a sparse instance to a problem is embedded in an appropriate dense pseudo-random graph. The core of the proof is based on Theorem 4.2 below. Let H be an arbitrary graph, and denote the graph property of excluding a (not necessarily induced) copy of H by $\mathcal{P}_H$. The theorem shows that for any graph G with sufficiently large minimal degree, in order to make it H-free in the most economical way, one needs to modify G so that it becomes almost r-colorable, where $\chi(H) = r + 1$. The theorem simultaneously extends the results of Erdős-Stone-Simonovits (cf. [12] and [28]) and Andrásfai-Erdős-Sós [8].

Theorem 4.2. (part of Theorem 6.1 in [3], see also [7]) *Let H be a graph of chromatic number $r + 1 \geq 3$. There are constants $\gamma = \gamma(H) > 0$ and $\mu = \mu(H) > 0$ such that if $G = (V, E)$ is a graph on n vertices of minimum degree at least $(1 - \mu)n$, then*

$$\Delta(G, \mathcal{P}_{r,0}) - O(n^{2-\gamma}) \leq \Delta(G, \mathcal{P}_H) \leq \Delta(G, \mathcal{P}_{r,0})$$

where $\mathcal{P}_{r,0}$ denotes the graph property of being r-colorable.

We now sketch the main ingredients of the reduction. Let $\mathcal{M}$ be an arbitrary monotone property, and suppose H is a graph not satisfying $\mathcal{M}$ with minimal chromatic number $\chi(H) = r+1 \geq 3$. The reduction is from the problem of computing the (exact) edit distance of an input graph from being r-colorable, which is known to be NP-hard for $r \geq 2$. Hence, the instance is a graph G for which one has to decide how many edges need to be removed in order to make it r-colorable. The input to the edge-modification problem, F is roughly constructed by taking the "Boolean Or" of two graphs. The first graph is a pseudo-random dense graph, with degrees $(1 - \mu)n$. The second is a disjoint union of r copies of a blow-up of the original graph by a polynomial factor. The "Boolean Or" of the two graphs is obtained by identifying their vertices arbitrarily (the order has no significance, as one of them is pseudo-random) and connecting two vertices if they are connected in at least one of the graphs. Thus, each edge e of G results in F in a slightly higher edge density (1 instead of $1 - \mu$) of the bipartite graph between the clusters which correspond to the endpoints of e in G. Theorem 4.2 implies that approximating the edit distance of F from $\mathcal{M}$ within at most $n^{2-\gamma(H)}$ enables one to compute the exact number of edges that need to be removed from G in order to make it r-colorable. Intuitively, this last observation follows from the fact that the closest graph in $\mathcal{M}$ is essentially r colorable with equal sized sets. Therefore, the contribution of the pseudo-random graph to the number of edges that need to be deleted is independent of the partition into r color classes. Hence altogether the distance of F from $\mathcal{M}$ mainly depends on the distance of G from being r-colorable.

4.2 A method for reducing hereditary properties

We would like to borrow some ideas from the above method to establish a reduction for hereditary properties. As we shall see shortly, most of the ingredients have to be modified. We first need to extend the usual notion of graph coloring as follows.

Definition 4.3. *For any pair of integers (r, s), such that $r+s > 0$, we say that a graph $G = (V, E)$ is* ***(r,s)-colorable*** *if there is a partition of V into $r+s$ (possibly empty) subsets $I_1, \ldots, I_r, C_1, \ldots, C_s$ such that each I_k induces an independent set in G, and each C_k induces a clique in G.*

Thus, in particular, $(r, 0)$-colorable graphs are r-colorable graphs. We denote by $\boldsymbol{\mathcal{P}_{r,s}}$ the graph property comprising all the (r, s)-colorable graphs.

Going back to the reduction, we narrow our discussion to properties of the family $\mathcal{P}^*_H$, and apply the following changes to the method of [3]:

- Let (r, s) be such that H is *not* (r, s) colorable. The reduction is from the problem of determining what is the exact edit distance of an input graph from being (r, s)-colorable, instead of the distance from being r-colorable. There is a simple reduction showing these problems are NP-hard when $\max\{r, s\} \geq 2$. [2] In addition, it is also needed in this case that it is hard to compute the edit distance from having an (r, s)-coloring in which all color classes have equal sizes. This slight refinement can be deduced by a straightforward reduction.
- The dense pseudo-random graph is replaced by a pseudo-random graph with edge density $\frac{1}{2}$.
- The "Boolean Or" operation is changed to adding/removing (pseudo-)random edges in order to increase/decrease the edge density of the pseudo-random graph according to the input graph. We consider a blow-up of the original graph G, identify its vertices with some vertices of the pseudo-random graph, and then slightly decrease (increase) the edge-density between clusters that correspond to vertices that are not connected (connected) in G. The modifications are applied so that the resulting graph is very close to the original pseudo-random graph, that is, the number of changes applied to edges touching each vertex in the pseudo-random graph is bounded by δn, where δ is a positive constant that depends on H.

Yet, the most important ingredient for proving the correctness of the above reduction is an extremal result, analogous to Theorem 4.2. Roughly, it states that modifying the random graph into an induced H-free graph in the most economical way is achieved by making it (r, s)-colorable. This basic idea is illustrated in the following theorem for the particular case $H = C_4$. Note that $\mathcal{P}_{1,1} \subset \mathcal{P}^*_{C_4}$.

[2] In any other case, namely $\max\{r, s\} \leq 1$, the problem is polynomial. For $(1, 0)$ and $(0, 1)$ this is trivial, while for $(1, 1)$ there is a simple polynomial time algorithm for computing the edit distance from being a split graph ($(1, 1)$-colorable), as described in [23].

Theorem 4.4. **([6])** *Let $\hat{G}$ be the closest graph in $\mathcal{P}^*_{C_4}$ to $G = G(n, \frac{1}{2})$, i.e. $\hat{G}$ satisfies $\Delta(G, \hat{G}) = \Delta(G, \mathcal{P}^*_{C_4})$. Then, w.h.p., $\hat{G}$ is $(1,1)$-colorable.*

In a more general form of Theorem 4.4, C_4 can be replaced by $K_{r+1}(s+1)$ - the complete $(r+1)$-partite graph with $s+1$ vertices in each part, in which case $\hat{G}$ will be (r,s)-colorable. In order to use the ideas of [3], Theorem 4.4 - which, in a sense, is an analog of Turán's Theorem - should be extended in the following ways:

- Theorem 4.4 holds even if one applies arbitrary changes to G, such that for any vertex v in the graph, the number of modifications in edges touching v does not exceed δn for some $\delta > 0$. This may seem similar to the minimal degree restriction of Theorem 4.2.
- Replacing the random graph by a pseudo-random graph with edge density $\frac{1}{2}$, the above results still hold. Hence, in this case the statement of Theorem 4.4 *always* holds (not just w.h.p.).

We note that the choice of $G(n, 1/2)$ makes the proof of Theorem 4.4 possible, since fundamental parts of the proof rely on the symmetry in the number of edge additions and deletions one has to apply in order to turn a subgraph of $G(n, 1/2)$ into either a clique or an independent set.

Recall that our construction results in a graph which is close to a pseudo-random graph, and thus its edit distance from being induced H-free equals its distance from being (r,s)-colorable with essentially equal sized sets. Therefore, similar to the argument for monotone properties, computing an approximation of its distance from being (r,s)-colorable enables us to accurately compute the distance of G from being (r,s)-colorable with equal sized sets. This therefore completes the sketch of the reduction.

4.3 The limits of this approach

It is clear that following this method, the extent of the hardness results strongly depends on the ability to generalize Theorem 4.4. So far, we were able to extend the reduction for properties $\mathcal{P}^*_H$ for a specific family of graphs, namely $K_{r+1}(s+1)$ ($\max\{r,s\} \geq 2$). These graphs seem to capture some special extremal properties which generalize the special role that the complete graphs play in various Turán type problems.

On the other hand, it is observed in [6] that in some cases an extremal result of this type is not possible. Namely, for some natural graphs H, there might be two graphs with a very different structure which are essentially the closest graphs in $\mathcal{P}^*_H$ to $G(n, \frac{1}{2})$. Hence, it is not possible to relate the edit distance from satisfying $\mathcal{P}$ to the edit distance from having some fixed structure (such as being (r,s)-colorable), and this approach will not provide an appropriate reduction. Therefore the extent of our approach here is limited.

A possible extension can be obtained by a similar reduction in which the starting point is a pseudo-random graph with some edge density $p \neq \frac{1}{2}$. To this end, two steps have to be taken. The first is proving an analogue of Theorem 4.4 (with its extensions) for $G(n,p)$ and other hereditary properties. In other words, it may be possible that for any hereditary property there exist some $0 \le p \le 1$ for which an analogue of the theorem holds. The second step is modifying the reduction appropriately in order to obtain a reduction which suits this setting. However, both problems seems to be difficult, and at the moment it seems that this approach is less promising than the one discussed in the previous section.

5 Conjectures and concluding remarks

The first conjecture is the following:

Conjecture 5.1. *For any graph H on at least 3 vertices, it is NP-hard to compute $\Delta(G, \mathcal{P}^*_H)$ for an input graph G.*

Recall that Lemma 2.2 proves this conjecture for any cycle. The NP-hardness of the edge modification problem on Clique graphs [27] imply the same for $P_3 = K_{1,2}$. Nevertheless, let us point out that the construction of Lemma 2.2 can be extended to show that for many other graphs H a similar result holds. For instance, if H is a disjoint union of a cycle and a tree or when H is a blow-up of a complete graph (where at least two vertices are replaced by independent sets of size at least 2) and in particular when H is any complete bipartite graph $K_{p,q}$ with $\min\{p,q\} \ge 2$.

The way to obtain this type of results is to construct a graph G'' as follows. We start with the graph G' which is constructed in Lemma 2.2 from G, and then add an n^2-clique blow-up of $H \setminus C_\ell$ (i.e., the graph H from which ℓ vertices that span C_ℓ are removed). This blow-up should be carefully connected to the vertices of G' in a way that guarantees that (i) when removing edges from G' such that all the copies of C_ℓ are destroyed then all the copies of H in G'' are destroyed, and (ii) any copy of C_ℓ in G' participates in n^2 edge-disjoint copies of H in G''. Therefore, we are again guaranteed that the most economical way to destroy all the copies of H in G'' is by removing edges in G' that correspond to a vertex cover of G.

Using Corollary 3.3 and assuming Conjecture 5.1, the hardness of obtaining additive approximation for various properties $\mathcal{P}^*_H$ follows . Nevertheless, our main conjecture is much stronger, and in fact we believe that an amplification is possible for every graph, namely:

Conjecture 5.2. *For any graph H on at least 3 vertices, and any positive $\eta > 0$, it is NP-hard to approximate the edit distance $\Delta(G, \mathcal{P}^*_H)$ for an input graph G on n vertices within an additive error $n^{2-\eta}$.*

Note that when $v(H) = 2$ the problem is trivially polynomial. For $v(H) = 3$ this conjecture follows for the triangle (and its complement) by the results of [3] and for $K_{1,2}$ (and its complement) by a combination of [27] and Corollary 3.3. Hence it follows for every graph on 3 vertices. We believe that combining arguments similar to those described in the previous couple of sections might prove this conjecture.

It is far more complicated to decide if the above holds for other hereditary properties. Intuitively, it seems plausible that for *most* of the hereditary properties Theorem 1.1 is essentially tight. This coincides with the case of monotone properties, for which such a result is proved in [3]. There are also various specific hereditary (non-monotone) such properties. One natural approach would be to extend proofs for a single forbidden induced subgraph (like the properties we considered in this work) to other hereditary properties which are defined by several forbidden induced subgraphs. On the other hand, it is worth mentioning that forbidding several induced subgraphs might end up in a more degenerate property for which determining the edit distance becomes easy. For instance, it is known that split graphs are exactly the graphs that are induced $\{C_4, \overline{C_4}, C_5\}$-free ([18], see also [21] pp. 151-152), where the computation of the exact edit distance from being a split graph can be done in polynomial time (cf. [23]). Yet computing (or even approximating in an appropriate sense) the edit distance from being induced H-free for any $H \in \{C_4, \overline{C_4}, C_5\}$ is NP-hard.

References

[1] N. Alon, Ranking tournaments, SIAM J. Discrete Math. 20 (2006), 137-142.

[2] N. Alon and A. Shapira, A characterization of the (natural) graph properties testable with one-sided error, Proc. of the 46^{th} IEEE FOCS (2005), 429-438.

[3] N. Alon, A. Shapira and B. Sudakov, Additive approximation for edge-deletion problems, Proc. of the 46^{th} IEEE FOCS (2005), 419-428. Also: Annals of Mathematics, to appear.

[4] N. Alon and U. Stav, What is the furthest graph from a hereditary property?, Random Structures and Algorithms 33 (2008), 87-104.

[5] N. Alon and U. Stav, The maximum edit distance from hereditary graph properties, Journal of Combinatorial Theory, Ser. B 98 (2008), 672-697.

[6] N. Alon and U. Stav, Stability type results for hereditary properties, Journal of Graph Theory, to appear.

[7] N. Alon and B. Sudakov, H-free graphs of large minimum degree, The Electronic J. of Combinatorics 13(1) (2006), R19, 9pp.

[8] B. Andrásfai, P. Erdős and V. Sós, On the connection between chromatic number, maximal clique and minimal degree of a graph, Discrete Math. 8 (1974), 205-218.

[9] T. Asano, An application of duality to edge-deletion problems, SIAM J. on Computing, 16 (1987), 312-331.

[10] T. Asano and T. Hirata, Edge-deletion and edge-contraction problems, Proc. of the 14^{th} ACM STOC (1982), 245-254.

[11] N. Bansal, A. Blum and S. Chawla, Correlation clustering, Proc. of the $43^{\rm rd}$ IEEE FOCS (2002), 238-247.

[12] B. Bollobás, **Extremal Graph Theory**, Academic Press (1978).

[13] L. Cai, Fixed-parameter tractability of graph modification problems for hereditary properties, Information Processing Letters, 58 (1996), 171-176.

[14] K. Cirino, S. Muthukrishnan, N. Narayanaswamy and H. Ramesh, Graph editing to bipartite interval graphs: exact and asymptotic bounds, Proc. of 17^{th} FSTTCS (1997), 37-53.

[15] I. Dinur and S. Safra, On the hardness of approximating minimum vertex-cover, Annals of Mathematics 162(1) (2005),439-485.

[16] E. S. El-Mallah and C. J. Colbourn, The complexity of some edge-deletion problems, IEEE transactions on circuits and systems, 35 (1988), 354-362.

[17] E. Fischer and I. Newman, Testing versus Estimation of Graph Properties, Proc. of the 37^{th} ACM STOC (2005), 138-146. Also: SIAM J. Computing 37(2)(2007), 482-501.

[18] S. Földes, P.L. Hammer, Split Graphs, Proc. 8^{th} Southeastern Conf. on Combinatorics, Graph Theory and Computing (F. Hoffmann et al., eds.), Lousiana State Univ., Baton Rouge, Louisiana (1977), 311-315.

[19] M.R. Garey and D.S. Johnson, **Computers and Intractability: A guide to the Theory of *NP*-Completeness**, W.H. Freeman and Co., San Francisco, 1979.

[20] P. W. Goldberg, M. C. Golumbic, H. Kaplan and R. Shamir, Four strikes against physical mapping of DNA, Journal of Computational Biology 2 (1995), 139–152.

[21] M.C. Golumbic, **Algorithmic Graph Theory and Perfect Graphs**, Academic Press, New York (1980).

[22] M. C. Golumbic, H. Kaplan and R. Shamir, On the complexity of DNA physical mapping, Advances in Applied Mathematics, 15 (1994), 251-261.

[23] P. L. Hammer and B. Simeone, The splittance of a graph, Combinatorica 1 (1981), 275-284.

[24] A. Natanzon, R. Shamir and R. Sharan, Complexity classification of some edge modification problems, Discrete Applied Mathematics 113 (2001), 109–128.

[25] S. Poljak, A note on the stable sets and coloring of graphs, Comment. Math. Univ. Carolin. 15 (1974), 307-309.

[26] J. D. Rose, A graph-theoretic study of the numerical solution of sparse positive-definite systems of linear equations, Graph Theory and Computing, R.C. Reed, ed., Academic Press, N.Y. (1972), 183-217.

[27] R. Shamir, R. Sharan and D. Tsur, Cluster graph modification problems, Discrete Applied Mathematics 144(1-2)(2004), 173-182.

[28] D. B. West, **Introduction to Graph Theory**, Prentice Hall, 2001.

[29] M. Yannakakis, Edge-deletion problems, SIAM J. Comput. 10 (1981), 297-309.