algorithms

MDPI

Article

Efficient Approximation for Restricted Biclique Cover Problems

Alessandro Epasto [1,*,†] **and Eli Upfal** [2]

[1] Google Research, New York, NY 10011, USA
[2] Department of Computer Science, Brown University, Providence, RI 02912, USA; eli@cs.brown.edu
[*] Correspondence: aepasto@google.com; Tel.: +1-212-565-0000
[†] Work partially done while at Brown University.

Received: 31 March 2018; Accepted: 6 June 2018; Published: 12 June 2018

Abstract: Covering the edges of a bipartite graph by a minimum set of bipartite complete graphs (bicliques) is a basic graph theoretic problem, with numerous applications. In particular, it is used to characterize parsimonious models of a set of observations (each biclique corresponds to a *factor* or *feature* that relates the observations in the two sets of nodes connected by the biclique). The decision version of the minimum biclique cover problem is NP-Complete, and unless $P = NP$, the cover size cannot be approximated in general within less than a sub-linear factor of the number of nodes (or edges) in the graph. In this work, we consider two natural restrictions to the problem, motivated by practical applications. In the first case, we restrict the number of bicliques a node can belong to. We show that when this number is at least 5, the problem is still NP-hard. In contrast, we show that when nodes belong to no more than two bicliques, the problem has efficient approximations. The second model we consider corresponds to observing a set of independent samples from an unknown model, governed by a possibly large number of factors. The model is defined by a bipartite graph $G = (L, R, E)$, where each node in L is assigned to an arbitrary subset of up to a constant f factors, while the nodes in R (the independent observations) are assigned to random subsets of the set of k factors where k can grow with size of the graph. We show that this practical version of the biclique cover problem is amenable to efficient approximations.

Keywords: biclique cover; approximation algorithms; probabilistic models

1. Introduction

We study the approximability of two variants of the minimum biclique cover problem, motivated by the problem of characterizing parsimonious models in computational biology settings.

Given a bipartite graph $G = (L, R, E)$, a *biclique* in G is a set of edges $H \subseteq E$ that induces a complete bipartite subgraph of G. A *biclique cover* of G is a collection of bicliques that cover all the edges of G. The *biclique cover number* or the *biclique dimension* of G is the minimum number of bicliques that cover E.

Covering a bipartite graph with a minimum number of bicliques is a fundamental graph theory problem that has received significant attention in theoretical computer science [1–6]. The computational problem has numerous applications, ranging from computational biology [7–9], data mining [10], and machine learning [11], to automata theory [4], communication complexity [6], and graph drawing [12].

The problem of determining whether a bipartite graph G has a biclique cover of size k is known to be NP-Complete even if the graph is a chordal bipartite graph [5,13]. Chalermsook et al. [1] recently showed that the biclique cover problem is also not approximable in polynomial time to less than an $O(|V|^{1-\epsilon})$ or $O(|E|^{1/2-\epsilon})$ factor for any $\epsilon > 0$ unless $P = NP$. This improved previous inapproximability results [3,4]. Moreover, unless NP $\subseteq$ BPTIME$(2^{\text{poly}(\log(n))})$, there are also no polynomial time $O\left(\frac{|V|}{2^{\log^{7/8+\epsilon}(|V|)}}\right)$ or $O\left(\frac{\sqrt{|E|}}{2^{\log^{7/8+\epsilon}(|V|)}}\right)$ approximation algorithms [1]. These results are almost tight in terms of lower-order factors of $|V|$ as an $O\left(\min(\frac{|V|}{\sqrt{\log(|V|)}}, |E|\frac{(\log(\log(|E|))^2}{\log^3(|E|)})\right)$ approximation algorithm is known [1].

Given these strong impossibility results, theoretical work has focused on developing polynomial time solutions for specific graph subclasses, such as C4-free graph [13], and domino-free graphs [14]. Another line of research developed parametrized complexity results, showing that it is possible to determine in time $O(f(k)\text{poly}(n))$ whether a bipartite graph G admits a cover with at most k bicliques [2,7] where $f(k)$ is an exponential function of k, but it does not depend on n. Other work has focused on the version of the problem where bicliques are edge-disjoint or vertex-disjoint [1,2]. Dawande et al. [15] studied the size of maximum cardinality of bicliques in random bipartite graph. The related problem of clique cover in non-bipartite graphs has also received significant attention [5].

Finally, a problem related to our work here, is the local biclique cover problem [16], where for a given (not necessarily bipartite) graph G and a parameter f we want to determine whether there exists a biclique cover of the edges of G (using an arbitrary number of bicliques) in which each node belongs to at most f bicliques. Arora et al. [17] studied a generalization of clique cover with applications to community detection where nodes are limited to participate in a few cliques.

We note that the biclique cover problem is equivalent to the binary matrix factorization problem [10], where the goal is to decompose an $n \times m$ binary matrix A into a product of $n \times k$ and $k \times m$ binary matrices. The minimum k for which such a decomposition exists, known as the *Boolean rank* of A, equals the minimum biclique cover number of a bipartite graph G with adjacency matrix A. Biclique cover can also be seen as a form of biclustering [18].

Motivation

The two restricted biclique cover problems studied here are motivated by the application of minimum biclique cover to the fundamental inference problem of characterizing parsimonious models. We present combinatorial abstractions of two computational biology problems that motivates this work [7–9] . We note that similar applications are found in other fields [10].

The goal of HLA (human leukecyte antigen) serology studies [9] is to identify antigens responsible for acceptance and rejections of transplanted tissues. For the combinatorial model, it is sufficient to know that an implant is rejected if it has an *antigen* that has a specificity in common with some *antibody* in the recipient serum. Thus, a donor and recipient are not compatible if the donor has an antigen that has specificity in common with at least one of the recipient antibodies. Given a set of donors and recipients, we construct a bipartite graph $G = (L, R, E)$, where L is the set of donors, R is the set of recipients, and $v \in L$ is connected with $u \in R$ if v and u are not compatible. The goal is to determine the minimum number of (antigens, antibodies) pairs with common specificity that explains the observed compatibility structure between the sets of donors and recipients. It is easy to verify that a (antigens, antibodies) pair with common specificity defines a biclique in the graph G and therefore the minimum number of bicliques corresponds to the minimum number of pairs that explain the observed data.

A second problem is the Mod-Resc Parsimony Inference problem [7]. Here, the bipartite graph $G = (L, R, E)$ represents reproductive incompatibility between strains of male L, and female R insects. An edge (u, v) indicates that a cross between $u \in L$ and $v \in R$ is incompatible. The biology model postulates that incompatibility is a result of a bacteria (mod-factor) in the male sperm and a lack of the corresponding (resc-factor) in the female. The goal is to find the minimum number of (mod, resc) factor pairs that explains the data. Since the edges of the graph correspond to incompatible pairs, it is easy to verify that a biclique corresponds to a factor, and a minimum biclique cover gives a parsimonious model.

In these two applications, it is reasonable to assume that the total number of different factors (antigens or bacteria) is growing with the size of the sample. Thus, a parameterized solution, exponential in the total number of bicliques, is not practical. On the other hand, it is reasonable to assume that the number of different antigens or antibodies of a given person, or the number of different bacteria infecting one insect, is not a function of the total sample size. Furthermore, in particular in the Mod-Resc model, evaluation considerations imply that the number of factors per sample must be very small, justifying the first restricted model we study here.This is also consistent with the observation in [8] that in their data for a weighted version of the biclique cover problem each node was covered by at most 2 *weighted* bicliques.

A second characteristic of the observed data is that it corresponds to independent samples. To incorporate the effect of random samples, we distinguish between two cases. Two factors can be highly correlated. For example, in the case of HLA serology, two antigens may appear together in donors and their corresponding antibodies may appear together in recipients. In that case we cannot distinguish between the two factors, as they correspond to the same biclique. Otherwise, we can assume that appearances of factors are relatively independent in at least one of the two sets of nodes. This observation motivates the second, stochastic model we study here.

2. Overview of New Results

Motivated by the above applications, we study two variations of the biclique cover problem. We present almost tight positive and negative results for each of the variants.

2.1. Model I: A Node belongs to a Small Number of Bicliques

In this model, we consider bipartite graphs with minimum biclique covers in which each node participates in only a small number of different bicliques. More formally, we introduce the (f_L, f_R, k)-biclique cover decision problem for bipartite graphs:

Problem 1 ((f_L, f_R, k)-biclique cover problem)**.** *Given a bipartite graph $G = (L, R, E)$, and parameters f_L, f_R and k, distinguish between the following two cases:*
Yes-instance: There is a biclique cover of G with $\leq k$ bicliques, such that $\forall u \in L$ (resp. $v \in R$), u belongs to at most f_L distinct bicliques in the cover (resp. v belongs to at most f_R distinct bicliques in the cover).
No-instance: No such cover exists.

When $f_L = f_R = f$, we write (f, k)-biclique cover instead of (f, f, k)-biclique cover.

Notice that this problem is related to the local biclique cover problem where one is only interested in minimizing the number of distinct bicliques covering each node without minimizing the total number of bicliques [16]. Clearly, if a graph satisfies the (f_L, f_R, k)-biclique cover condition it also satisfies the $(\max(f_L, f_R))$-local biclique cover condition, but the converse is not always true.

Results 1. *Our first result (Section 3.3) shows that the (f, k)-biclique problem remains computationally hard even for small values of f. Specifically, we prove that the (f, k)-biclique problem is NP-Complete for any $f \geq 5$.*

On the positive side, we first note that the $(1, k)$-biclique case is trivial since the graph must consist of disjoint bicliques. We present two results for the non-trivial $(2, k)$-biclique case. (1) For any graph with a $(2, k)$-biclique cover, we present a polynomial time algorithm that finds a biclique cover with at most $\approx 2.83k$ bicliques. (2) For any graph with a $(2, k)$-biclique cover, we present a polynomial time algorithm that covers at least a $1 - \frac{1}{e}$ fraction of the edges of the graph with k bicliques. These algorithms do not guarantee that each node is covered by at most 2 bicliques.

2.2. Model II: A One-Side Stochastic Model

In this model, we capture the fact that nodes represent independent observations, but the bipartite graph has more structure than a standard random bipartite graph. This structure is dictated by the semi-random assignment of factors to nodes.

Formally, we define the $(f, \bar{p}, k)$-biclique cover problem, for a constant f, a vector of k probabilities $\bar{p}$, and an arbitrary large number of factors k. An input to the $(f, \bar{p}, k)$-biclique cover problem is a bipartite graph $G = (L, R, E)$ generated by the following process: Factors are assigned arbitrarily to nodes in L, subject to a limit of f factors per node. Nodes in R are assigned factors by a random process in which each node is assigned factor i with probability p_i, independent of any other choice. A node $u \in L$ and a node $v \in R$ are connected if they share a factor. Given the observed graph $G = (L, R, E)$, and the parameters f, $\bar{p}$ and k, we want to find a k-biclique cover of the graph.

Results 2. *We show that for an arbitrary large constant f, and under some simple conditions on the distribution $\bar{p}$, there is a $O((|L| + |R|)^{O(1)})$ time algorithm that w.h.p. obtains a k-biclique cover of the graph, as long as $|R| \geq c \log(|L|)$ for some large enough constant $c > 0$. Here, the $O(1)$ depends only on the constant f.*

For an arbitrary probability vector $\bar{p}$, we present an algorithm that in $O((|L| + |R|)^{O(1)})$ time with high probability finds a set of bicliques of size k covering at least a $1 - \frac{1}{e}$ fraction of the edges, or $O(k \log(|R|))$ bicliques covering all edges of G, as long as $|R| \geq c \log(|L|)$ for some large enough constant $c > 0$. Here, the $O(1)$ depends on the distribution $\bar{p}$ and f but not on the size of the graph.

3. Model I: (f, k)-Biclique Cover

In this section, we study the (f, k)-biclique problem introduced. For the rest of the paper we use $N(u)$ to indicate the neighborhood of any node $u \in L \cup R$. Also, we represent (bipartite) subgraphs of G with capital letters, and we use $A = (L_A, R_A)$ to indicate that subgraph A has nodes L_A (resp. R_A) in the left (resp. right) side of the graph.

3.1. $(1, k)$-Biclique Cover

First, we observe that the $(1, k)$-biclique problem can be trivially solved in polynomial time.

Lemma 1. *A bipartite graph $G = (L, R, E)$ can be covered with a $(1, k)$-biclique cover iff it has $\leq k$ connected components and each of them is a biclique.*

Proof. Suppose each connected component of G is a biclique. It is easy to see that then all edges can be covered by using each connected component as the biclique in the cover. Also, each node is covered by a single biclique.

Then, assume G is a yes-instance. Let $B_1, \ldots B_t$, be the set of bicliques covering all the edges of G, in the solution. Notice that the bicliques define a partition of the nodes in G as each node can belong to a single biclique. We show that there is no edge that connects two nodes in two distinct bicliques, and hence the graph consists of $t \leq k$ connected components, each a complete subgraph. Suppose to get a contradiction that $(u, v) \in E$ is an edge connecting the nodes in B_i, B_j for $i \neq j$. Since (u, v) does not belong to any of the two bicliques, the edge must be covered by a third biclique. However, all edges of u are covered by a single biclique. □

This condition can be easily checked in polynomial time.

3.2. Results for the $(2, k)$-Biclique Cover

In this section, we assume that we are given in input a bipartite graph $G = (L, R, E)$ and a parameter k such that G is a yes-instance of the $(2, k)$-biclique cover problem. Our goal it to compute approximate solution to the following two optimization problems: covering all edges of G with as few bicliques as possible and covering as many edges of G as possible with k bicliques. We denote the union of the sets of node by $V = L \cup R$.

3.2.1. Maximize Edge Coverage with k Bicliques

We prove the following result:

Theorem 1. *There exists a* $O\left(\min\left(|V|^3 + k^{15}|E|, k|V|^7|E|\right)\right)$ *time algorithm that given in input a bipartite graph* $G = (L, R, E)$ *and a parameter* k*, where* G *is a yes-instance of the* $(2, k)$*-biclique cover problem, it outputs a set of at most* k *bicliques in* G *covering at least* $(1 - \frac{1}{e})|E|$ *edges of* G*.*

We first provide an overview of the main ideas on which Algorithm 1, which maximizes the number of edges covered with k bicliques, is based. The algorithm has four steps. First, we use a kernelization method to obtain a smaller graph G' which has a $(2, k)$-biclique cover iff G has a $(2, k)$-biclique cover. Then, we exploit the properties of the $(2, k)$-biclique covers for the kernelized graph G' to identify a set $\mathcal{B}'$ of all large enough bicliques that are part of any cover of G'. Then, we will show how to enumerate the set $\mathcal{C}'$ containing all possible small bicliques of any cover of G' not in $\mathcal{B}'$. We will keep a mapping from bicliques in the smaller graph G' to bicliques in the original graph G. Then, we cast the problem of selecting $\leq k$ bicliques as a monotone submodular maximization problem hence showing that we can cover at least $(1 - \frac{1}{e})|E|$ edges of the original graph G with k bicliques.

We now prove formally Theorem 1, showing an algorithm that outputs a set of at most k bicliques in G covering at least $(1 - \frac{1}{e})|E|$ edges of G. We describe each step of Algorithm 1 in details.

Step 1: Kernelization.

We apply the kernelization technique introduced by Fleischner et al. [2] for showing the parametrized complexity of biclique cover, which we now recall.

Given an instance (G, k) of biclique cover, the kernelization obtains an instance (G', k) where G' is constructed by applying the following two rules to G: (1) remove all nodes from G with no neighbors; (2) for each set of vertices that have the same neighborhood in G, keep only one of them. We call G' the kernelized graph of G if G' is constructed from G by applying the two rules. Fleischner et al. [2] show that G admits a k-biclique cover iff G' admits a k-biclique cover. We now show that using the same kernelization, G admits a (f_L, f_R, k)-biclique cover iff G' admits a (f_L, f_R, k)-biclique cover.

Lemma 2. $G = (L, R, E)$ *is a yes-instance of the* (f_L, f_R, k)*-biclique cover problem iff the kernelized graph* $G' = (L', R', E')$ *of* G *is a yes-instance of the* (f_L, f_R, k)*-biclique cover problem.*

Proof. The lemma is a corollary of the results for general kernelization of biclique cover [2,7].

Notice that nodes with no neighbors can be ignored without complications. We show that one application of the rule that removes one node u if another node v has the same neighborhood transforms yes-instances into yes-instances and no-instances to no-instance. The result will follow then by induction (over the number of application of the rule). Let G' be the graph obtained by removing $v \in L$ from G when there is a $v \neq u \in L$ s.t. $N(u) = N(v)$ (the case with $v \in R$ is similar and omitted).

Consider a (f_L, f_R, k)-biclique cover of G. Removing a node v from G and from all the bicliques involving v gives a graph that can be covered with $\leq k$ bicliques each covering at most f_L nodes on the left side and f_R on the right side.

Consider a (f_L, f_R, k)-biclique cover of G'. As node u has the same neighbors of v, we can extend each of the (at most f_L bicliques) covering u to include v on the left side. This preserves the fact that they are bicliques. Also v is covered by at most f_L bicliques, and in total we use k bicliques. The nodes on the right side are still covered by the same number of bicliques. □

The algorithm first applies the kernelization to obtain the bipartite graph $G' = (L', R', E')$. This can be done in time $O(|V|^3)$. We define $V' = L' \cup R'$. We will store a mapping $\texttt{kern} : V' \to 2^V$

that records for each node in $v \in V'$ the set $\texttt{kern}(v)$ of nodes in V that had the same neighborhood and of which v is the only node left in G' – i.e., $\texttt{kern}(v)$ is an equivalence class of the nodes with the same neighborhood. Notice that for each node $v \in V'$, $\texttt{kern}(v)$ contains at least one node and that all nodes in G with non-zero degree are in a unique $\texttt{kern}(v)$ set.

Algorithm 1 ApproxMaxCover($G = (L, R, E), k$)

```
Assumes $G$ is a yes-instance for $(2,k)$-biclique cover.
 — Step 1 —
Let $G' = (L', R', E')$ be the graph obtained applying the Fleischner et al. [2] kernelization.
Store mapping kern.
Let $V' = L' \cup R'$
if $|V'| \geq 2k^2$ then
    report no $(2,k)$-biclique cover exists for $G$ and quit.
end if
 — Step 2 —
Let $\mathcal{B}' \leftarrow \emptyset$.
For all $A \subseteq V'$ such that $G[A] = K_{3,4}$ or $G[A] = K_{4,3}$.
  • Let $A_L = A \cap L'$ and $A_R = A \cap R'$
  • Let $A_L \leftarrow \bigcap_{v \in A_R} N(v)$
  • Let $A_R \leftarrow \bigcap_{v \in A_L} N(v)$
  • Let $A' = (A_L, A_R)$ and add biclique $A'$ to the set $\mathcal{B}'$
if $|\mathcal{B}'| \geq k$ then
    report no $(2,k)$-biclique cover exists for $G$ and quit.
end if
 — Step 3 —
Let $\mathcal{C}'$ be the set containing all bicliques of $G'$ of the form:
  • All the bicliques $(\{u\},N(u))$ for $u \in L'$ or $(N(u), \{u\})$ for $u \in R'$
  • All the bicliques $(\{u,v\}, N(u) \cap N(v))$ for $u,v \in L'$ or $(N(u) \cap N(v), \{u,v\})$ for $u,v \in R'$
  • And all the $K_{3,3}$ in $G'$.
 — Step 4 —
Let $\mathcal{B} = \{\texttt{kern}(B') | B' \in \mathcal{B}'\}$.
Let $\mathcal{C} = \{\texttt{kern}(C') | C' \in \mathcal{C}'\}$.
Let $\mathcal{O} = \mathcal{B}$.
while $|\mathcal{O}| < k$ do
    Select $C \in \mathcal{C}$ s.t. the number of edges covered in $G$ by the union of the bicliques in $\mathcal{O} \cup \{C\}$ is maximized.
    Add C to $\mathcal{O}$.
end while
return $\mathcal{O}$.
```

First, notice that given a biclique $B' = (B'_L, B'_R)$ of G', we can construct a biclique $B = (B_L, B_R)$ of G using the mapping kern: $B_L = \bigcup_{v \in B'_L} \texttt{kern}(v)$ and $B_R = \bigcup_{v \in B'_R} \texttt{kern}(v)$. It is possible to observe that (B_L, B_R) is a biclique in G if (B'_L, B'_R) is a biclique in G'. Given a biclique B' in G', we write $\texttt{kern}(B')$ for the biclique B in G' obtained in this way.

The following properties of the kernelized graph G' of G will be exploited in the rest of the paper.

Lemma 3. *Let $G = (L, R, E)$ be a yes-instance of the $(2, k)$-biclique cover problem and let $G' = (L', R', E')$ be the kernelized version of the graph G. Then the following statements are true:*

- *G' admits a $(2, k)$-biclique cover.*
- *$|L' \cup R'| \leq O(k^2)$.*

- *In any $(2,k)$-biclique cover of G' there are no two nodes on the same side of the graph covered by the exact same set of bicliques.*

Proof. The first statement is a corollary of Lemma 2. Notice that all nodes on the same side of G', by construction, have distinct neighborhoods. Hence, they cannot be covered by the same set of 2 bicliques in any $(2,k)$-biclique cover. Finally, since each node in one side is covered by distinct subsets of at most 2 bicliques out of k, there are at most $2k^2 = O(k^2)$ nodes in G'. □

Step 2: Identifying Large Bicliques in the Cover.

We call $K_{l,r}$ biclique any biclique that has exactly l nodes in the left side of the bipartite graph and exactly r nodes in the right side. For a graph G and a set of vertices A, we call $G[A]$ the induced subgraph containing all nodes A and the edges between them.

We will also use the following notation. Given a $(2,k)$-biclique cover $\mathcal{B}$ of the graph G, where $\mathcal{B}$ is a set of bicliques covering all the edges of G, we define $F_{\mathcal{B}}(u)$ as the set of bicliques covering the node u in the cover $\mathcal{B}$. We have $1 \leq |F_{\mathcal{B}}(u)| \leq 2$. We will omit $\mathcal{B}$ from $F(\cdot)$ when it is clear which cover we are discussing.

We now make a series of important observations for the structure of the $(2,k)$-biclique cover of the kernelized graph G' of a $(2,k)$-biclique cover yes-instance G.

Lemma 4. *Let G be a yes-instance for the $(2,k)$-biclique cover problem and let G' be the kernelized graph of G. Suppose $A \subseteq L'$ (resp. $A \subseteq R'$) is such that $|A| \geq 3$. Let $\mathcal{B}$ be a $(2,k)$-biclique cover of G'. If there is a single biclique $B \in \mathcal{B}$ that covers all the nodes in A, then, in the cover $\mathcal{B}$, all nodes $S \subseteq R'$ ($S \subseteq L'$) connected to all nodes in A are covered by the same biclique B.*

Proof. Fix a solution $\mathcal{B}$ of the $(2,k)$-biclique cover of G'. Let B be the biclique that covers all nodes in A in the solution $\mathcal{B}$. We prove the case for $A \subseteq L'$, the other case is similar.

First, fix three distinct nodes a_1, a_2, a_3 in A. All the nodes are covered by biclique B. Notice that by the third item of Lemma 3, each node in A is covered by a different set of (at most 2) bicliques in $\mathcal{B}$. Hence, there is no other single biclique B' that covers two nodes in A (otherwise there will be two nodes both covered by the pair B, B' of bicliques).

Now consider a node $v \in R'$ connected to all nodes in A. If v is not part of biclique B, in the biclique cover $\mathcal{B}$, all the edges $(a_i, v) \in E$ for $i \in [3]$ are covered by a distinct biclique. This is a contradiction, as v is covered by at most 2 bicliques in $\mathcal{B}$. □

We use the previous lemma to prove the following useful result.

Lemma 5. *Let $G' = (L', R', E')$ be the kernelized graph of a $(2,k)$-biclique cover yes-instance G. Suppose that $A \subseteq L' \cup R'$ is such that $G'[A]$ is a $K_{3,4}$ biclique (or a $K_{4,3}$ biclique). Let $A_L = A \cap L'$ and $A_R = A \cap R'$. Let $A'_L = \bigcap_{v \in A_R} N(v)$ and let $A'_R = \bigcap_{v \in A'_L} N(v)$. Then the biclique $A' = (A'_L, A'_R)$ is part of* any *$(2,k)$-biclique cover of G' and it is the unique biclique that covers all nodes in A.*

Proof. Suppose to get a contradiction that in a given $(2,k)$-biclique cover $\mathcal{B}$ of G' there is no such unique biclique A' covering all nodes in A. From now on in this proof, when we use $F(u)$, we refer to $F_{\mathcal{B}}(u)$, i.e., the set of bicliques covering u in the cover $\mathcal{B}$.

W.l.o.g., we assume that $|A_L| = 3$ and $|A_R| = 4$. The proof follows similarly in the opposite case.

Notice that if any set of at least 3 nodes in either side of A is covered by the same biclique $B \in \mathcal{B}$ then, B covers all nodes in A. This is because, by Lemma 4, if three nodes in side L' in A (resp. R') are covered by B then all nodes in side R' in A (resp. L') are covered by B, and this implies that the all nodes in side L' (resp. R') are covered by B.

Consider any node $u \in A$. If $|F(u)| = 1$, than there is a contradiction. In fact, all neighbors of u are covered by a single biclique and hence we get that more than 3 on one side are covered by the same biclique. By the previous argument than there is a biclique covering all nodes in A. So we assume for the rest of the proof that $|F(u)| = 2$ for all nodes $u \in A$. Fix a node u in $A \cap L'$ and let B_1, B_2 be the two distinct bicliques covering the node u in $\mathcal{B}$. All nodes in $A \cap R'$ are covered by either B_1 or B_2.

We now show that there is no node $v \in A \cap R'$ such that $F(v) = \{B_1, B_2\}$. In fact, if such a node exists, then there are no assignments of B_1 and B_2 to nodes in $A \cap R'$ such that all nodes in $A \cap R'$ belong to one of the two bicliques but no bicliques covers 3 nodes in $A \cap R'$ (recall that if a biclique covers 3 nodes in $A \cap R'$ we get a contradiction because then all nodes in $A \cap L'$ are covered by the same biclique and hence the fourth node in $A \cap R'$ is covered as well by that biclique). Also similarly, this implies that B_1 and B_2 both cover exactly two distinct nodes $A \cap R'$. Let $r_1, \ldots r_4 \in A \cap R'$. W.l.o.g., we assume that $F(r_1) = \{B_1, C_1\}$ and $F(r_2) = \{B_1, C_2\}$, $F(r_3) = \{B_2, D_1\}$ and $F(r_4) = \{B_2, D_2\}$ where by kernelization we have $C_1 \neq C_2$ and $D_1 \neq D_2$.

Now, we have that all nodes in $A \cap L' \setminus \{u\}$ do not belong to both B_1 and B_2. For sake of contradiction, let, w.l.o.g., $v \in A \cap L \setminus \{u\}$ be a node that belongs to B_1. We know that v does not belong to B_2 (by kernelization). So $\{B_1, D_1, D_2\} \subseteq F(v)$, as it is connected to r_3, r_4 and hence $F(v) > 2$, giving a contradiction.

This implies, for any node $v \in A \cap L' \setminus \{u\}$ to be connected to all nodes in $A \cap R'$, that the following conditions are met: $\{C_1, C_2\} = \{D_1, D_2\}$ and $F(v) = \{C_1, C_2\}$. By kernelization, this gives a contradiction, as two distinct nodes in $A \cap L' \setminus \{u\}$ belong to the same set of bicliques.

We have shown that in any $(2,k)$-biclique cover $\mathcal{B}$ of G' there is at least one biclique A' covering all nodes in A. Now notice that there can be only one biclique covering all nodes in A, as otherwise there will be two nodes with the same set of bicliques in the cover on the same side of the graph. Hence, A' is the unique biclique covering all nodes in A in any $(2,k)$-biclique cover. Finally, by applying Lemma 4 to all nodes connected to each node in $A \cap L'$ (resp. $A \cap R'$), we know that the unique biclique A' is (A'_L, A'_R), as in the statement of the lemma. □

The previous lemma means that if A is a $K_{3,4}$ (or $K_{4,3}$) of G' then the biclique A' obtained as in the statement is part of any optimal solution. Also notice that clearly if in a given $(2,k)$-biclique cover of G' there is a biclique of size at least 3 in the smallest side, and at least 4 in the largest side, this biclique induces a $K_{3,4}$ (or $K_{4,3}$) in G'.

Hence, by enumerating all $K_{3,4}$ and $K_{4,3}$ in G' in polynomial time we can find a set $\mathcal{B}'$ of bicliques in G' such that: $\mathcal{B}'$ contains all large enough bicliques—i.e., with ≥ 3 nodes (resp. ≥ 4) in the smaller (resp. larger) side that are part of all $(2,k)$-covers of G'; and no biclique not in $\mathcal{B}'$ which is large enough (again it contains a $K_{3,4}$ or a $K_{4,3}$) is part of any $(2,k)$-biclique cover. This is done in Step 2 of Algorithm 1.

Step 3: Small Bicliques

We know that all the bicliques necessary to cover the remaining edges of G' (besides the one in $\mathcal{B}'$) need to involve at most 3 nodes in the smaller side. We can then enumerate a set $\mathcal{C}'$ of bicliques, in polynomial time, containing a super-graph of each biclique not in $\mathcal{B}'$ that is part of any $(2,k)$-biclique cover. It is possible to verify that at the end of Step 3 of the algorithm, $\mathcal{C}'$ contains at least one biclique that covers all edges of any biclique with at most 3 nodes in the smallest side of the biclique. Step 3 runs in polynomial time.

Step 4: Max Coverage

Finally, we have a set $\mathcal{B}'$, which contains bicliques that are necessary part of any optimal solution of G' and a set $\mathcal{C}'$ of bicliques G' that are sufficient to cover all edges of G' not covered by bicliques in $\mathcal{B}'$.

Let $\mathcal{B}$ be the set of bicliques of G obtained by $\mathcal{B} = \{\mathtt{kern}(B') | B' \in \mathcal{B}'\}$ (recall the definition of the mapping). Similarly we obtain the set $\mathcal{C}$ from $\mathcal{C}'$.

The algorithm will output all the bicliques in $\mathcal{B}$. For $t = |\mathcal{B}| \leq k$ the algorithm will obtain the remaining $k' = k - t$ bicliques in the following way. First, let $\mathcal{O} = \mathcal{B}$. The algorithm will add greedily a single biclique from $\mathcal{C}$ to $\mathcal{O}$ for k' times choosing each time the biclique that maximizes the number of newly covered edges in G'. It is easy to see that, since this function is monotone submodular and k' additional bicliques are sufficient to cover all edges not covered by $\mathcal{B}$, at least $(1 - \frac{1}{e})|E|$ of the edges of G are covered at the end of the process. We can now complete the proof of Theorem 1.

Proof of Theorem 1. Notice that by the previous considerations, the set $\mathcal{O}$ at the end of Algorithm 1 contains at most k bicliques, and these bicliques covers $(1 - \frac{1}{e})|E|$ edges of G.

The algorithm requires $O(|V|^3)$ time for the kernelization. Then, the remaining graph G' has at most $n = \min(2k^2, |V|)$ nodes and at most $m = \min(4k^4, k|E|)$ edges, by Lemma 3.

The most expensive operation is enumerating all the $K_{3,4}$, $K_{4,3}$ of the graph which can be done in $O(n^7)$ time. Hence, Step 2 takes $O(n^7 m)$ time.

Step 3 takes at most $O(n^6)$ time.

Finally, Step 4 does at most k iterations over a set of size $O(n^6)$ each step of the iteration taking at most $O(|E|)$ time.

The total is hence $O(|V|^3 + kn^7|E|)$, and the result follows. □

3.2.2. Covering All Edges with the Minimum Number of Bicliques

In this section we prove Theorem 2, the existence of an algorithm that outputs a set of at most $H(9)k$ bicliques covering all edges of the graph.

Theorem 2. *There exists a $O\left(\min\left(|V|^3 + k^{15}|E|, k|V|^7|E|\right)\right)$ time algorithm that given in input a bipartite graph $G = (L, R, E)$ and a parameter k, where G is a yes-instance of the $(2,k)$-biclique cover problem, outputs: a set of at most $H(9)k$ bicliques in G covering all edges of G.*

Here, $H(i) = \sum_{1 \le j \le i} j^{-1}$ is the i-th harmonic number and $H(9) \approx 2.83$.

For this problem, we use similar techniques described before. We outline the main differences from Algorithm 1. As before, we will construct the set $\mathcal{B}'$, which contains large bicliques that are part of any optimal solution. In this set, besides the bicliques including a $K_{3,4}$ (or $K_{4,3}$), we will identify as well certain bicliques with size $K_{2,a}$ for $a \ge 5$ and $K_{1,b}$ for $b \ge 10$ (and the cases $K_{a,2}$, $K_{b,1}$).

This will ensure that all remaining bicliques necessary for covering all edges in G' will have at most 9 edges in G' (i.e., they will be $K_{3,3}$, $K_{2,a}$, $K_{a,2}$ for $a \le 4$ and $K_{1,b}$, $K_{b,1}$ for $b \le 9$.)

Then, we cast the problem of selecting the minimum number of bicliques in $\mathcal{C}'$ that cover all remaining edges of G' (besides the one in $\mathcal{B}'$) as a set cover problem with sets of size at most 9. For this problem, it is known that the greedy algorithm gives a $H(9)$ approximation factor showing that all remaining edges in G' are covered with additional $\le H(9)(k - k')$ bicliques plus the $k' \le k$ in $\mathcal{B}'$ for a total of $\le H(9)k$ bicliques.

Finally, from the biclique cover of G', we can reconstruct the biclique cover in G with the same techniques of the previous section, outputting a cover with at most $H(9)k$ bicliques.

Properties of the Kernelized Graph of a $(2,k)$-Biclique Cover Yes-Instance

We show some additional properties of the kernelized graph of a $(2,k)$-biclique cover yes-instance that will be exploited by the algorithm. In this subsection, we always assume G' is the kernelized graph of $(2,k)$-biclique cover yes-instance G.

Lemma 6. *Let (L_A, R_A) be a $K_{2,5}$ biclique in G' such that L_A is not included in any $K_{3,4}$ biclique in G'. Then, in any $(2,k)$-biclique cover $\mathcal{B}$, there is always a unique biclique that covers all the nodes in L_A and no other node in the left side. (A similar result holds for $K_{5,2}$ bicliques.)*

Proof. We first show that in any $K_{2,5}$ biclique (L_A, R_A), all nodes in the small side L_A need to belong to the same biclique in any $(2,k)$-biclique cover .

Fix a $(2,k)$-biclique cover $\mathcal{B}$. $F(u)$ refers to the set of bicliques covering u in the cover $\mathcal{B}$.

Let $L_A = \{u, v\}$. Suppose to get a contradiction that $F(u) \cap F(v) = \emptyset$, then each node in R_A contains one biclique from $F(u)$ and one from $F(v)$. It is possible to verify that at most $2 \times 2 = 4$ distinct sets of size at most 2 can be formed by taking one element from $F(u)$ and one from $F(v)$. Hence, there are two nodes in R_A that are covered by the same set of bicliques, giving a contradiction as the graph is kernelized.

So we know that there is a biclique $A \in F(u) \cap F(v)$. Consider a node $x \in R_A$ such that $A \notin F(x)$. It is easy to see that there can be only one such node, as this node must be covered by the other biclique of $F(u)$ and the other biclique of $F(v)$ not equal to A. Again, for the kernelization, only one such node can exist.

So at least 4 nodes in R_A belong to a unique biclique. Finally, notice that if there is another node in the side of L_A in the same biclique, that would form a $K_{3,4}$ biclique. □

Lemma 7. *Let (L_A, R_A) be a $K_{1,10}$ biclique in G' such that R_A is not in any $K_{2,5}$ biclique including the node in L_A in the left part of the biclique. Then, in any optimal solution, there is a single biclique that covers the node in L_A and no other node on that side.*

Proof. Fix a $(2, k)$-biclique cover $\mathcal{B}$. $F(u)$ refers to the set of bicliques covering u in the cover $\mathcal{B}$.

Let $u \in L_A$. If $F(u) = A$, then all node in R_A belong to A and we can apply Lemma 4. Let $F(u) = \{A, B\}$, we have that all nodes in R_A belong either to A or to B or both. W.l.o.g., let A be the more common of the two. We have that A appears at least 5 times. So, if any other node in u side belongs to A, that would form a $K_{2,5}$ biclique. □

Algorithm

We now state how the algorithm for this problem constructs the set $\mathcal{B}'$. First, we add to $\mathcal{B}'$ all bicliques including a $K_{3,4}$, as done in Step 2 of Algorithm 1. Then, we add to $\mathcal{B}'$ all bicliques of the form $(\{u, v\}, N(u) \cap N(v))$ (and similar form $(N(u) \cap N(v)), \{u, v\})$ for $u \neq v \in L'$ such that u, v are part of a $K_{2,5}$ and not of a $K_{3,4}$. Finally, we add bicliques $(\{u\}, N(u))$ (and the form $(N(u), \{u\})$) such that $u \in L'$ has at least 10 edges not covered by the bicliques added before.

We claim the following property for $\mathcal{B}'$.

Lemma 8. *There is a $(2, k)$-biclique cover of G' which contains all bicliques in $\mathcal{B}'$ and only other bicliques with at most 9 edges.*

Proof. The lemma follows by combining Lemmas 5 and 6. By Lemma 5, we know that there is no other biclique in any $(2, k)$-biclique cover with at least 3 nodes on the smaller side and at least 4 on the largest side.

Suppose there is a $(2, k)$-biclique cover of G' containing all bicliques of $\mathcal{B}'$ and another biclique $B = (L_B, R_B) \notin \mathcal{B}'$ of the form $K_{2,a}$ for $a \geq 5$. If nodes in L_B are part of a $K_{3,4}$ biclique, we get a contradiction, as we have two nodes in L_B that are both covered by the biclique B and the other biclique in $\mathcal{B}'$ identified by the algorithm (by kernelization this is not possible). So we know that L_B is not part of a $K_{3,4}$ biclique. However, in this case, by construction of the algorithm, there is another biclique $\mathcal{B}'$ covering L_B, so again we get a contradiction. (The same proof works for the $K_{a,2}$ case for $a \geq 5$).

Finally notice that, by construction, at most 9 edges of each node are left uncovered by bicliques in $\mathcal{B}'$ so there is a $(2, k)$-biclique cover that includes all bicliques in $\mathcal{B}'$ and other bicliques with at most 9 edges each. □

Finally, notice that we can enumerate the set $\mathcal{C}'$ of all bicliques with at most 9 edges necessary to cover the edges left uncovered by the bicliques in $\mathcal{B}'$ in time $O(n^6)$, where n is the number of nodes in the graph G'.

The proof of Theorem 2 follows by the properties of set cover.

3.3. NP-Complete Result for $f \geq 5$

Theorem 3. *(f, k)-biclique cover is NP-Complete for any $f \geq 5$.*

3.4. Proof of Theorem 3

Recall that the (Δ, k)-clique vertex partition problem is defined as follows. Given a graph $G = (V, E)$, with maximum degree $\leq \Delta$ and an integer k, determine whether there is a partition of nodes in V in $t \leq k$ subsets $C_1, \ldots, C_t$, such that each subset C_i is a clique.

Král et al. [19] showed a reduction to clique vertex partition from the NP-Complete Clause-Linked Planar 3-SAT problem [20]. The latter is the following decision problem: given a CNF formula, where each clause contains either 2 or 3 literals and each literal appears in exactly 3 clauses (twice negated and once positive), determine whether the formula is satisfiable.

Král et al. [19] showed that given such formula ϕ we can construct a graph $G_\phi = (V_\phi, E_\phi)$ such that nodes in V_ϕ can be partitioned in $|X| + 3|C|$ cliques iff ϕ is satisfiable (where X and C are the sets of variables and clauses of ϕ, respectively). The graph G_ϕ is constructed as follows. For each variable $x \in X$, create the subgraph $G_x = (\{x, x^+, x_1^-, x_2^-\}, \{xx^+, xx_1^-, xx_2^-, x_1^- x_2^-\})$. For each clause $c \in C$, create a 7-node cycle G_c. If the clause c has 3 literals x,y,z, fix three non pairwise adjacent nodes: $c(x)$, $c(y)$, $c(z)$ in the cycle (e.g., let $c(x)$ be at distance two from $c(y)$, and let $c(y)$ be at distance two from $c(z)$ and let $c(z)$ be at distance three from $c(x)$). If the clause has two variables, ignore $c(z)$. Now, for each literal x if c is the clause in which it appears positive, let the $c(x)$ node in G_c be coincident with the node x^+ in G_x. For the clauses in which x appears negated, we similarly use nodes x_1^- and x_2^-.

The following result is a corollary of those of Král et al. [19].

Lemma 9. *(Δ, k)-clique vertex partition is NP-Complete for $\Delta \geq 4$.*

Proof. By Král et al. [19], we know that ϕ is satisfiable iff the nodes in V_ϕ can be partitioned in $|X| + 3|C|$ cliques. Notice that graph G_ϕ has maximum degree $\Delta \leq 4$. Hence, $(4, k)$-clique vertex partition is NP-Complete as well. □

Now, using the reduction of Orlin [5], we show that (Δ, k)-clique vertex partition can be reduced to $(\Delta + 1, \Delta + 1, k')$-biclique cover for some $k' \geq k$.

Given a graph $G = (V, E)$, construct a bipartite graph $B_G = (L, R, E_B)$ as follows. For each node $v_i \in V$, create two nodes $x_i \in L, y_i \in R$. Moreover, for each edge $v_i v_j \in E$, make two nodes $x_{ij} \in L$ and $y_{ij} \in R$. For each $v_i \in V$. we have $x_i y_i \in E_B$. Also for each edge $v_i v_j \in E$. we have the following edges in E_B $x_i y_j, x_{ij} y_j, x_i y_{ij}, x_{ij} y_{ij}$. Orlins [5] showed that a graph G admits a clique vertex partition in k cliques iff the edges of B_G can be cover by at most $k + |E|$ bicliques. The following lemma can also be shown.

Lemma 10. *Let G be a graph with maximum degree at most Δ, and let B_G the corresponding bipartite graph defined before, there exists a $(\Delta + 1, k + |E|)$-biclique cover of B_G iff there is k-clique vertex partition of G.*

Proof. Suppose there is no k-clique vertex partition of G. Then, by Orlins [5], we know that there is no biclique cover of the edges of B_G with $k + |E|$ bicliques (irrespectively of how many bicliques a node belongs). Suppose there is a k-clique vertex partition of nodes in G. Let C be any set in the partition. Then, it is possible to see that $B_C = (\{x_i | v_i \in C\}, \{y_i | v_i \in C\})$ is a biclique in B_G and that k such bicliques covers all edges of the form $x_i y_i$. Also, as the sets in the partition are disjoint, each node in $V(B_G)$ belongs to at most one such biclique.

Now, consider the edges of the form $x_i y_j \in E_B$, for $v_i v_j \in E$. Any such edge can be covered with the biclique $B_{ij} = (\{x_{ij}, x_i\}, \{y_j, y_{ij}\})$. Notice that by using $|E|$ such bicliques we can complete the cover of all the remaining edges in E_B (including the edges of the form $x_{ij} y_{ij}$). Now, each node of the form x_{ij} (or y_{ij}) belongs to only one biclique in this cover. Nodes x_i belong to one biclique of the form

B_C as well as other $deg(v_i)$ bicliques of the form B_{ij} such that $v_iv_j \in E$. So we have that each node is part of at most $\Delta + 1$ bicliques because of the degree of G. □

Theorem 3 is a corollary of the above lemma.

4. Model II: One-Side Stochastic Model

We recall the definition of the model. In the $(f, \bar{p}, k)$-biclique cover model, there is a constant f, a vector of k probabilities $\bar{p} = p_1, \ldots p_k$, and a number of factors k. The input is a graph $G = (L, R, E)$ where each node in L is assigned to $\leq f$ factors from $[k]$ arbitrary and each node in R is assigned independently to factor i with probability $0 < p_i < 1$. A node $u \in L$ and a node $v \in R$ are connected iff they share a common factor.

We show that, under certain conditions on the parameters f and $\bar{p}$, we can recover an optimal or close to optimal (f, k, k)-biclique cover of the graph.

We will make the following two assumptions throughout the section. First, let $L_i \subseteq L$ be the set of nodes assigned to factor i in L. We will always assume that for no pair $i \neq j \in [k]$, $L_i \subseteq L_j$. Notice that if $L_i \subseteq L_j$ for $i \neq j$ then the biclique of factor i is completely covered by the biclique of factor j and there is a equivalent $(f, \bar{p'}, k-1)$-biclique cover model without the factor i.

Let $p_m = \min_{i \in [k]}(p_i)$, $p_M = \max_{i \in [k]}(p_i)$. Second, we assume that the right side R is sufficiently large w.r.t. the left side. In particular we assume $|R| > C_{f,p_m,p_M} \log(|L|)$ for a sufficiently large constant C_{f,p_m,p_M} which depends on f, p_m, p_M only. This is to allow extracting enough information from the random assignment in the right part.

In the rest of the section, we present two results. First, we show that under an additional condition on the constants $\bar{p}$ that ensures that no factor has much higher probability than another factor and that no factor has too large a probability, then an exact $(f, \bar{p}, k)$-biclique cover can be obtained in polynomial time. We then show that for any vector of constants $\bar{p}$, it is possible to obtain an approximate cover.

4.1. Optimal $(f, \bar{p}, k)$-Biclique Cover

We first state the additional condition on $\bar{p}$:

$$\frac{p_m}{2} > 1 - (1 - p_M)^{2f-1}(1 + (2f-1)p_M) \tag{1}$$

Notice that for any f there is a non-empty set of vectors $\bar{p}$ for which the condition holds. For instance, it holds for any vector s.t. $0 < p_m = p_M < c_f$ where $c_f < 1$ is constant depending on f. We will prove the following theorem.

Theorem 4. *Let $G = (L, R, E)$ be the graph obtained by the $(f, \bar{p}, k)$-biclique cover model, such that $|R| \geq \Omega(\log(|L|))$ and where the condition in Equation (1) holds for $\bar{p}$. There exists a $O(k|L|^{f+1}|E|)$ time algorithm that given G, f, k and p_m outputs a (f, k, k)-biclique cover of G w.h.p. In the paper, we say that an event happens with high probability if it happens with probability $\geq 1 - |V|^c$ for some constant $c > 0$.*

Main Ideas

Before entering into the details of the proof here we present the main ideas of the proof. First, we show that, under the condition in Equation (1) and for $|R| \geq \Omega(\log(|L|))$, it is possible to distinguish in polynomial time w.h.p., whether in any given set $S \subseteq L$ there exists at least a factor $i \in [k]$, such that all nodes in S belong to the L_i (are assigned to the factor). In particular, we show that this is possible by just looking at the number of the common neighbors of all the nodes in S.

We will then assume to have an oracle $\mathcal{I} : 2^L \to \{0, 1\}$ that returns, w.h.p. $\mathcal{I}(S) = 1$, iff there is factor to which all nodes in S belong.

We will show that there is an algorithm that makes a polynomial numbers of calls to this oracle and finds an (f,k,k)-biclique cover of G. We stress that this algorithm could be used also for other probabilistic models where the function $\mathcal{I}$ can be *computed* with sufficient accuracy.

The intuition of the algorithm is that we can construct each biclique in the cover, one at a time, once we identify a set of nodes $S \subseteq L$ such that there exists a *unique* factor $i \in [k]$ to which all nodes in S belong. Furthermore, we show that for any factor there is at least one set of size $\leq f$ of nodes in L for which this happens. We can then identify each biclique in polynomial time, w.h.p.

We now proceed to prove formally the results the this model.

4.1.1. Determining Whether Nodes in $S \subseteq L$ Share a Common factor

Recall that in this section we assume we are given a bipartite graph $G = (L, R, E)$ obtained by the $(f, \bar{p}, k)$-biclique cover where the condition Equation (1) holds, $|R| \geq \Omega(\log(|L|))$ and no L_i set is subset of L_j for two distinct factors $i, j \in [k]$.

First, we show that, given a set $S \subseteq L$, it is possible to distinguish based on the graph G whether there is at least a common factor assigned to all nodes in S. In this section, we use $F(u)$ to indicate the set of bicliques covering node u in the model (i.e., each biclique corresponds to the nodes with one factor).

Lemma 11. *Given a set $S \subseteq L$, if $\bigcap_{u \in S} F(u) = \emptyset$ then the probability that*

$$|\bigcap_{u \in S} N(u)| > (1+\epsilon)2(1-(1-p_M)^{2f} - 2fp_M(1-p_M)^{2f-1}))|R|$$

for $0 < \epsilon < 1$ is smaller than $\exp\left(-1/3\epsilon^2 C|R|\right)$, for some constant C not dependent on $|R|$.

Proof. Let X_w be the indicator variable of the event that $\forall u \in S\ (u,w) \in E$.

We claim that the probability of X_w is upper-bounded by twice the probability of the following event: w is part of at least two bicliques selected from a set F' of cardinality at most $2f$. Suppose in fact that w is part of a single biclique B, then as $\bigcap_{u \in S} F(u) = \emptyset$, w cannot be connected to all nodes in S. w must be part of at least two bicliques. Now, fix a single node $u \in S$. $|F(u)| \leq f$, and w is part of at least one biclique in $F(u)$. If w is part of two bicliques in $F(u)$, then our bound holds (use $F' = F(u)$ as a the sets from which the bicliques must be selected). If w is part of a single biclique $B \in F(u)$, then fix $v \in S$, a node such that $B \notin F(v)$ (the node exists for the empty intersection in S). We have that w is part of another biclique from $F(v)$ with $|F(v)| \leq f$. In this case, w is part of at least two bicliques from the set $F' = F(v) \cup F(u)$ of size at most $2f$. By union bound on these two cases

$$\Pr\left(X_w\right) \leq 2(1-(1-p_M)^{2f} - 2fp_M(1-p_M)^{2f-1}))$$

where the inequality comes from the fact that the probability is maximized when the set F' contains exactly $2f$ bicliques of maximum probability p_M, the probability of selecting at least 2 bicliques is the opposite of the probability of not selecting 0 of 1 biclique.

Let X be the random variable that measures the cardinality of $|\bigcap_{u \in S} N(u)|$. Then, it is easy to see that

$$\mathbf{E}\left[X\right] \leq |R|2(1-(1-p_M)^{2f} - 2fp_M(1-p_M)^{2f-1}),$$

and by the Chernoff bound, for $0 < \epsilon < 1$, we have

$$\Pr\left(X > (1+\epsilon)\mathbf{E}\left[X\right]\right) < \exp\left(-1/3\epsilon^2 2(1-(1-p_M)^{2f} - 2fp_M(1-p_M)^{2f-1}|R|)\right).$$

□

Similarly, we can show

Lemma 12. *Let $S \subseteq L$ such that $\bigcap_{u \in S} F(u) \neq \emptyset$ then the probability that $|\bigcap_{u \in S} N(u)| < (1-\epsilon)p_m|R|$ for $0 < \epsilon < 1$ is smaller than* $\exp(-1/3\epsilon^2 p_m|R|)$.

Proof. Let X_w be the indicator variable of the event that $\forall u \in S$, we have $(u, w) \in E$. Let A be one of the common biclique of the nodes. We have

$$\Pr(X_w) \geq \Pr(A \in F(w)) = p_A \geq p_m,$$

the result follows by the Chernoff bound. □

Lemma 13. *Let $p_m > 2(1-(1-p_M)^{2f} - 2fp_M(1-p_M)^{2f-1})$, given a set $S \subseteq L$, then there is a constant $0 < \epsilon < 1$ and a constant $c > 0$ such that, for $|R| \in \Theta(\log(|L|))$ with probability $\geq 1 - \exp(-c|R|)$*

$$|\bigcap_{u \in S} F(u)| > 0$$

if and only if $|\bigcap_{u \in S} N(u)| \geq (1-\epsilon)p_m|R|$.

Proof. If $p_m > 2(1-(1-p_M)^{2f} - 2fp_M(1-p_M)^{2f-1}$ then there is an $\epsilon > 0$ for which $(1-\epsilon)p_m > (1+\epsilon)2(1-(1-p_M)^{2f} - 2fp_M(1-p_M)^{2f-1}$.

Notice that $(1/3)\epsilon^2 p_m > 1/3\epsilon^2 2(1-(1-p_M)^{2f} - 2fp_M(1-p_M)^{2f-1}$.

Then we can apply the previous lemma and show that that with probability at least

$$1 - \exp(-(1/3)\epsilon^2 p_m|R|)$$

we have $|\bigcap_{u \in S} N(u)| \geq (1-\epsilon)p_m|R|$ iff $|\bigcap_{u \in S} F(u)| > 0$. □

Given the previous results, we can implement a function $I(S)$ that given a set S determines correctly, in time $O(|E|)$, whether $|\bigcap_{u \in S} F(u)| > 0$ with probability at least

$$1 - \exp(-(1/3)\epsilon^2 p_m|R|).$$

4.1.2. Algorithm Based on the Oracle I

In this section, we will assume that there is a function $I(S)$ that returns $I(S) = 1$ iff $|\bigcap_{u \in S} F(u)| > 0$. We will assume that the function returns the correct answers all times. Given the previous results, since we make $O(|L|^c)$ calls to such function, this happens w.h.p. as long as $|R| \geq C_{f,p_m,p_M}\log(|L|)$ for such sufficiently large constant C_{f,p_m,p_M} depending on f, p_m, p_M.

The algorithm is presented in Algorithm 2.

Algorithm 2 OptCoverOneSideRandom($G = (L, R, E), f, p_m, k$)

```
Mark all edges in $E$ as uncovered.
Let $\mathcal{B} \leftarrow \emptyset$ be the set of bicliques in the cover
while $\exists u \in L$ adjacent to an uncovered edge do
    — Step 1 —
    $L_B \leftarrow \{u\}$.
    — Step 2 —
    while $\exists S \subseteq L \setminus L_B$ s.t. $1 \leq |S| \leq f, \mathcal{I}(S \cup L_B) = 1$ and there is no biclique in $\mathcal{B}$ that covers all nodes in $S \cup L_B$ do
        $L_B \leftarrow L_B \cup S$.
    end while
    — Step 3 —
    for $\forall v \in L$ do
        if $v \notin L_B$ and $\mathcal{I}(L_B \cup v) = 1$ then
            $L_B \leftarrow L_B \cup \{u\}$.
        end if
    end for
    — Step 4 —
    $R_B = \bigcap_{u \in L_B} N(u)$
    $B \leftarrow (L_B, R_B)$
    $\mathcal{B} \leftarrow \mathcal{B} \cup B$
    Mark all edges in $B$ as covered.
end while
return $\mathcal{B}$
```

Intuition of the Algorithm

The main idea of the algorithm is to isolate a single new biclique B at a time from the optimal cover $\mathcal{B}$ given by the model. We do so by finding any subset $L_B \subseteq L$ of left nodes such that: any node in S belong to biclique B; and there is no other biclique to which all nodes in S belongs. Then we will identify all the left nodes that belong to biclique B.

Correctness of the Algorithm

We first show the following lemma.

Lemma 14. *For any biclique $B = (L_B, R_B)$ in the cover $\mathcal{B}$ given by the model, there is a set of $S \subseteq L$ of size at most f such that $\bigcap_{u \in S} F(u) = B$.*

Proof. Consider a node $u \in L_B$. This node can belong to at most other $f - 1$ bicliques $B_1, \dots B_c \in \mathcal{B}$ with $c \leq f - 1$. Notice that for any biclique B_i there exists a node u_i s.t. $B \in F(u_i)$ and $B_i \notin F(u_i)$ (by assumption in the model left nodes in biclique B are not a subset of biclique B_i). Hence the set $S = \{u, u_1, \dots u_c\}$ (where u_i might not be distinct) has the property that $\bigcap_{u \in S} F(u) = B$ and $|S| \leq f$. □

We now show that Algorithm 2 is correct.

Theorem 5. *The Algorithm 2 computes a biclique cover for the graph G with k bicliques.*

Proof. We show this by induction on the order of the bicliques determined by the algorithm. More precisely, we will show that for any $k' \geq 0$, there exists an optimal cover containing all the first k' bicliques determined by the algorithm. The base case for $k' = 0$ is obvious (no clique has been determined). Suppose the statement holds for the first $k' - 1$ bicliques. If the algorithm ends without adding another biclique, this concludes the proof. Otherwise, let B be the k'-th biclique determined. Notice that for the algorithm to not stop, this means that a node $u \in L$ had at least an edge uncovered. So we know that u must belong to another biclique in an optimal cover. We will now show that until the end of Step 3, the set L_B preserves the property that there is a biclique B in the optimal cover (not already found) containing all nodes in L_B. This is clearly true at the beginning (Step 1). In Step 2, we check this property explicitly. Notice that in Step 3 we add nodes in L_B only if they maintain the property $\mathcal{I}(L_B \cup v) = 1$, so the property continues to hold.

We now show that at the end of Step 3 there is a *unique* biclique not in the cover determined so far that covers all nodes in L_B in the optimal solution. Suppose there exists a set of 2 or more distinct new bicliques that covers all the nodes in the set L_B obtained after the end of Step 3. Notice that by Lemma 14 there is set $S \subseteq L$ of size $|S| \leq f$ such that only one of the two bicliques covers all nodes in S. S respects the condition of the while, so we would not have exited the loop of Step 2 giving a contradiction. Hence, at the end of Step 2, there is a unique new biclique B, such that all nodes in L_B belong to the biclique.

Hence, it follows that all left nodes of biclique B are found in Step 3, (they intersect because of biclique B, and the loop greedily adds all nodes that intersect).

Finally, during Step 4, it is clear that we construct a biclique B that covers all the edges covered by the biclique in the optimal solution with left side given by L_B. So there is an optimal solution that contains the biclique B computed. □

Time Complexity

Notice that the outer loop can be executed only $O(k)$ times (once for each biclique found). In Step 2, the loop is executed at most $O(n)$ times and requires at most $O(|L|^f)$ calls to $\mathcal{I}$ each of cost $O(|E|)$ for a total of $O(|L|^{f+1}|E|)$. Step 3 requires $O(|L|)$ calls to $\mathcal{I}$ for a total of $O(L|E|)$. Finally, Step 4 costs $O(|E|)$. The total cost is $O(k|L|^{f+1}|E|)$.

The correctness of Theorem 4 follows from the previous results.

4.2. Approximate Biclique Cover

In this section, we show that for any constants in $\bar{p}$ subject only to $0 < p_i < 1$ for $i \in [k]$ we can construct an approximate solution to the problem.

Theorem 6. *Let $G = (L, R, E)$ be the graph obtained by the $(f, \bar{p}, k)$-biclique cover model, such that $|R| \geq \Omega(\log(|L|))$. There exists a $O((|L| + |R|)^{\alpha_{f,p_m,p_M}})$ time algorithm where α_{f,p_m,p_M} is a constant depending only on f, p_m, p_M but not on the size of the graph, such that, given G, f, k, and p_m the algorithm outputs 1) a set of $k' = O(\log(|L| + |R|)k)$ bicliques covering all edges of G; and 2) a set of k bicliques covering $(1 - \frac{1}{e})|E|$ edges of G.*

Before entering into the details of the proof, we present the main idea of the algorithm, which is the following. There exists a constant c_{f,p_m,p_M} depending on f, p_m, p_M such that if we sample u.a.r. a set $S \subseteq R$ of size $|S| = c_{f,p_m,p_M} \log(|L|)$ we have that for each factor $i \in [k]$ there is a set $S_i \subseteq S$ such that $|S_i| \geq c' \log(|L|)$ for some constant $c' < c_{f,p_m,p_M}$ and all nodes in S_i belong to the factor i. Moreover, it is possible to show that the set of common neighbors of S_i in L is, w.h.p., exactly the set of nodes L_i. Given such set S_i, we can reconstruct the biclique B_i that covers all nodes with factor i. Now, we can list all subsets $S' \subseteq S$ of size $\leq c' \log(|L|)$ in polynomial time and construct a single candidate biclique B' from each subset S'. We have that k of these bicliques cover all edges of G, so we can cast the problem as a maximum coverage problem and find a set of k bicliques covering $(1 - \frac{1}{e})|E|$ edges or as a set cover problem and find a set of $\log(|E|)k$ bicliques covering all edges.

Algorithm

We now outline the algorithm (Algorithm 3). Suppose we are given a bipartite graph $G = (L, R, E)$ obtained by the $(f, \bar{p}, k)$-biclique cover model.

Algorithm 3 ApproxCoverOneSideRandom($G = (L, R, E), f, p_m, p_M, k$)

```
Let
    $\bar{c} = \max\left(16p_m^{-1}, 2p_m^{-1}\log_{(1-(1-p_M)^f)^{-1}}(3)\right)$
Sample a set $S \subseteq R$ of $\bar{c}\log(|L|)$ nodes u.a.r. from $R$ (without replacement).
Let $\mathcal{C} \leftarrow \emptyset$
for Each subset $S' \subseteq S$ do
    Let $L' \leftarrow \bigcap_{v \in S'} N(v)$.
    Let $S'' \leftarrow \bigcap_{v \in L'} N(v)$.
    Let $B$ be the set of the edges covered by biclique $(L', S'')$.
    $\mathcal{C} \leftarrow \mathcal{C} \cup B$.
end for
Apply the greedy max coverage (or set cover) algorithm using $\mathcal{C}$ as the input sets and $E$ as the universe set to cover.
```

Let $p_m = \min_i p_i$ and $p_M = \max_i p_i$. Let $\bar{c}$ be a constant to be determined subsequently. The algorithm first samples a subset S of R by selecting $c\log(|L|)$ u.a.r. nodes in R without replacement. In the next lemma, we show that, for each biclique B_i (in the cover defining the model), we have a large number of nodes in S covered by that biclique.

Lemma 15. *Suppose for $\bar{c} \geq 16p_m^{-1}$ a set of S of $\bar{c}\log(|L|)$ nodes is drawn u.a.r. without replacement. No biclique $i \in [k]$, $B_i = (L_i, R_i)$ of the model is such that $|R_i \cap S| < \frac{1}{2}\bar{c}p_m\log(|L|)$ with probability at least $1 - \frac{1}{|L|}$.*

Proof. Consider a single biclique B_t, $t \in [k]$. Let X_i be an indicator random variable of the event that the i-th node drawn in S is covered by B in the cover of the model. Let $X = \sum_{i=1}^{\bar{c}\log(|L|)} X_i$. Notice that as nodes in R are assigned to biclique B_t independently with probability p_t, and the sampling without replacement maintains the independence, we have $\mathbf{E}[X] = \bar{c}\log(|L|)p_t \geq \bar{c}\log(|L|)p_m$ and by Chernoff bound

$$\Pr\left(X < \frac{1}{2}\bar{c}\log(|L|)p_m\right) \leq \exp(-\frac{1}{8}\bar{c}p_m\log(|L|)).$$

For $\bar{c} \geq 16p_m^{-1}$, we have

$$\Pr\left(X < \frac{1}{2}\bar{c}\log(|L|)p_m\right) \leq |L|^{-2},$$

and hence by union bound as $k \leq |L|$ with probability $\geq 1 - \frac{1}{|L|}$ no biclique appears in less than $\frac{1}{2}\bar{c}\log(|L|)p_m$ nodes in S. □

Then, we show that if a subset $S' \subseteq S \subseteq R$ of size $> c'\log(n)$, where c' is a constant, is such that all nodes of S' belong to biclique B_i then no node $v \in L$ is connected to all nodes in S', w.h.p, unless it belongs to the same biclique.

Lemma 16. *Suppose $c' \geq \log_{(1-(1-p_M)^f)^{-1}}(3)$. Fix a set $S' \subseteq S$ of size $c'\log(|L|)$ of nodes in R. Suppose that all nodes in S are such that they all belong to biclique $B_t = (L_t, R_t)$ in the cover defining the model, then*

$$\bigcap_{v \in S'} N(v) = L_t,$$

with probability $\geq 1 - \frac{1}{|L|^2}$.

Proof. Notice that $L_t \subseteq \bigcap_{v \in S'} N(v)$, as all nodes in L_t are connected to nodes in S' by definition. We prove that a given node $v \in L \setminus L_t$ is connected to all nodes in S' with probability $\leq \frac{1}{|L|^3}$.

Notice that all bicliques are assigned to nodes in R independently. So consider a node $u \in L$ such that $B_t \notin F(u)$.

$$\begin{aligned}
\Pr\left(u \in \bigcap_{v \in S'} N(v)\right) &= \prod_{v \in S'} \Pr\left(F(u) \cap F(v) \neq \emptyset\right) \\
&= \prod_{v \in S'} \left(1 - \Pr\left(F(u) \cap F(v) = \emptyset\right)\right) \\
&= \prod_{v \in S'} \left(1 - \prod_{B_t \in F(u)} (1 - p_t)\right) \\
&\leq \prod_{v \in S'} \left(1 - (1 - p_M)^f\right) \\
&\leq \left(1 - (1 - p_M)^f\right)^{c' \log(|L|)} \\
&\leq |L|^{-3}
\end{aligned} \tag{2}$$

By union bound on the n nodes in L, the result follows. □

Now, we fix a constant $\bar{c}$ such that $\frac{1}{2}\bar{c}p_m \geq \log_{(1-(1-p_M)^f)^{-1}}(3)$ and such that the requirement of Lemma 15 is satisfied.

It is easy to see that for all $i \in k$ the bicliques in the optimal cover there is a set $S'_i \subset S$ that by Lemma 16 w.h.p. is such that for $L'_i = \bigcap_{v \in S'_i} N(v)$, $S''_i = \bigcap_{v \in L'_i} N(v)$, the biclique (L'_i, S''_i) covers more edges than the i-th biclique in the cover defining the model, w.h.p.

Notice that we can obtain all subsets $S' \subseteq S$ in polynomial time $O(2^{\bar{c}\log(|L|)}) = O(|L|^{\bar{c}})$.

The correctness of Algorithm 3 follows from the fact that the set $\mathcal{C}$ contains w.h.p. each biclique from the cover defining the model, and the set $\mathcal{C}$ has size $O(|L| + |E|)^{\alpha_{f,p_m,p_M}})$ for some constant α_{f,p_m,p_M} depending on f, p_m, p_M but not on the size of the graph or the number of bicliques.

5. Conclusions

We have studied two natural restrictions to the well-known problem of biclique cover in bipartite graphs that are motivated by practical applications in computational biology and other fields. For these restrictions, we have shown some non-trivial optimal or approximation algorithms which run in polynomial time.

As a future work, it would be interesting to determine whether (f,k)-biclique cover is NP-Complete for $2 \leq f \leq 4$ and to determine inapproximability results for general values of f.

Author Contributions: A.E. and E.U. contributed equally to the paper.

Acknowledgments: We thank Marie-France Sagot for introducing us to this problem, and Blerina Sinaimeri and Mary Arnaud for many insightful discussions on the problem. The research was partially supported by NSF grant IIS-1247581 and NIH grant R01-CA180776.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Chalermsook, P.; Heydrich, S.; Holm, E.; Karrenbauer, A. Nearly Tight Approximability Results for Minimum Biclique Cover and Partition. In *Algorithms-ESA 2014*; Springer: Berlin, Germany, 2014; pp. 235–246.
2. Fleischner, H.; Mujuni, E.; Paulusma, D.; Szeider, S. Covering graphs with few complete bipartite subgraphs. *Theor. Comput. Sci.* **2009**, *410*, 2045–2053. [CrossRef]
3. Simon, H.U. On approximate solutions for combinatorial optimization problems. *SIAM J. Discret. Math.* **1990**, *3*, 294–310. [CrossRef]
4. Gruber, H.; Holzer, M. Inapproximability of nondeterministic state and transition complexity assuming P $\neq$ NP. In *Developments in Language Theory*; Springer: Berlin, Germany, 2007; pp. 205–216.
5. Orlin, J. Contentment in graph theory: Covering graphs with cliques. In *Indagationes Mathematicae (Proceedings)*; Elsevier: NewYork, NY, USA, 1977; Volume 80, pp. 406–424.
6. Jukna, S.; Kulikov, A.S. On covering graphs by complete bipartite subgraphs. *Discret. Math.* **2009**, *309*, 3399–3403. [CrossRef]

7. Nor, I.; Hermelin, D.; Charlat, S.; Engelstadter, J.; Reuter, M.; Duron, O.; Sagot, M.F. Mod/Resc parsimony inference: Theory and application. *Inf. Comput.* **2012**, *213*, 23–32. [CrossRef]
8. Nor, I.; Engelstädter, J.; Duron, O.; Reuter, M.; Sagot, M.F.; Charlat, S. On the genetic architecture of cytoplasmic incompatibility: Inference from phenotypic data. *Am. Nat.* **2013**, *182*, E15–E24. [CrossRef] [PubMed]
9. Nau, D.S.; Markowsky, G.; Woodbury, M.A.; Amos, D.B. A mathematical analysis of human leukocyte antigen serology. *Math. Biosci.* **1978**, *40*, 243–270. [CrossRef]
10. Miettinen, P.; Mielikainen, T.; Gionis, A.; Das, G.; Mannila, H. The discrete basis problem. *IEEE Trans. Knowl. Data Eng.* **2008**, *20*, 1348–1362. [CrossRef]
11. Mishra, N.; Ron, D.; Swaminathan, R. Learning Theory and Kernel Machines. In Proceedings of the 16th Annual Conference on Learning Theory and 7th Kernel Workshop COLT/Kernel 2003, Washington, DC, USA, 24–27 August 2003; pp. 448–462.
12. Hirsch, M.; Meijer, H.; Rappaport, D. Biclique edge cover graphs and confluent drawings. In *Graph Drawing*; Springer: Berlin, Germany, 2006; pp. 405–416.
13. Müller, H. On edge perfectness and classes of bipartite graphs. *Discret. Math.* **1996**, *149*, 159–187. [CrossRef]
14. Amilhastre, J.; Vilarem, M.C.; Janssen, P. Complexity of minimum biclique cover and minimum biclique decomposition for bipartite domino-free graphs. *Discret. Appl. Math.* **1998**, *86*, 125–144. [CrossRef]
15. Dawande, M.; Keskinocak, P.; Swaminathan, J.M.; Tayur, S. On bipartite and multipartite clique problems. *J. Algorithms* **2001**, *41*, 388–403. [CrossRef]
16. Javadi, R.; Maleki, Z.; Omoomi, B. Local Clique Covering of Graphs. *arXiv* **2012**, arXiv:1210.6965.
17. Arora, S.; Ge, R.; Sachdeva, S.; Schoenebeck, G. Finding overlapping communities in social networks: Toward a rigorous approach. In Proceedings of the 13th ACM Conference on Electronic Commerce, Valencia, Spain, 4–8 June 2012; pp. 37–54.
18. Cheng, Y.; Church, G.M. *Biclustering of Expression Data*; ISMB: Leesburg, VA, USA, 2000; Volume 8, pp. 93–103.
19. Král', D.; Kratochvíl, J.; Tuza, Z.; Woeginger, G. Complexity of coloring graphs without forbidden induced subgraphs. In *Graph-Theoretic Concepts in Computer Science*; Springer: Berlin, Germany, 2001; pp. 254–262.
20. Fellows, M.R.; Kratochvíl, J.; Middendorf, M.; Pfeiffer, F. The complexity of induced minors and related problems. *Algorithmica* **1995**, *13*, 266–282. [CrossRef]