Machine Learning, 56, 115–151, 2004

A New Conceptual Clustering Framework

NINA MISHRA* nina.mishra@cs.stanford.edu
HP Labs, Palo Alto, CA 94304, USA; Department of Computer Science, Stanford University, Palo Alto, CA 94305, USA

DANA RON† danar@eng.tau.ac.il
Department of EE-Systems, Tel-Aviv University, Tel Aviv 69978, Israel

RAM SWAMINATHAN swaram@hpl.hp.com
HP Labs, Palo Alto, CA 94304, USA

Editor: Philip Long

Abstract. We propose a new formulation of the conceptual clustering problem where the goal is to explicitly output a collection of simple and meaningful conjunctions of attributes that define the clusters. The formulation differs from previous approaches since the clusters discovered may overlap and also may not cover all the points. In addition, a point may be assigned to a cluster description even if it only satisfies most, and not necessarily all, of the attributes in the conjunction. Connections between this conceptual clustering problem and the maximum edge biclique problem are made. Simple, randomized algorithms are given that discover a collection of approximate conjunctive cluster descriptions in sublinear time.

Keywords: conceptual clustering, maximum edge biclustering

1. Introduction

Clustering is the problem of grouping similar objects together. In this paper we study the problem of identifying tight descriptions of large groups of points. As a motivating application, consider the problem of selling two or more items at a single combined price, also known as product bundling. For a collection of d products $\mathcal{P}$, we are given customer purchases $P_1, \ldots, P_n \subset \mathcal{P}$ and the goal is to identify k conjunctions of products, or product bundles, $C_1, \ldots, C_k$, with the property that many customers purchase the bundle C_i for each i. Such product bundles can then be used on a promotional basis by offering a discount to individuals who purchase all items in the bundle in one transaction. While we use this motivating example throughout, our formulation is useful in other application contexts such as text clustering (Dhillon, 2001; Dhillon, Mallela, & Modha, 2003).

In Conjunctive Clustering, the goal is to identify long conjunctive cluster descriptions that cover a dense region of space. More formally, a conjunctive cluster is a conjunction of attributes c together with the points Y in the data set that satisfy the conjunction c. In general

*Research partially supported by NSF Grant EIA-0137761.
†Research partially supported by an HP Labs Research Grant.

we are interested in longer, more specific conjunctions since then the points that satisfy the conjunction have more in common. We are also interested in having a large number of points satisfy that conjunction. A natural way to combine these objectives is to maximize $|c| \cdot |Y|$ so that we simultaneously ensure many points are in the cluster and also that those points have much in common. There may be other ways to combine these objectives as we discuss in Section 8.

A convenient way to think about a conjunctive cluster is as a biclique in a bipartite graph. For a bipartite graph $G = (U, W, E)$, let U be the points to be clustered (for the sake of simplicity, assume $U \subset \{0, 1\}^d$), and let W correspond to the attributes or dimensions. Let there be an edge between $u \in U$ and $w \in W$ if the wth dimension of u is 1.[1] A biclique is a subgraph (U^*, W^*) where each vertex in U^* is adjacent to each vertex in W^*. A biclique naturally corresponds to a conjunctive cluster since each point u in U^* satisfies the conjunction of attributes in W^*. A maximum edge biclique corresponds to the best conjunctive cluster, since $|W^*|$ is precisely the length of the conjunction and $|U^*|$ is the number of points that satisfy the conjunction. We define the k best conjunctive clusters as the k largest clusters that do not overlap too much. A formal definition can be found in Section 2.

Since the maximum edge biclique problem is NP-hard (Peeters, 2003) and there is evidence that it is difficult to approximate (Feige, 2002), we relax the problem to allow our algorithm to produce bisubgraphs (U', W') where each vertex in W' connects to most vertices in U'. In practice such a relaxation is quite natural since it allows a point to be assigned to a conjunctive cluster description even if it does not satisfy all attributes in the conjunction, but rather most of the attributes. Moreover, our algorithm only outputs one side of each approximate biclique—the one that corresponds to the description, or the common attributes of the cluster. The points that belong to the cluster can then be determined using this description. The running time required to obtain the description depends only logarithmically on the number of data points.

The conjunctive clustering formulation possesses some characteristics that are desirable in certain applications. The first characteristic is that conjunctive cluster descriptions are identified. Returning to the product bundling example, the algorithm will directly output what is required: a conjunctive description of products that many customers have purchased. A second characteristic is that not all points are clustered. In this application, it is in fact desirable to ignore the customers with unusual buying patterns. Another characteristic is that clusters overlap; it may be desirable to allow some shopping baskets to belong to multiple clusters since customers may actually purchase multiple bundles in one shopping excursion. In addition, the algorithm's running time is sublinear in the number of points to be clustered. This may be desirable for companies with a large number of customers. Finally, instead of requiring an input metric space, the formulation allows for clustering categorical data which is more suitable for the product bundling scenario.

While there have been approaches to clustering that either find cluster descriptions (Pitt & Reinke, 1987; Mishra, Oblinger, & Pitt, 2001), do not require a strict partition (Dempster, Laird, & Rubin, 1997), have algorithms with sublinear running time (Indyk, 1999; Mishra, Oblinger, & Pitt, 2001; Alon et al., 2003), or cluster categorical data (Guha, Rastogi, & Shim, 2000; Gibson, Kleinberg, & Raghavan, 2000; Ganti, Gehrke, & Ramakrishnan, 1999),

conjunctive clustering is an attempt at designing a problem formulation with all of these characteristics.

1.1. Our results

Maximum conjunctive cluster/maximum edge biclique. We start by considering the problem of finding a maximum conjunctive cluster, that is, a biclique (U^*, W^*) with the most edges. As mentioned, since this problem is hard, we consider a relaxation where the goal is to identify a subgraph that is both very dense and also very large. For $\hat{U} \subseteq U$, $\hat{W} \subseteq W$ and $0 < \epsilon \leq 1$, we say that $(\hat{U}, \hat{W})$ is ϵ-*close* to being a biclique if every vertex in $\hat{U}$ neighbors at least $(1 - \epsilon)$ of the vertices in $\hat{W}$. We say in such a case that it is an ϵ-*biclique*. For any given ϵ, our algorithm outputs a subgraph that is both an ϵ-biclique and also has almost as many edges as an optimum biclique. Actually, rather than outputting the subgraph $(\hat{U}, \hat{W})$, the algorithm only outputs $\hat{W}$ which corresponds to the cluster description. $\hat{U}$ is implicitly determined from $\hat{W}$, i.e., $\hat{U}$ contains all vertices in U that neighbor at least $(1 - \epsilon)$ of the vertices in $\hat{W}$.

Our algorithm runs efficiently provided that the fraction of points in both U^* and W^* is sufficiently large. Indeed, if $|U^*| \geq \rho_{\mathrm{U}} \cdot |U|$ and $|W^*| \geq \rho_{\mathrm{W}} \cdot |W|$, for certain parameters $\rho_{\mathrm{U}}, \rho_{\mathrm{W}}$ where $0 < \rho_{\mathrm{U}}, \rho_{\mathrm{W}} \leq 1$, then our algorithm draws a sample from U of size polynomial in the input parameters $1/\epsilon$, $1/\rho_U$, and $1/\rho_{\mathrm{W}}$, and runs in time linear in $|W|$, quasi-polynomial in $1/\rho_{\mathrm{U}}$ and $1/\rho_{\mathrm{W}}$, and exponential in poly$(1/\epsilon)$. Thus, the number of sample points is independent of $|U|$. The running time depends logarithmically on $|U|$ only because $O(\log |U|)$ bits are required to specify an element of U. Since $\log |U| \leq |W|$ there is no explicit dependence on $|U|$ in the running time.

While it would be more desirable to have an algorithm with running time polynomial in all problem parameters, we cannot expect to have polynomial dependence on $1/\epsilon$ since in such a case we could use the algorithm to solve the original NP-hard problem in polynomial time by setting $\epsilon < \frac{1}{|U| \cdot |W|}$. We leave open the question of whether it is possible to obtain an algorithm with polynomial dependence on $1/\rho_{\mathrm{U}}$ and $1/\rho_{\mathrm{W}}$. This paper addresses the situation when both $1/\rho_{\mathrm{U}}$ and $1/\rho_{\mathrm{W}}$ are small, for example, when ρ_{U}, ρ_{W} are constants like $1/3$. Such a situation has practical motivation. For instance, product bundling schemes are often designed to affect large portions, e.g., 10–30%, of the customer population.

Collection of large conjunctive clusters. We next discuss the more general problem of identifying a collection of large ϵ-bicliques. For a subgraph A, let $E(A)$ denote the edges in A. We say that a subgraph A *dominates* another subgraph B if the ratio of $|E(A) \backslash E(B)|$ to $|E(A) \cup E(B)|$ is small. The goal of Conjunctive Clustering is to find k large ϵ-bicliques where no ϵ-biclique dominates another. More precisely, given parameters ρ_{U} and ρ_{W} and a parameter k that denotes the number of desired clusters, we give an algorithm that outputs k subsets, $\hat{W}_1, \ldots, \hat{W}_k$ where $\hat{W}_i \subseteq W$, for which the following holds: (1) For every $1 \leq i \leq k$, $|\hat{W}_i| \geq \rho_{\mathrm{W}} \cdot |W|$. (2) For every $1 \leq i \leq k$, the subset $\hat{U}_i$ consists of all vertices in U that neighbor at least $(1 - \epsilon)$ of the vertices in $\hat{W}_i$. (3) The different ϵ-bicliques do not dominate each other. (4) For every biclique (U', W') such that $|U'| \geq \rho_{\mathrm{U}} \cdot |U|$ and $|W'| \geq \rho_{\mathrm{W}} \cdot |W|$, either there is an ϵ-biclique in our collection that dominates (U', W') or all

ϵ-bicliques in our collection are almost as large as (U', W'). If the fourth condition is satisfied, we say that our collection *swamps* the large bicliques. The running time of our algorithm is quasi-polynomial in k, $1/\rho_{\mathrm{U}}$ and $1/\rho_{\mathrm{W}}$, exponential in poly$(1/\epsilon)$, and linear in $|W|$.

Finding approximations to ϵ-bicliques. The problems described so far assume that the optimum true single cluster is a biclique. But in practice, the optimum true cluster is likely to be an ϵ-biclique. Thus we also consider the problem of closely approximating the largest ϵ-biclique. Specifically, we describe an algorithm that outputs a $4\epsilon^{1/3}$-biclique that is almost as large as the largest ϵ-biclique.

Finding very small, implicit representations of dense subgraphs. By slightly modifying our algorithms we can obtain very small, implicit representations of dense large subgraphs. In the case of finding an approximation to the maximum biclique, the algorithm will find small subsets $S \subset U$ and $S' \subset W$ that can be used to obtain a bisubgraph $(\hat{U}, \hat{W})$ that contains at least $(1 - O(\epsilon)) \cdot |\hat{U}| \cdot |\hat{W}|$ edges. In addition, the number of edges is almost as large as the size of a maximum biclique. Thus, we improve our running time at the expense of the quality of the output, since it is no longer the case that every vertex in $\hat{U}$ neighbors almost all vertices in $\hat{W}$, but rather that almost all vertices in $\hat{U}$ neighbor almost all vertices in $\hat{W}$. This result is more appealing in the general context of sublinear graph algorithms than in the particular context of conjunctive clustering, since in the latter we would like to know that every two members of the cluster share much in common.

Data streams. A data stream is a sequence of points $u_1, \ldots, u_i, \ldots, u_{|U|}$ that can only be read once in increasing order of the indices i. For many modern applications, the notion of a data stream is more appropriate than a static dataset. We give an algorithm that stores a sketch of the stream that can be used to identify conjunctive clusters. If the data is actually arriving in sequenced chunks, $C_1, \ldots, C_J$, where each C_i is a collection of points, then the size of the sketch is quasi-polynomial in $\frac{J}{\rho_{\mathrm{U}}}$ and $\frac{J}{\rho_{\mathrm{W}}}$, linear in $|W|$, and exponential in poly$(1/\epsilon)$.

1.2. The algorithms

Our algorithms are based on the following idea. Consider a fixed biclique (U^*, W^*) such that $|U^*| \geq \rho_{\mathrm{U}}|U|$ and $|W^*| \geq \rho_{\mathrm{W}}|W|$. In particular, this may be a maximum edge biclique. Assume that W^* is maximal in the sense that it is not possible to add any vertex to W^* and still obtain a biclique. Then, by definition of bicliques, W^* is simply the intersection of the sets of neighbors that vertices in U^* have in W. Let us denote this intersection by $\Gamma(U^*)$.

Suppose, as a mental experiment, we were given the ability to sample from U^*. Then for any sample $S \subseteq U^*$, we have that $W^* \subseteq \Gamma(S)$. The problem is that $\Gamma(S)$ may include additional vertices that do not belong to W^*, so that it is not immediately clear how to use the fact that $\Gamma(S)$ contains all vertices in W^*, as W itself also contains all vertices in W^*. However, it can be shown that if we uniformly select a sample of sufficiently large size then the following holds with high probability: If we take all vertices in U that neighbor all but at most an ϵ-fraction of the vertices in $\Gamma(S)$, then we obtain an ϵ-biclique that has at least as many edges as (U^*, W^*). We shall say in such a case that S is a *good seed* U^*.

Since we cannot actually sample from U^*, we instead consider all subsets of the sample whose size is lower bounded by a certain threshold. This technique is sometimes referred to as exhaustive sampling (Fernandez de la Vega, 1996; Arora, Karger, & Karpinski, 1995; Goldreich, Goldwasser, & Ron, 1998; Frieze & Kanan, 1999). It can be verified that if the sample is sufficiently large, then with high probability one of these subsets is a good seed. However, now we have to address a new problem: How do we decide which subset is the good seed? We could of course check the resulting ϵ-biclique for each subset, but this would take time linear in $|U|$, and we are interested in an algorithm having time sublinear in $|U|$. As one may guess at this point, we solve this by sampling again from U. While we described the ideas for the case of finding an approximate biclique, the ideas can be extended to finding a collection of bicliques.

1.3. Related work

Clustering. The field of clustering is quite extensive so we briefly comment on some of the more popular and/or recent clustering problems. One of the most widely studied clustering objectives is known as k-Median. Here we are given a set of points S in a metric space and must find a collection of k centers, which are themselves points in the metric space, such that the average distance from a point in S to its nearest center is minimized. Numerous clustering algorithms have been proposed for identifying approximately good clusterings, including (Jain & Vazirani, 1999; Charikar & Guha, 1999; Arya et al., 2001; Thorup, 2001). Sublinear versions of these algorithms can be found in Indyk (1999) and Mishra, Oblinger, and Pitt (2001). Related to the k-Median objective is the k-Median-squared objective of minimizing the average squared distance from a point its nearest center. Theoretical results for k-Median typically also apply to the k-Median-squared objective, but approximation algorithms have also been explicitly derived for this objective (Kanungo et al., 2002).

Among the more widely-used practical algorithms are k-Means (Duda, Hart, & Stork, 2000), which is known to find a local optimum solution to the k-Median[2] objective (Selim & Ismail, 1984), and EM (Dempster, Laird, & Rubin, 1997) which can for example be used to estimate the parameters of a mixture of Gaussians.

In the k-Center problem, the goal is to minimize the maximum radius of a cluster where the radius is defined to be the largest distance from a point to its nearest center. This problem is known to have a 2-approximation (Feder & Greene, 1988; Hochbaum & Shmoys, 1986). Sublinear algorithms are also known for the k-center problem (Alon et al., 2003).

Many other clustering objectives have recently been proposed for which approximation algorithms are also known, e.g., correlation clustering (Bansal, Blum, & Chawla, 2002; Charikar, Guruswami, & Wirth, 2003; Emmanuel & Fiat, 2003; Demaine & Immorlica, 2003), conductance-based clustering (Kannan, Vempala, & Vetta, 2000), and catalog segmentation (Kleinberg, Papadimitriou, & Raghavan, 1998).

Simultaneously clustering points and attributes, commonly known as biclustering, was introduced in Hartigan (1972), and has regained renewed interest recently. Specifically, a polynomial-time algorithm for identifying bounded width axis-parallel hyper-rectangles containing a constant fraction of the input points was described in Procopiuc et al. (2002). Biclustering has also been used to cluster gene expression data (e.g., Cheng & Church, 2000; Tanay, Sharan, & Shamir, 2002; Murali & Kasif, 2003). For a survey see Madeira

and Oliveira (2004). Finally, Dhillon, Mallela, and Modha (2003) provide an information-theoretic formulation of biclustering and present an iterative algorithm for reaching a local minimum.

Graph algorithms. As mentioned previously, the maximum-edge biclique problem is NP-hard (Peeters, 2003). Under the assumption that refuting 3SAT is hard on average, Feige (2002) shows that it is hard to approximate the maximum-edge biclique problem to within a certain constant. Furthermore, for certain constants $\alpha < \beta$, it is hard to distinguish between the case in which the maximum biclique has size at least $\beta \cdot |U| \cdot |W|$ and the case in which the maximum biclique has size less than $\alpha \cdot |U| \cdot |W|$. This implies that it is hard to obtain a $(1 + \gamma)$-approximation for a sufficiently small constant γ.

If the graph G is weighted, then approximating the weight of a maximum weighted edge biclique is as hard as the problem of approximating the size of a maximum clique in an arbitrary graph. The NP-hardness proof given in Hochbaum (1998) essentially shows that approximation is hard. Hochbaum (1998) gives a 2-approximation to the problem of minimizing the number of edges that need to be deleted so that the remaining graph is a biclique. Finding the biclique with the largest number of vertices is NP-hard when both sides of the biclique have to be of the same size (Garey & Johnson, 1979), but can be solved in polynomial time when this constraint is removed (Yannakakis, 1981).

Another problem concerning the search for large bicliques is that of finding a small collection of bicliques that form a partition of the edge set of the graph, also known as a *biclique partition*. The *order* of the partition is the sum of the numbers of vertices in the different bicliques, and the goal is to find a biclique partition having minimum order. Feder and Motwani (1995) study this problem in the context of graph compression. They show that the problem is NP-hard, and provide an efficient algorithm that finds a partition of order $O(\frac{m \log(n^2/m)}{\log n})$ where n is the number of vertices and m is the number of edges.

Related to the problem of finding relaxed bicliques is finding subgraphs with maximum average degree, sometimes called dense subgraphs. Finding a maximum density subgraph of a particular size is NP-hard, since CLIQUE is NP-hard. The algorithm in Peleg, Feige, and Kortsarz (2001) gives an approximation factor of $O(n^{1/3})$ and the algorithm of Arora, Karger, and Karpinski (1995) gives a polynomial-time approximation scheme for dense graphs. A maximum density subgraph without size constraints can be found in polynomial time (Goldberg, 1984; Charikar, 2000).

Property testing. Our algorithms are related to Property Testing algorithms on dense graphs (Goldreich, Goldwasser, & Ron, 1998), and in particular, are inspired by the CLIQUE-testing algorithm in that paper. Such algorithms are designed to decide whether a given dense graph has a certain property or whether many edge modifications should be performed so that it obtains the property. In a manner that is similar to the approximate solutions studied in this paper, many testing algorithms can be modified so as to obtain approximate solutions to the corresponding search problems. However, none of the known property testing algorithms, or their extension to approximation algorithms, directly apply to our problem. In particular, the most general family of graph properties studied in Goldreich, Goldwasser, and Ron (1998) does not capture our definition of clustering which allows for overlapping subsets of vertices. Other related work on approximation algorithms and

testing algorithm on dense graphs includes (Arora, Karger, & Karpinski, 1995; Frieze & Kanan, 1999; Alon et al., 2000).

Conceptual clustering. The general notion of finding cluster descriptions is known as conceptual clustering (Michalski, 1980; Fisher & Langley, 1985). Pitt and Reinke (1987) show that the Hierarchical Agglomerative Clustering (HAC) algorithm finds an optimum clustering under particular conditions on intra cluster distance, described by the distance between points within a cluster, and inter cluster distance, described by the distance between clusters. We discuss one instantiation of their results that is relevant to conjunctive clustering. Let the distance between two conjunctions c_i, c_j be the number of literals x such that x is in c_i and $\bar{x}$ is in c_j. For a clustering $c_1, \ldots, c_k$, let the intercluster distance be the minimum distance between two conjunctions, and let the intracluster distance be the length of the shortest conjunction. If the goal is to maximize the difference between the inter and intra cluster distance, then the HAC algorithm can find the optimum clustering. One advantage of this algorithm is that the number of clusters k need not be specified a priori. A separate conjunctive clustering problem, considered in Mishra, Oblinger, and Pitt (2001), is that of finding $k \geq 2$ conjunctive descriptions $c_1, \ldots, c_k$ such that $\sum_{i=1}^{k} |c_i| \cdot |Y_i|$ is maximized, and no point satisfies both c_i and c_j. An algorithm for finding the optimum solution was shown to have running time $O(d^{O(k^2)})$, where d is the number of attributes. These two results are not applicable to our problem since we do not require that each point be assigned to a cluster, that the clusters be disjoint, or that a point exactly satisfy a conjunction in order to be assigned to it. A separate paper on identifying descriptions of clusters by Agrawal et al. (1998) gives algorithms for identifying DNF descriptions for each cluster. In this work the objective function is different in that a cluster is a union of connected regions with more density within the region than the area around it.

Frequent itemsets. The frequent itemset problem (Agrawal, Imielinski, & Swami, 1993; Gunopulos et al., 1997) is also closely related to conjunctive clustering. Given a collection of points U in $\{0, 1\}^d$, the frequent itemset problem is that of identifying all subsets of variables that satisfy a sufficiently large fraction of U, also known as *support*. Since all the frequent subsets can be deduced from the maximally frequent subsets, many algorithms are specifically designed to identify the smaller collection of maximally frequent subsets. A large conjunctive cluster is in some sense a maximally frequent itemset. The key difference between the two formulations is in the required output. In the frequent set formulation, one is interested in identifying all frequent itemsets, or possibly all maximally frequent itemsets. As such, the identification of a border separating the frequent from the infrequent is crucial to solving the problem. In contrast, our objective is to find a collection of k conjunctions that do not overlap too much and that swamp all the large conjunctions. Thus, we seek a subset of the frequent itemsets which may be significantly smaller in cardinality than the number of maximally frequent sets and also serve a different purpose in terms of overlapping with or being larger than all the frequent sets. Finally, while the Apriori (Agrawal, Imielinski, & Swami, 1993) algorithm could be used to find all conjunctive clusters, the running time would be prohibitive.

Web communities. Research on discovering web communities (Kumar et al., 1999; Gibson, Kleinberg, & Raghavan, 1998; Flake, Lawrence, & Giles, 2000) is also related

to conjunctive clustering. A web community is a set of web pages that are all relevant to each other. One way to view the community discovery problem is as a bipartite graph $G = (U, W, E)$ where $U = W$ are the pages on the web and E consists of edges (u, w) if there is a hyperlink from u to w or if $u = w$. A biclique (U', W') forms a community since each page in U' is linked to each page in W'. Our results can be used to identify a collection of communities that dominate the large communities on the web. In contrast, our algorithms are not designed to find small communities, also known as "cores" as studied by Kumar et al. (1999), where the goal is to, for example, find all $K_{3,2}$'s.

1.4. Overview

In Section 2 we define bicliques and their relationship to conjunctive clusters. We then define the conjunctive clustering problem that we study. In Section 3, we describe a method called *Good Seed* for almost reconstructing a biclique given the ability to sample from one side of the biclique. In Section 4, we use the Good Seed algorithm as a basis for identifying a collection of conjunctive clusters. In the event that there are no 'true' bicliques in the bipartite graph but only 'relaxed' bicliques, we show in Section 5 that the algorithm can find a collection of more-relaxed bicliques that dominate the relaxed bicliques. In Section 6, we present a variant of our algorithm for finding an approximate maximum biclique that runs in time logarithmic in both $|U|$ and $|W|$. Finally, in Section 7, we describe a streaming version of the algorithm.

2. Preliminaries and problem definitions

As noted in the introduction, it will be convenient to define our problems using a graph-theoretic formulation. Given a bipartite graph $G = (U, W, E)$ and two subsets $U' \subseteq U$ and $W' \subseteq W$, we denote by $E(U', W')$ the subset of all edges between vertices in U' and vertices in W'. We refer to such a pair (U', W') as a *bisubgraph*. The *size* of a bisubgraph is the number of edges contained in the bisubgraph. For a vertex v we denote the neighbor set of v by $\Gamma(v)$. For a subset S of vertices, we let $\Gamma(S) \stackrel{\text{def}}{=} \cap_{v \in S} \Gamma(v)$ denote the set of vertices each of which neighbor *every* vertex in S. For a subset S and a parameter $\epsilon \leq 1/2$, we let $\Gamma_\epsilon(S) \stackrel{\text{def}}{=} \{w : |\Gamma(w) \cap S| \geq (1 - \epsilon)|S|\}$ denote the set of vertices that neighbor all but an ϵ-fraction of S. For an illustration of $\Gamma(S)$ and $\Gamma_\epsilon(S)$ see figure 1.

Definition 1. Given a bipartite graph $G = (U, W, E)$, a bisubgraph (U', W') is a biclique if $E(U', W') = U' \times W'$. The size of a biclique (U', W') is $|U'| \cdot |W'|$, and a maximum biclique is a biclique (U', W') for which $|U'| \cdot |W'|$ is maximized over all bicliques.

A maximum biclique is uniquely determined by either of its sides: U^* or W^*. Given W^*, we can obtain $U^* = \Gamma(W^*)$, and vice versa.

As noted previously, the maximum biclique problem is NP-hard. Here we suggest a relaxation of the maximum biclique problem which allows the output to be close to a biclique.

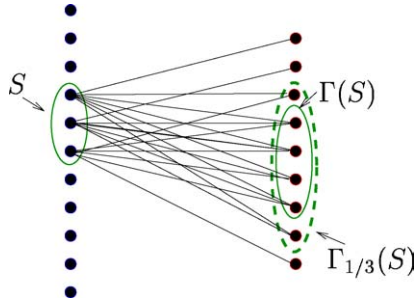

Figure 1. An illustration of a set S, and of $\Gamma(S)$ and $\Gamma_\epsilon(S)$, for $\epsilon = 1/3$. The set $\Gamma_\epsilon(S)$ is enclosed in a dashed ellipse. To avoid clutter, only the edges incident to vertices in S are depicted.

Definition 2. We say that (U', W') is ϵ-close to being a biclique, for $0 \leq \epsilon \leq 1$, if every vertex in U' neighbors at least $(1 - \epsilon)$ of the vertices in W'. For the sake of succinctness, we say that (U', W') is an ϵ-biclique.

In the context of conjunctive clusters, an ϵ-biclique corresponds to a pair (Y, c) such that every point in Y satisfies at least $(1 - \epsilon)$ of the attributes in c. The asymmetry between U' and W' in the definition of an ϵ-biclique corresponds to our needs in the context of clustering where the two sides of the ϵ-biclique in fact have a different role. Similarly to what was noted above for bicliques, if an ϵ-biclique (U', W') is maximal then it is completely determined by W', that is, $(U', W') = (\Gamma_\epsilon(W'), W')$. This is especially useful in the context of clustering since $W' = c$ is the description of the cluster, and therefore we do not need to output explicitly all points Y in the cluster.

Our first problem formulation follows.

Problem 1. Given a bipartite graph $G = (U, W, E)$, find a subset $W' \subseteq W$ such that the ϵ-biclique $(\Gamma_\epsilon(W'), W')$ is at least $(1 - b\epsilon)$ times as large as the maximum biclique for a small constant b.

We present an algorithm that solves this problem for $b = 2$. The running time of the algorithm depends linearly on $|W|$ and only logarithmically on $|U|$. If we allow running time linear in $|U|$, the ϵ-biclique output by the algorithm contains at least as many edges as the maximum biclique, in other words, $b = 0$.

Collections of large bicliques. The above relaxation addresses the issue of finding a single approximate maximum biclique. We now turn to defining a good collection of at most k bicliques.

A natural way to define the k best conjunctive clusters is as the k largest clusters, i.e., the k bicliques with the most total number of edges. The problem with this definition is that clusters may overlap in terms of both the attributes that define the conjunctions and in terms of the points that satisfy the conjunctive descriptions. Thus the two largest clusters (U_1, W_1) and (U_2, W_2) may be such that $U_1 \cap U_2$ is almost identical to both U_1, U_2 and

that $W_1 \cap W_2$ is almost identical to both W_1, W_2. In such a case, we would prefer that the algorithm output just one of these bicliques since they are so similar.

Before defining the k best clusters, we first discuss what it means for one cluster to "dominate" another. Any definition of dominate must address overlap in terms of both the descriptions and the points satisfied by the description. In other words, it is possible that two cluster descriptions are very similar, but that the descriptions satisfy disjoint point sets. In such a case, we would say that the clusters do not overlap since they satisfy distinct sets of points. In the other direction, if two very similar sets of points are satisfied by completely different conjunctions, then we would also say that the clusters do not overlap. Thus any definition of dominates should address overlap in terms of both U and W.

Next we observe that "dominates" is not symmetric. Consider the situation when $U_1 \subseteq U_2$, $W_1 \subseteq W_2$, $|U_2| \gg |U_1|$, and $|W_2| \gg |W_1|$. In such a case, we would say that (U_2, W_2) dominates (U_1, W_1) since each point/attribute combination of (U_1, W_1) appears in (U_2, W_2). However, we would not say that (U_1, W_1) dominates (U_2, W_2) since most of the point/attribute combinations from (U_2, W_2) are missing from (U_1, W_1).

Having established that we want a definition of dominates that takes into account both U and W and is asymmetric, there may be several natural definitions of dominates. Since our clusters correspond to bisubgraphs with large edge density, we define dominates in terms of the subsets of edges in each bisubgraph.

Definition 3. Let $G = (U, W, E)$ be a bipartite graph and let $U_1, U_2 \subseteq U$ and $W_1, W_2 \subseteq W$. We say that (U_1, W_1) δ-dominates (U_2, W_2) if the number of edges in $E(U_2, W_2)$ that do not belong to $E(U_1, W_1)$ is at most a δ fraction of the union of the two sets of edges:

$$\frac{|E(U_2, W_2) \backslash E(U_1, W_1)|}{|E(U_2, W_2) \cup E(U_1, W_1)|} \leq \delta.$$

In the above definition, the size of the uncovered set of edges, $|E(U_2, W_2) \backslash E(U_1, W_1)|$ is normalized by the size of the union of the two sets of edges, $|E(U_2, W_2) \cup E(U_1, W_1)|$, rather than by the the number of edges in the covered bisubgraph, $|E(U_2, W_2)|$. While the latter seems to be the more natural choice, we have chosen the former for technical reasons (e.g., it obeys a certain triangle-like inequality). Note that if the size of the dominating bisubgraph, $|E(U_1, W_1)|$ is not much larger than the size of the dominated bisubgraph, $|E(U_2, W_2)|$, then the two measures are roughly the same. Since we will be interested in dominating relatively large bisubgraphs, namely, such that $|U_2| \geq \rho_{\mathrm{U}} \cdot |U|$ and $|W_2| \geq \rho_{\mathrm{W}} \cdot |W|$, the ratio between the two measures is always upper bounded by $1/(\rho_{\mathrm{U}} \cdot \rho_{\mathrm{W}})$.

We use the definition of dominates in two ways: to make sure that the clusters our algorithm outputs dominate the clusters in the optimum solution, and also to ensure that the clusters we output are distinct enough from each other.

Since our clustering objective is to identify a collection of k ϵ-bicliques, we need some method of comparing our discovered ϵ-bicliques to the true bicliques. Consider a fixed large true biclique (U', W'). We say that our collection of k ϵ-bicliques swamps a fixed large true biclique (U', W') if either (a) we have one in our collection that δ-dominates (U', W') or (b) (U', W') is smaller than every ϵ-biclique we find. Actually, since our interest is in sublinear

algorithms, condition (b) is replaced with a relaxation where we ensure that (U', W') is not much larger than every ϵ-biclique we find.

Definition 4. Let $G = (U, W, E)$ be a bipartite graph and let $\mathcal{C} = \{(U_i, W_i)\}_{i=1}^{k}$ be a collection of pairs of vertex subsets where $U_i \subseteq U$, and $W_i \subseteq W$. We say that $\mathcal{C}$ ($\boldsymbol{\delta}$,$\boldsymbol{\epsilon}$)-swamps a pair (U', W') if either there exists a pair $(U_i, W_i) \in \mathcal{C}$ that δ-dominates (U', W'), or $|E(U', W')| \leq (1+\epsilon) \cdot \min_j \{|E(U'_j, W'_j)|\}$.

We will also sometimes say that a collection $\mathcal{C}_1$ of subgraphs swamps another collection of subgraphs $\mathcal{C}_2$ if $\mathcal{C}_1$ swamps each subgraph $(U_2, W_2) \in \mathcal{C}_2$.

We introduce one more important definition that is based on the notion of dominates. This definition ensures that in the collection of subgraphs output by the algorithm no subgraph dominates another.

Definition 5. Let $G = (U, W, E)$ be a bipartite graph and let $\mathcal{C} = \{(U_i, W_i)\}_{i=1}^{k}$ be a collection of pairs of vertex subsets where $U_i \subseteq U$, and $W_i \subseteq W$. We say that $\mathcal{C}$ is $\boldsymbol{\delta}$-diverse if for every two different pairs (U_i, W_i) and (U_j, W_j) in $\mathcal{C}$, neither δ-dominates the other.

Since this paper is focused on identifying large conjunctive clusters, we introduce two lower-bound parameters, ρ_{U} and ρ_{W}, which the algorithm is provided with, and consider only bicliques (U', W') such that $|U'| \geq \rho_{\text{U}} \cdot |U|$ and $|W'| \geq \rho_{\text{W}} \cdot |W|$. In the conjunctive clustering formulation, these parameters are quite natural since a bisubgraph with too few attributes or too few points may not be an interesting cluster. Let $\mathcal{B}(\rho_{\text{U}}, \rho_{\text{W}})$ denote the set of all bicliques (U', W') in G such that $|U'| \geq \rho_{\text{U}} \cdot |U|$ and $|W'| \geq \rho_{\text{W}} \cdot |W|$.

Given the above definitions, a natural problem is to find a collection of at most k bicliques in $\mathcal{B}(\rho_{\text{U}}, \rho_{\text{W}})$ that is both δ-diverse and (δ, ϵ)-swamps every $(U', W') \in \mathcal{B}(\rho_{\text{U}}, \rho_{\text{W}})$. Here we define a relaxation of this problem where we can output ϵ-bicliques. Let $\mathcal{B}_\epsilon(\rho_{\text{U}}, \rho_{\text{W}})$ denote the set of all ϵ-bicliques (U', W') in G such that $|U'| \geq \rho_{\text{U}} \cdot |U|$ and $|W'| \geq \rho_{\text{W}} \cdot |W|$.

Problem 2. Let $G = (U, W, E)$ be a given bipartite graph, $0 < \rho_{\text{U}}, \rho_{\text{W}} \leq 1$ two size parameters, k a positive integer, $0 \leq \delta \leq 1$ a diversity/dominating parameter, and $0 \leq \epsilon \leq 1$ an approximation parameter. Find a collection $\tilde{\mathcal{C}}$ of at most k ϵ-bicliques in $\mathcal{B}_\epsilon(\rho_{\text{U}}, \rho_{\text{W}})$ such that $\tilde{\mathcal{C}}$ is δ-diverse and for every $(U', W') \in \mathcal{B}(\rho_{\text{U}}, \rho_{\text{W}})$, (U', W') is $(b \cdot (\delta + \epsilon), b' \cdot \epsilon)$-swamped by $\tilde{\mathcal{C}}$ for some small constants b and b'. Furthermore, the collection $\tilde{\mathcal{C}}$ must either be of size k or, for some δ-diverse collection $\mathcal{C}^* \subseteq \mathcal{B}(\rho_{\text{U}}, \rho_{\text{W}})$ of k bicliques (if such exists), every biclique $(U_i^*, W_i^*) \in \mathcal{C}^*$ is $(\delta + b'' \cdot \epsilon)$-dominated by some ϵ-biclique in $\tilde{\mathcal{C}}$.

As in Problem 1, it suffices that the algorithm output only the subset $W' \subseteq W$ in each ϵ-biclique (U', W'). For this problem, we present an algorithm that works for $b = b'' = 4$ and $b' = 2$. Observe that without the additional requirement concerning a lower bound on the size of $\tilde{\mathcal{C}}$, Problem 2 would be no harder than Problem 1 (as the output of Problem 1 swamps all bicliques in $\mathcal{B}(\rho_{\text{U}}, \rho_{\text{W}})$). By adding this requirement we ensure that the algorithm either outputs a diverse collection of k large ϵ-bicliques or it "well-dominates" a diverse collection of k bicliques. Hence, the algorithm is allowed to output just a single ϵ-biclique

but in such a case this ϵ-biclique should dominate many diverse bicliques. As an extreme example, suppose (U, W) is itself an ϵ-biclique. Then by outputting (U, W), the algorithm actually dominates *all* bicliques.

3. A good seed

In this section we discuss a central building block of our algorithms. Consider a fixed biclique (U^*, W^*) and assume it is maximal. Suppose, as a mental experiment, that we can obtain a small, random sample S from U^*. We think of the sample S as being a "good seed" for the biclique (U^*, W^*). In this section we show how we can largely recover (U^*, W^*) with just a good seed by using the ideas discussed in Section 1.2. In the next section we remove the imaginary assumption that we can directly sample from U^* in order to obtain the good seed.

Let ρ_w be a lower bound on $|W^*|/|W|$, and let $\hat{m} = \frac{16}{\epsilon^2} \log \frac{80}{\rho_w \epsilon}$.

Good Seed Algorithm
1. Let S be a sample of size $\hat{m}$ drawn uniformly and independently from U^*. 2. Let $\hat{W} \leftarrow \Gamma(S)$. 3. Output $\hat{W}$.

Lemma 1. *Let $\hat{W}$ be as constructed in the Good Seed Algorithm. With probability at least $\frac{19}{20}$ over the choice of $S \subset U^*$, $|E(\Gamma_\epsilon(\hat{W}), \hat{W})| \geq |U^*| \cdot |W^*|$.*

In order to prove the lemma, it will be helpful to partition the vertices in $\hat{W}$ into three subsets. The first subset is W^*; while the second subset consists of vertices that do not belong to W^* but that neighbor a significant fraction of vertices in U^*. The second subset will be referred to as "high degree" vertices. The third subset consists of those vertices that neighbor relatively few vertices in U^*. These will be referred to as "low degree" vertices. We show that, with high probability, the size of the subset of vertices in $\hat{W}$ that have low degree with respect to U^* is small. Thus, since most of the vertices in $\hat{W}$ are either in the optimum biclique or have high degree with respect to U^*, we shall argue that the bisubgraph $(\Gamma_\epsilon(\hat{W}), \hat{W}))$ has at least as many edges as the optimum. We now precisely define the terms "high degree" and "low degree".

Definition 6. A vertex $w \in W$ has high degree with respect to U^* if

$$\frac{|\Gamma(w) \cap U^*|}{|U^*|} \geq 1 - (\epsilon/4)^2.$$

Otherwise it has low degree with respect to U^*.

Note that every $w \in W^*$ has high degree with respect to U^*, since for every $w \in W^*$, $\frac{|\Gamma(w) \cap U^*|}{|U^*|} = 1$. We will be interested in samples of U^* that are "good seeds" as defined below.

Definition 7. We say that a subset $S \subseteq U^*$ is a good seed of U^* if the number of vertices in $\Gamma(S) \subseteq W$ that have low degree with respect to U^* is at most $(\epsilon/4)|W^*|$.

Lemma 2. *With probability at least* $\frac{19}{20}$, *the sample S drawn in Step* 1 *of the Good Seed Algorithm is a good seed of* U^*.

Proof: Consider a fixed vertex $w \in W$ that has low degree with respect to U^*. Then, by definition of low degree vertices,

$$\Pr_S[w \in \Gamma(S)] < (1 - (\epsilon/4)^2)^{\hat{m}} < \exp(-(\epsilon/4)^2 \cdot \hat{m}) = \frac{\epsilon \cdot \rho_W}{80} \tag{1}$$

where the last inequality follows from the definition of $\hat{m} = \frac{16}{\epsilon^2} \log \frac{80}{\epsilon \cdot \rho_W}$. Hence, the expected number of vertices in $\Gamma(S)$ that have low degree with respect to U^* is bounded by $|W| \cdot \frac{\rho_W \cdot \epsilon}{80}$. By Markov's inequality, and using $|W^*| \geq \rho_W \cdot |W|$, the probability that $\hat{W} = \Gamma(S)$ contains more than $\frac{\epsilon}{4}|W^*|$ vertices with low degree with respect to U^* is at most $\frac{1}{20}$. □

We next show that if S is a good seed of U^* then the ϵ-biclique $(\Gamma_\epsilon(\hat{W}), \hat{W}) = (\Gamma_\epsilon(\Gamma(S)), \Gamma(S))$ has as many edges as (U^*, W^*).

Lemma 3. *Let S be a good seed of* U^* *and let* $\hat{W} = \Gamma(S)$. *Then we have the following*: (1) $|\Gamma_\epsilon(\hat{W})| \geq (1 - \epsilon/4)|U^*|$, (2) $(\Gamma_\epsilon(\hat{W}), \hat{W})$ $\epsilon/4$-*dominates* (U^*, W^*), *and* (3) $|E(\Gamma_\epsilon(\hat{W}), \hat{W})| \geq |U^*| \cdot |W^*|$.

An illustration for Lemma 3 can be found in figure 2.

Proof: The subset $\hat{W}$ consists of three parts: (1) the vertices of W^*; (2) a subset of vertices, denoted H, that do not belong to W^* and have high degree with respect to U^*; (3) a subset of vertices, denoted L, that have low degree with respect to U^*. Note that since S is a good seed of U^*, we have that $|L| \leq \frac{\epsilon}{4}|W^*|$. In what follows, let $\hat{U} \stackrel{\text{def}}{=} \Gamma_\epsilon(\hat{W})$. We consider two cases based on whether $|H|$ is small or large.

$|H| \leq \frac{\epsilon}{2}|W^*|$: In this case,

$$\frac{|W^*|}{|\hat{W}|} = \frac{|W^*|}{|H| + |L| + |W^*|} > \frac{|W^*|}{\frac{\epsilon}{2}|W^*| + \frac{\epsilon}{4}|W^*| + |W^*|} \tag{2}$$

$$= \frac{1}{1 + \frac{\epsilon}{2} + \frac{\epsilon}{4}} \geq (1 - \epsilon) \tag{3}$$

This implies that every vertex in U^*, which by definition neighbors every vertex in W^*, neighbors at least $(1 - \epsilon)$ of the vertices in $\hat{W}$. That is, $U^* \subseteq \hat{U}$, so that in particular, $|\hat{U}| \geq |U^*|$. Since we also have that $W^* \subseteq \hat{W}$ we get that $(\hat{U}, \hat{W})$ completely dominates (U^*, W^*) and so $|E(\hat{U}, \hat{W})| \geq |U^*| \cdot |W^*|$.

$|H| > \frac{\epsilon}{2}|W^*|$: In this case we first show that all but at most an $\epsilon/4$-fraction of the vertices in U^* have at least $(1 - \epsilon/4)|H|$ neighbors in H. Let the subset of vertices in U^* having at least $(1 - \epsilon/4)|H|$ neighbors in H be denoted Q^*. Thus we would like to show that $|Q^*| \geq (1 - \epsilon/4)|U^*|$.

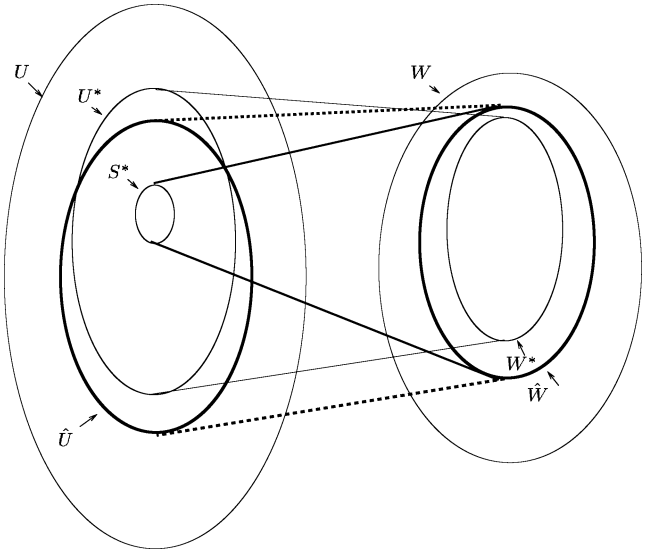

Figure 2. An illustration for Lemma 3. All subsets of vertices are denoted by ellipses. If a pair of subsets form a biclique, then there are two solid lines connecting them. If they form an ϵ-biclique, then there are dashed lines. Specifically, U^* and W^* are subsets of U and W, respectively, where (U^*, W^*) is a biclique. The small subset $S^* \subset U^*$ is a good seed of U^*. The subset $\hat{W} = \Gamma(S^*)$ $(\subseteq W)$, which is marked in bold, contains all vertices in W^* and possibly additional vertices, some having high degree with respect to U^* and very few having low degree. Finally, the subset $\hat{U} = \Gamma_\epsilon(\hat{W})$ $(\subseteq U)$, which is also marked in bold, contains almost all vertices in U^*. The resulting ϵ-biclique $(\hat{U}, \hat{W})$ has at least as many edges as the biclique (U^*, W^*).

Let α be such that $|Q^*| = (1-\alpha)|U^*|$. Observe that the total number of edges between U^* and H can be strictly upper bounded as follows:

$$\begin{aligned} |E(U^*, H)| &< |Q^*| \cdot |H| + |U^* \backslash Q^*| \cdot (1-\epsilon/4)|H| \\ &= (1-\alpha)|U^*| \cdot |H| \ + \ \alpha|U^*| \cdot (1-\epsilon/4)|H| \\ &= (1-\alpha+\alpha-(\alpha \cdot \epsilon/4))|U^*| \cdot |H| \end{aligned}$$

By definition of H, we can also lower bound the number of edges between U^* and H: $|E(U^*, H)| \geq |H| \cdot (1-(\epsilon/4)^2)|U^*|$. Combining the two previous equations, we have that

$$(1-(\epsilon/4)^2)|U^*| \cdot |H| \leq (1-(\alpha \cdot \epsilon/4))|U^*| \cdot |H|$$

which implies that $\alpha \leq \epsilon/4$ as desired.

Since $|L| \leq (\epsilon/4)|W^*|$, every vertex in Q^* has at least $(1-\epsilon)|\hat{W}|$ neighbors in $\hat{W}$, and hence $Q^* \subseteq \hat{U}$. Since we have shown that $|Q^*| \geq (1-\epsilon/4)|U^*|$ (where $Q^* \subseteq U^*$),

we have that $|\hat{U}| \geq (1 - \epsilon/4)|U^*|$ and $E(\hat{U}, \hat{W})$ contains all edges in $E(U^*, W^*)$ but at most $(\epsilon/4) \cdot |U^*| \cdot |W^*|$. Hence, $(\hat{U}, \hat{W})$ certainly $(\epsilon/4)$-dominates (U^*, W^*). Finally, by definition of Q^* and what we have shown concerning its size,

$$
\begin{aligned}
|E(Q^*, \hat{W})| &\geq |Q^*| \cdot \big(|W^*| + (1 - \epsilon/4)|H|\big) && (4)\\
&\geq (1 - \epsilon/4)|U^*| \cdot \big(|W^*| + |H|(1 - \epsilon/4)\big) \\
&> |U^*| \cdot |W^*| \cdot (1 - \epsilon/4) \cdot (1 + (\epsilon/2)(1 - \epsilon/4)) && (5)\\
&> |U^*| \cdot |W^*| && (6)
\end{aligned}
$$

Since $|E(\hat{U}, \hat{W})| \geq |E(Q^*, \hat{W})|$, together with Eq. (6), the lemma follows. □

The proof of Lemma 1 directly follows from Lemmas 2 and 3.

4. Conjunctive clustering algorithm

We now turn to the problem of identifying conjunctive clusters. We begin by considering the problem of finding one large approximate conjunctive cluster. Then we consider the problem of finding a good collection of approximate conjunctive clusters.

4.1. Approximate maximum biclique

Given ρ_U and ρ_W for which $\mathcal{B}(\rho_U, \rho_W)$ is non-empty we show how to find an ϵ-biclique in which the number of edges is almost as large as in a maximum biclique in $\mathcal{B}(\rho_U, \rho_W)$.[2] Solving this problem is interesting in its own right since no solution to the maximum edge relaxed biclique problem is known. In addition, the solution to this problem is used to identify k large, approximate, conjunctive clusters later in this section.

We cannot directly sample from the left-hand side, U^*, of a maximum biclique since the maximum biclique is unknown. Instead, we sample from U and consider all subsets S of size $\hat{m}$, where $\hat{m}$ is as defined in the Good Seed Algorithm. As we shall show, with high probability, at least one of these subsets, denoted S^*, is contained within U^* and furthermore, S^* is a good seed of U^*. In particular, for this subset S^*, the ϵ-biclique $(\Gamma_\epsilon(\Gamma(S^*)), \Gamma(S^*))$ has at least as many edges as the maximum biclique (U^*, W^*).

How do we determine which of the subsets S is the good seed S^*? One straightforward solution is to simply construct the sets $\Gamma_\epsilon(\Gamma(S))$, and to output $\Gamma(S)$ for which $|\Gamma_\epsilon(\Gamma(S))| \cdot |\Gamma(S)|$ (or $|E(\Gamma_\epsilon(\Gamma(S)), \Gamma(S))|$) is maximized. This however comes at a cost linear in $|U|$, while we are interested in an algorithm that only depends linearly on $|W|$, the set of attributes. Hence, we use additional sampling from U in order to approximate the size of $|\Gamma_\epsilon(\Gamma(S))| \cdot |\Gamma(S)|$ for every S.

Let $\hat{m}$ be as defined in the Good Seed Algorithm, let $m = \frac{2}{\rho_U} \cdot \hat{m}$, and let $t = \frac{96}{\rho_U \cdot \epsilon^2} \cdot m$.

Algorithm Approximate Maximum Biclique

1. Draw a sample X of m vertices uniformly and independently from U.
2. Draw an additional sample T of t vertices uniformly and independently from U.
3. For each subset S of X that has size $\hat{m}$ do:
   (a) $\hat{W}(S) \leftarrow \Gamma(S)$.
   (b) $\hat{T}(S) \leftarrow T \cap \Gamma_\epsilon(\hat{W}(S))$.
4. Among all subsets S considered by the algorithm for which $|\hat{W}(S)| \geq \rho_{\mathrm{w}}|W|$ and $|\hat{T}(S)| \geq (1-\epsilon/2)\rho_{\mathrm{U}}t$, let S^+ be the one for which $|\hat{T}(S^+)| \cdot |\hat{W}(S^+)|$ is maximized. Output $\hat{W}(S^+)$.

Theorem 1. *Let $\rho^* \cdot |U| \cdot |W|$ be the size of a maximum biclique in $\mathcal{B}(\rho_U, \rho_w)$. With probability at least* $2/3$, *Algorithm Approximate Maximum Biclique outputs a subset $\hat{W} = \hat{W}(S^+)$ so that*

$$|E(\Gamma_\epsilon(\hat{W}), \hat{W}))| \geq (1-2\epsilon) \cdot \rho^* \cdot |U^*| \cdot |W^*|$$

where $|\hat{W}| \geq \rho_w|W|$ and $|\Gamma_\epsilon(\hat{W})| \geq (1-\epsilon)\rho_U|U|$. The running time of the algorithm is exponential in $\mathrm{poly}(1/\epsilon)$, *quasi-polynomial in $\frac{1}{\rho_U}$ and $\frac{1}{\rho_w}$, and linear in $|W|$.*

In order to prove Theorem 1, we first introduce some notation and then prove a few lemmas from which the theorem will follow. Let (U^*, W^*) be a maximum biclique in $\mathcal{B}(\rho_{\mathrm{U}}, \rho_{\mathrm{w}})$. For any subset S, let $\hat{W}(S)$ and $\hat{T}(S)$ be as defined in the algorithm. Let $\hat{U}(S) = \Gamma_\epsilon(\hat{W}(S))$ and let $\hat{G}(S) = (\hat{U}(S), \hat{W}(S))$ be the bisubgraph determined by S. We define the *true relative size* of $\hat{G}(S)$ to be $\frac{|\hat{U}(S)|\cdot|\hat{W}(S)|}{|U|\cdot|W|}$ and the *estimated relative size* of $\hat{G}(S)$ to be $\frac{|\hat{T}(S)|\cdot|\hat{W}(S)|}{t\cdot|W|}$ (recall that $|T| = t$). We also define a *good subset* S_g to be one for which $\frac{|\hat{U}(S_g)|}{|U|} \geq (1-\epsilon)\rho_{\mathrm{U}}$ and a *bad subset* S_b to be one for which $\frac{|\hat{U}(S_b)|}{|U|} < (1-\epsilon)\rho_{\mathrm{U}}$.

The analysis of the algorithm works via the following reasoning. We first show that, with high probability, one of the subsets, S, considered in Step 3 of the algorithm is a good seed of U^*. Denote this subset by S^*. By Lemma 3 we know that (with high probability) $|\hat{U}(S^*)| \geq (1-\epsilon/4)\rho_{\mathrm{U}}|U|$, so that S^* is a good subset, and we know that the bisubgraph $(\hat{U}(S^*), \hat{W}(S^*))$ has at least as many edges as the maximum biclique (U^*, W^*). We then show that, with high probability, only good subsets are candidates for determining the set, $\hat{W}$, that is output by the algorithm. Furthermore, the estimated relative size of $\hat{G}(S_g)$ for every good subset S_g is close to its true relative size. Thus, the estimated relative size of $\hat{G}(S^*)$ for the seed S^* is close to its true relative size. It will then follow that for the set $\hat{W}(S^+)$ that is output by the algorithm, the bisubgraph $(\hat{U}(S^+), \hat{W}(S^+))$ must have almost as many edges as the maximum biclique (U^*, W^*).

We begin by showing that one of the subsets S considered in Step 3 of the algorithm is a good seed of U^*.

Lemma 4. *With probability at least* $9/10$, *one of the subsets considered in Step* 3 *of Algorithm Approximate Maximum Biclique is a good seed of U^*.*

Proof: Since $m = (2/\rho_{\mathrm{U}}) \cdot \hat{m}$, by a multiplicative Chernoff bound, with probability at least $1 - \exp(-(1/2)^3 \rho_{\mathrm{U}} m) = 1 - \exp(-\hat{m}/4) > 19/20$, at least $\hat{m}$ of the sampled vertices belong to U^*. Assume that this event in fact occurs. Since our algorithm enumerates over all subsets of size $\hat{m}$, it will come across a subset, call it S^*, that is contained in U^*. Since the m vertices drawn in Step 2 are drawn uniformly from U, the subset S^* is uniformly distributed in U^*. Recall that $|S^*| = \hat{m}$. Hence, by Lemma 2, with probability at least $19/20$, S^* is a good seed of U^*. Summing the probabilities of failure, the lemma follows. □

Lemma 5. *With probability at least* $19/20$, *for every bad subset* S_b, $|T(S_b)|/t < (1 - \epsilon/2)\rho_U$.

Proof: Consider a subset $R \subset W$ such that $|\hat{U}(R)|/|U|$ is exactly $(1 - \epsilon)\rho_{\mathrm{U}}$. For such a subset, by a multiplicative Chernoff bound

$$\begin{aligned} \Pr\left[\frac{|\hat{T}(R)|}{t} \geq (1 - \epsilon/2)\rho_{\mathrm{U}}\right] &\leq \Pr\left[\frac{|\hat{T}(R)|}{t} \geq (1 + \epsilon/2) \cdot (1 - \epsilon)\rho_{\mathrm{U}}\right] \\ &\leq e^{-(1/2)t(\epsilon/2)^2(1-\epsilon)\rho_{\mathrm{U}}} \end{aligned} \tag{7}$$

where the probability is taken over the choice of the sample T. Given our choice of $t = |T|$, and since $\epsilon \leq 1/2$, the above probability is upper bounded by e^{-4m}. Next consider any fixed bad subset S_b. Since the probability of obtaining a vertex in $\hat{U}(S_b)$ when uniformly selecting a vertex in U is at most $(1 - \epsilon)\rho_{\mathrm{U}}$, we have that

$$\Pr\left[\frac{|\hat{T}(S_b)|}{t} \geq (1 - \epsilon/2)\rho_{\mathrm{U}}\right] \leq \Pr\left[\frac{|\hat{T}(R)|}{t} \geq (1 - \epsilon/2)\rho_{\mathrm{U}}\right] \leq e^{-4m} \tag{8}$$

By the union bound, the probability that for any bad subset S_b we have $\frac{|\hat{T}(S_b)|}{t} \geq (1 - \epsilon/2)\rho_{\mathrm{U}}$ is at most $\binom{m}{\hat{m}} \cdot e^{-4m} < 2^m \cdot 2^{-4m}$, which for our choice of m is clearly strictly smaller than $1/20$ and the lemma follows. □

We next show that the estimated relative size of $\hat{G}(S_g)$ for any good subset S_g is close to its true relative size.

Lemma 6. *With probability at least* $19/20$, *for every good subset* S_g *of* X *of size* $\hat{m}$,

$$(1 - \epsilon/4)\frac{|\hat{U}(S_g)|}{|U|} \leq \frac{|\hat{T}(S_g)|}{t} \leq (1 + \epsilon/4)\frac{|\hat{U}(S_g)|}{|U|}.$$

Proof: For any fixed good subset S_g, by a multiplicative Chernoff bound, with probability at least $1 - \exp(-(1/2)t\frac{|\hat{U}(S_g)|}{|U|}(\epsilon/4)^2)$,

$$\frac{|\hat{T}(S_g)|}{t} \geq (1 - \epsilon/4)\frac{|\hat{U}(S_g)|}{|U|}. \tag{9}$$

Similarly, for any fixed good subset S_g, by a multiplicative Chernoff bound, with probability at least $1 - \exp(-(1/3)t\frac{|\hat{U}(S_g)|}{|U|}(\epsilon/4)^2)$,

$$\frac{|\hat{T}(S_g)|}{t} \leq (1+\epsilon/4)\frac{|\hat{U}(S_g)|}{|U|}. \tag{10}$$

Thus, by the union bound, the probability that for some good subset, S_g, the inequalities in the lemma do not hold is at most

$$\binom{m}{\hat{m}} \cdot \exp\left(-\frac{1}{2}t(1-\epsilon)\rho_{\mathrm{U}}(\epsilon/4)^2\right) + \binom{m}{\hat{m}} \exp\left(-\frac{1}{3}t(1-\epsilon)\rho_{\mathrm{U}}(\epsilon/4)^2\right).$$

For our choice of t and m, each of the above terms is clearly upper bounded by $1/40$. The lemma follows. □

By Lemmas 4–6, with probability at least $4/5 > 2/3$: (1) the algorithm comes across a good seed S^*; (2) no bad subset is a candidate for determining the output $\hat{W}(S^+)$, and in particular, $|\hat{U}(S^+)| \geq (1-\epsilon)\rho_{\mathrm{U}}|U|$; and (3) for all good subsets S_g, the inequalities in Lemma 6 hold. For the rest of the proof, assume that these events in fact occur.

First we show that, conditioned on the above events, the specific subset S^* will not be excluded in Step 4 of the algorithm. Observe that the subset S^* is good since by Lemma 3 $|\hat{U}(S^*)| \geq (1-\epsilon/4)|U^*| \geq (1-\epsilon/4)\rho_{\mathrm{U}}|U|$. Since S^* is good, we know that $\frac{|\hat{T}(S^*)|}{t} \geq (1-\epsilon/4)\frac{|\hat{U}(S^*)|}{|U|}$. Therefore,

$$\frac{|\hat{T}(S^*)|}{t} \geq (1-\epsilon/4)\frac{|\hat{U}(S^*)|}{|U|} \geq (1-\epsilon/4)(1-\epsilon/4)\rho_{\mathrm{U}} \geq (1-\epsilon/2)\rho_{\mathrm{U}},$$

and hence S^* will not be excluded in Step 4.

Next we show that the number of edges in $(\hat{U}(S^+), \hat{W}(S^+))$ is not much smaller than the number of edges in the maximum biclique (U^*, W^*).

$$\begin{aligned}
|E(\hat{U}(S^+), \hat{W}(S^+))| &\geq (1-\epsilon)\cdot|\hat{U}(S^+)|\cdot|\hat{W}(S^+)| && (11)\\
&\geq \frac{1-\epsilon}{1+\epsilon/4}\cdot\frac{|U|}{t}\cdot|\hat{T}(S^+)|\cdot|\hat{W}(S^+)| && (12)\\
&\geq \frac{1-\epsilon}{1+\epsilon/4}\cdot\frac{|U|}{t}\cdot|\hat{T}(S^*)|\cdot|\hat{W}(S^*)| && (13)\\
&\geq \frac{(1-\epsilon)(1-\epsilon/4)}{1+\epsilon/4}\cdot|\hat{U}(S^*)|\cdot|\hat{W}(S^*)| && (14)\\
&\geq (1-2\epsilon)\cdot|U^*|\cdot|W^*| && (15)
\end{aligned}$$

In the above sequence of equations, Eq. (11) follows from the definition of $\hat{U}(S^+) = \Gamma_\epsilon(\hat{W}(S^+))$; Eq. (12) follows from Eq. (10), which holds for every good subset (recall that by the second event we condition on, S^+ cannot be a bad subset); Eq. (13) holds by our choice of S^+ as the subset for which $|\hat{T}(S^+)|\cdot|\hat{W}(S^+)|$ is maximized and the fact that S^*

is considered by the algorithm since it is good; Eq. (14) follows from the fact that $\frac{|\hat{T}(S^*)|}{t} \geq (1-\epsilon/4)\frac{|\hat{U}(S^*)|}{|U|}$, and Eq. (15) follows from the fact that $|\hat{U}(S^*)| \cdot |\hat{W}(S^*)| \geq |U^*| \cdot |W^*|$ (recall that $(\hat{U}(S^*), \hat{W}(S^*))$ contains at least as many edges as (U^*, W^*) where the latter is a biclique), and from elementary mathematical manipulations.

The bound on the running time follows from the fact that we enumerate over all subsets of size $\hat{m}$ of the m vertices drawn in Step 1. The total number of such subsets is $\binom{m}{\hat{m}}$. For each subset S we compute $\Gamma(S)$ and $T \cap \Gamma_\epsilon(\Gamma(S))$. Thus for each subset S, the algorithm spends time $O(t \cdot |W|)$. Hence the total running time is

$$O(m^{\hat{m}}|W|t) = O\left(\left(\frac{1}{\rho_{\mathrm{U}}\epsilon^2}\log\frac{1}{\rho_{\mathrm{W}}\epsilon}\right)^{O\left(\frac{1}{\epsilon^2}\log\frac{1}{\rho_{\mathrm{W}}\epsilon}\right)}|W|\right).$$

4.2. *Conjunctive clustering*

Recall that given ρ_{U}, ρ_{W}, k, ϵ and δ, our goal is to output a collection $\tilde{\mathcal{C}}$ of at most k ϵ-bicliques that is δ-diverse and that $(b(\delta+\epsilon), b'\epsilon)$-swamps every biclique in $\mathcal{B}(\rho_{\mathrm{U}}, \rho_{\mathrm{W}})$ for small constants b and b'. Furthermore, either $|\tilde{\mathcal{C}}| = k$ or, for some δ-diverse collection $\mathcal{C}^* \subseteq \mathcal{B}(\rho_{\mathrm{U}}, \rho_{\mathrm{W}})$ of k bicliques, every biclique $(U_i^*, W_i^*) \in \mathcal{C}^*$ is $(\delta + b\epsilon)$-dominated by some ϵ-biclique in $\tilde{\mathcal{C}}$.

We reset $\hat{m}, m$ and t as follows: $\hat{m} = c_1 \cdot \frac{1}{\epsilon^2} \cdot \log\frac{k}{\rho_{\mathrm{W}}\cdot\epsilon}$, $m = c_2 \cdot \frac{\log k}{\rho_{\mathrm{U}}} \cdot \hat{m}$, and $t = c_3 \cdot \frac{\log(1/\epsilon)}{(\rho_{\mathrm{U}}\cdot\rho_{\mathrm{W}}\cdot\epsilon)^3} \cdot m$. Here c_1, c_2, and c_3 are constants that can be determined from the analysis.

Theorem 2. *With probability at least* 4/5, *Algorithm Conjunctive Clustering outputs a collection $\tilde{\mathcal{V}}$ of at most k subsets such that $\tilde{\mathcal{C}} = \{(\Gamma_\epsilon(\hat{W}), \hat{W}) : \hat{W} \in \tilde{\mathcal{V}}\}$ is δ-diverse, and $\tilde{\mathcal{C}}$ $((2\delta+4\epsilon), 3\epsilon)$-swamps every biclique in $\mathcal{B}(\rho_U, \rho_W)$. The collection $\tilde{\mathcal{C}}$ is either of size k or for some δ-diverse collection $\mathcal{C}^* \subseteq \mathcal{B}(\rho_U, \rho_W)$ of k bicliques, every biclique $(U_i^*, W_i^*) \in \mathcal{C}^*$ is $(\delta+4\epsilon)$-dominated by some ϵ-biclique in $\tilde{\mathcal{C}}$. Furthermore, for every $\hat{W} \in \tilde{\mathcal{V}}$, $|\hat{W}| \geq \rho_w|W|$ and $|\Gamma_\epsilon(\hat{W})| \geq (1-\epsilon)\rho_U|U|$. The running time of the algorithm is exponential in* poly$(1/\epsilon)$, *quasi-polynomial in k, $1/\rho_U$, and $1/\rho_W$, and linear in $|W|$.*

Before proving the theorem, we prove a couple of lemmas. For the sake of simplicity, in all that follows we use the term "with high probability" to mean with probability at least $1-\gamma$ for some sufficiently small constant γ. By selecting the constants c_1, c_2, and c_3 to be sufficiently large, we can ensure that the sum of all the γ's is at most 1/5 as required.

We shall use the same notations and terms as in the previous subsection. Specifically, for $S \subset X$, $\hat{U}(S) = \Gamma_\epsilon(\hat{W}(S))$ (so that $\hat{T}(S) = T \cap \hat{U}(S)$) and $\hat{G}(S) = (\hat{U}(S), \hat{W}(S))$. We say that a subset S is good if $|\hat{U}(S)|/|U| \geq (1-\epsilon)\rho_{\mathrm{U}}$; otherwise it is bad. Recall that by Lemmas 5 and 6, with high probability, for every $\hat{W}(S) \in \hat{\mathcal{V}}$, the subset S is good, and for every good subset S,

$$(1-\epsilon/4)\frac{|\hat{U}(S)|}{|U|} \leq \frac{|\hat{T}(S)|}{t} \leq (1+\epsilon/4)\frac{|\hat{U}(S)|}{|U|}. \tag{16}$$

Conjunctive Clustering Algorithm

1. Draw a sample X of m vertices uniformly and independently from U.
2. Draw an additional sample T of t vertices uniformly and independently from U.
3. Let $\hat{\mathcal{W}} \leftarrow \emptyset$.
4. For each subset S of X that has size $\hat{m}$ do
   (a) $\hat{W}(S) \leftarrow \Gamma(S)$.
   (b) $\hat{T}(S) \leftarrow T \cap \Gamma_\epsilon(\hat{W}(S))$.
   (c) If $|\hat{W}(S)| \geq \rho_{\mathrm{w}}|W|$ and $|\hat{T}(S)| \geq (1-\epsilon/2)\rho_{\mathrm{U}}t$ then add $\hat{W}(S)$ to $\hat{\mathcal{W}}$.
5. Order the subsets $\hat{W}(S)$ in $\hat{\mathcal{W}}$ according to the magnitude of $|\hat{T}(S)|\cdot|\hat{W}(S)|$. Perform the following at most k times: Add to $\tilde{\mathcal{W}}$ the next subset $\hat{W}(S)$ (according to the above order) such that $(\hat{T}(S), \hat{W}(S))$ is not yet $(\delta+2\epsilon)$-dominated by any $(\hat{T}(S'), \hat{W}(S'))$ where $\hat{W}(S') \in \tilde{\mathcal{W}}$.

Assume from this point on that these events in fact occur. The lemma below follows from Lemma 3 using similar arguments to those used to prove Theorem 1.

Lemma 7. *Let $\mathcal{C}^* = \{(U_i^*, W_i^*)\}_{i=1}^k$ be a fixed collection of bicliques in $\mathcal{B}(\rho_U, \rho_w)$. With high probability over the choice of the samples X and T, for every $(U_i^*, W_i^*) \in \mathcal{C}^*$, there exists a subset $\hat{W}(S) \in \hat{\mathcal{W}}$, such that $\hat{G}(S) = (\hat{U}(S), \hat{W}(S))$ $(\epsilon/4)$-dominates (U_i^*, W_i^*), and furthermore, $|E(\hat{U}(S), \hat{W}(S))| \geq |U_i^*| \cdot |W_i^*|$.*

Proof: For any fixed $(U_i^*, W_i^*) \in \mathcal{C}^*$, given the setting of the sample size and the size of the subsets, S, considered by the algorithm, with probability at least $1 - \frac{1}{ck}$ (where c is a large constant), there is a subset $S \subset X$ that is a good seed of U_i^*. This is proved using the same basic arguments as those used in the proof of Lemma 4. Let us denote this subset by S_i^*.

By Lemma 3 we know that $|\hat{U}(S_i^*)| \geq (1-\epsilon/4)|U_i^*|$, $\hat{G}(S_i^*) = (\hat{U}(S_i^*), \hat{W}(S_i^*))$ $(\epsilon/4)$-dominates (U_i^*, W_i^*) and $|E(\hat{U}(S_i^*), \hat{W}(S_i^*))| \geq |U_i^*| \cdot |W_i^*|$. Since $|\mathcal{C}^*| = k$, by a union bound, with probability at least $1 - 1/c$, each biclique in $\mathcal{C}^*$ is $(\epsilon/4)$-dominated by some ϵ-biclique $\hat{G}(S)$. Since for each such S_i^*, $|\hat{U}(S_i^*)| \geq (1-\epsilon/4)\rho_{\mathrm{U}}|U|$, by Eq. (16) $\hat{T}(S_i^*)$ is sufficiently large, and hence $\hat{W}(S_i^*)$ will be added to $\hat{\mathcal{W}}$ as required. □

For each subset S_i, we denote $T_i \stackrel{\text{def}}{=} \hat{T}(S_i)$ and $W_i \stackrel{\text{def}}{=} \hat{W}(S_i)$.

Lemma 8. *With high probability over the choice of T, for every pair of subsets S_1 and S_2 such that $W_1, W_2 \in \hat{\mathcal{W}}$,*

$$\left| \frac{|E(T_2, W_2) \backslash E(T_1, W_1)|}{t \cdot |W|} - \frac{|E(U_2, W_2) \backslash E(U_1, W_1)|}{|U| \cdot |W|} \right| \leq \frac{\epsilon \rho_U \rho_W}{c}, \tag{17}$$

and

$$\left| \frac{|E(T_2, W_2) \cup E(T_1, W_1)|}{t \cdot |W|} - \frac{|E(U_2, W_2) \cup E(U_1, W_1)|}{|U| \cdot |W|} \right| \leq \frac{\epsilon \rho_U \rho_W}{c}, \tag{18}$$

where c is a large constant.

Proof: We prove that Eq. (17) holds with high probability. The argument for Eq. (18) is proved analogously. Consider the edges in $E(\hat{U}(S_2), \hat{W}(S_2)) \backslash E(\hat{U}(S_1), \hat{W}(S_1))$. They consist of two disjoint sets of edges: (1) edges between $\hat{U}(S_2) \backslash \hat{U}(S_1)$ and $\hat{W}(S_2)$, and (2) edges between $\hat{U}(S_2) \cap \hat{U}(S_1)$ and $\hat{W}(S_2) \backslash \hat{W}(S_1)$. To simplify notation, let $U_{2\backslash 1} = \hat{U}(S_2) \backslash \hat{U}(S_1)$, $U_{2\cap 1} = \hat{U}(S_2) \cap \hat{U}(S_1)$, $W_2 = \hat{W}(S_2)$, and $W_{2\backslash 1} = \hat{W}(S_2) \backslash \hat{W}(S_1)$. Finally, let $T_{2\backslash 1} = T \cap U_{2\backslash 1}$, and $T_{2\cap 1} = T \cap U_{2\cap 1}$. We would like to show that $\frac{1}{t \cdot |W|} \cdot |E(T_{2\backslash 1}, W_2)|$ is a good estimate of $\frac{1}{|U| \cdot |W|} \cdot |E(U_{2\backslash 1}, W_2)|$, and that $\frac{1}{t \cdot |W|} \cdot |E(T_{2\cap 1}, W_2)|$ is a good estimate of $\frac{1}{|U| \cdot |W|} \cdot |E(U_{2\cap 1}, W_2)|$.

Consider first the edges between $U_{2\backslash 1}$ and W_2. Note that every vertex in $U_{2\backslash 1}$ neighbors at least $(1 - \epsilon)$ of the vertices in W_2. Hence, if we ensure that the number of vertices in $T_{2\backslash 1}$ is close to its expected value, then the number of edges between $T_{2\backslash 1}$ and W_2 is close to its expected value. Specifically, by applying an additive Chernoff bound we can ensure that with high probability, for every good S_1 and S_2,

$$\left| \frac{|T_{2\backslash 1}|}{t} - \frac{|U_{2\backslash 1}|}{|U|} \right| \leq \frac{\epsilon \rho_U \rho_W}{c'}, \tag{19}$$

for some sufficiently large constant c'. Using the fact that every vertex in $U_{2\backslash 1}$ (and hence in $T_{2\backslash 1}$), neighbors at least $(1 - \epsilon)$ of the vertices in W_2 (and at most all vertices in W_2), we can derive that

$$\left| \frac{|E(T_{2\backslash 1}, W_2)|}{t \cdot |W|} - \frac{|E(U_{2\backslash 1}, W_2)|}{|U| \cdot |W|} \right| \leq \frac{\epsilon \rho_U \rho_W}{2c} \tag{20}$$

where c is the constant stated in the lemma.

Consider next the edges between $U_{2\cap 1}$ and $W_{2\backslash 1}$. Since the vertices in $U_{2\cap 1}$ may vary in the number of neighbors they have in $W_{2\backslash 1}$, we need to do a slightly more careful analysis. Let $\epsilon' = \rho_U \rho_W \epsilon / 6c$. Suppose we partition the vertices in $U_{2\cap 1}$ into $M = \lceil 1/\epsilon' \rceil$ subsets such that in each subset all vertices have roughly the same number of neighbors in $W_{2\backslash 1}$. More precisely, for $j = 1, \ldots, M$, let $U^j_{2\cap 1} \subseteq U_{2\cap 1}$ consist of all vertices in $U_{2\cap 1}$ that neighbor between $(j-1) \cdot \epsilon'$ of the vertices in $W_{2\backslash 1}$, and $j \cdot \epsilon'$ of these vertices. Let $T^j_{2\cap 1} = T \cap U^j_{2\cap 1}$. By a Chernoff bound we can ensure that, with high probability, for every $U^j_{2\cap 1}$ such that $\frac{|U^j_{2\cap 1}|}{|U|} \geq (\epsilon')^2 / c'$ (where c' is a sufficiently large constant),

$$(1 - \epsilon') \cdot \frac{\left|U^j_{2\cap 1}\right|}{|U|} \leq \frac{\left|T^j_{2\cap 1}\right|}{t} \leq (1 + \epsilon') \cdot \frac{\left|U^j_{2\cap 1}\right|}{|U|}$$

while for every $U^j_{2\cap 1}$ such that $\frac{|U^j_{2\cap 1}|}{|U|} < (\epsilon')^2/c'$, we have $\frac{|T^j_{2\cap 1}|}{t} < 2(\epsilon')^2/c'$. Assuming that these inequalities hold, we have that

$$\begin{aligned}
&\frac{|E(T_{2\cap 1}, W_{2\setminus 1})|}{t \cdot |W|} \\
&\leq \frac{\sum_{j=1}^{M} \left|T^j_{2\cap 1}\right| \cdot |W_{2\setminus 1}| \cdot j \cdot \epsilon'}{t \cdot |W|} && (21)\\
&\leq \frac{(1+\epsilon') \cdot \sum_{j=1}^{M} \left|U^j_{2\cap 1}\right| \cdot |W_{2\setminus 1}| \cdot j \cdot \epsilon'}{|U| \cdot |W|} + M \cdot \frac{2(\epsilon')^2}{c'} \cdot \frac{|W_{2\setminus 1}|}{|W|} && (22)\\
&\leq \frac{\sum_{j=1}^{M} \left|U^j_{2\cap 1}\right| \cdot |W_{2\setminus 1}| \cdot (j-1) \cdot \epsilon'}{|U| \cdot |W|} + \frac{\sum_{j=1}^{M} \left|U^j_{2\cap 1}\right| \cdot |W_{2\setminus 1}| \cdot \epsilon'}{|U| \cdot |W|} \\
&\quad + \frac{\epsilon' \cdot |W_{2\setminus 1}|}{|W|} \cdot \frac{\sum_{j=1}^{M} \left|U^j_{2\cap 1}\right| \cdot j \cdot \epsilon'}{|U|} + \frac{2 \cdot M \cdot (\epsilon')^2}{c'} && (23)\\
&\leq \frac{|E(U_{2\cap 1}, W_{2\setminus 1})|}{|U| \cdot |W|} + \epsilon' \left(1 + M \cdot \epsilon' + \frac{2 \cdot M \cdot \epsilon'}{c'}\right) && (24)\\
&\leq \frac{|E(U_{2\cap 1}, W_{2\setminus 1})|}{|U| \cdot |W|} + 3\epsilon' && (25)\\
&= \frac{|E(U_{2\cap 1}, W_{2\setminus 1})|}{|U| \cdot |W|} + \frac{\epsilon \rho_{\mathrm{U}} \rho_{\mathrm{W}}}{2c}. && (26)
\end{aligned}$$

Similarly,

$$\frac{|E(T_{2\cap 1}, W_{2\setminus 1})|}{t \cdot |W|} \geq \frac{|E(U_{2\cap 1}, W_{2\setminus 1})|}{|U| \cdot |W|} - \frac{\epsilon \rho_{\mathrm{U}} \rho_{\mathrm{W}}}{2c}. \quad (27)$$

The lemma follows from Eqs. (20), (26) and (27). □

As a corollary of Lemma 8:

Corollary 9. *With high probability over the choice of T, for every pair of good subsets S_1 and S_2 that are considered by the algorithm in Step 5, if $\hat{G}(S_1)$ α-dominates $\hat{G}(S_2)$ then $\hat{G}_T(S_1)$ $(\alpha+\epsilon)$-dominates $\hat{G}_T(S_2)$, and if $\hat{G}(S_1)$ does not α-dominate $\hat{G}(S_2)$ then $\hat{G}_T(S_1)$ does not $(\alpha-\epsilon)$-dominate $\hat{G}_T(S_2)$.*

Proof: As before, let $T_i \stackrel{\text{def}}{=} \hat{T}(S_i)$, $U_i \stackrel{\text{def}}{=} \hat{U}(S_i)$, and $W_i \stackrel{\text{def}}{=} \hat{W}(S_i)$, Suppose that $\hat{G}(S_1)$ α-dominates $\hat{G}(S_2)$, that is,

$$\frac{|E(U_2, W_2) \setminus E(U_1, W_1)|}{|E(U_2, W_2) \cup E(U_1, W_1)|} \leq \alpha.$$

Then, by Lemma 8

$$\begin{aligned}
&\frac{|E(T_2, W_2)\backslash E(T_1, W_1)|}{|E(T_2, W_2) \cup E(T_1, W_1)|} \\
&\quad = \frac{|E(T_2, W_2)\backslash E(T_1, W_1)|/(t \cdot |W|)}{|E(T_2, W_2) \cup E(T_1, W_1)|/(t \cdot |W|)} \\
&\quad \leq \frac{|E(U_2, W_2)\backslash E(U_1, W_1)|/(|U| \cdot |W|) + \epsilon\rho_{\mathrm{U}}\rho_{\mathrm{W}}/c}{|E(U_2, W_2) \cup E(U_1, W_1)|/(|U| \cdot |W|) - \epsilon\rho_{\mathrm{U}}\rho_{\mathrm{W}}/c} \\
&\quad \leq \frac{|E(U_2), W_2)\backslash E(U_1, W_1)|/(|U| \cdot |W|) + \epsilon\rho_{\mathrm{U}}\rho_{\mathrm{W}}/c}{\left(1 - \frac{4\epsilon}{c}\right) \cdot |E(U_2, W_2) \cup E(U_1, W_1)|/(|U| \cdot |W|)} \qquad (28) \\
&\quad \leq \alpha + \epsilon. \qquad (29)
\end{aligned}$$

Equation (28) follows from the fact that

$$\frac{|E(\hat{U}(S_2), \hat{W}(S_2)) \cup E(\hat{U}(S_1), \hat{W}(S_1))|}{|U| \cdot |W|} \geq \rho_{\mathrm{U}}\rho_{\mathrm{W}}/4 \qquad (30)$$

which is true because $\hat{W}(S_2), \hat{W}(S_1) \in \hat{\mathcal{W}}$, and S_1 and S_2 are good. Equation (29) follows for an appropriate setting of the constant c.

By a similar argument, we establish the second part of the corollary. □

Proof of Theorem 2: Recall that $\tilde{\mathcal{C}}$ is the set of all bisubgraphs $\hat{G}(S) = (\hat{U}(S), \hat{W}(S))$ such that $\hat{W}(S) \in \tilde{\mathcal{W}}$. In what follows, we assume that the events that are proved to hold with high probability in Lemma 7 and Corollary 9, in fact hold. When we say that we apply Lemma 7 or Corollary 9, then we mean that we rely on the events stated in them.

We first prove that $\tilde{\mathcal{C}}$ is δ-diverse. Recall that for any subset S, $\hat{G}_T(S) = (\hat{T}(S), \hat{W}(S))$ denotes the bisubgraph that is actually considered by the algorithm in Step 5. Consider any iteration in Step 5 of the algorithm, where we add to $\tilde{\mathcal{W}}$ a new subset $\hat{W}(S_{\mathrm{new}})$. Let $S_{\mathrm{prev}} \subset X$ be any subset such that $\hat{W}(S_{\mathrm{prev}})$ already belongs to $\tilde{\mathcal{W}}$. We need to show that: (1) $\hat{G}(S_{\mathrm{prev}})$ does not δ-dominate $\hat{G}(S_{\mathrm{new}})$, and (2) $\hat{G}(S_{\mathrm{new}})$ does not δ-dominate $\hat{G}(S_{\mathrm{prev}})$. By definition of Step 5 of the algorithm we know that $\hat{G}_T(S_{\mathrm{prev}})$ does not $(\delta + 2\epsilon)$-dominate $\hat{G}_T(S_{\mathrm{new}})$. By Corollary 9, we have that $\hat{G}(S_{\mathrm{prev}})$ does not $(\delta + \epsilon)$-dominate $\hat{G}(S_{\mathrm{new}})$—and hence does not δ-dominate $\hat{G}(S_{\mathrm{new}})$.

We now turn to showing that $\hat{G}(S_{\mathrm{new}})$ does not δ-dominate $\hat{G}(S_{\mathrm{prev}})$. Note first that if $\hat{G}(S_{\mathrm{prev}})$ contains at least as many edges as $\hat{G}(S_{\mathrm{new}})$, then we are done. Specifically, in this case:

$$\begin{aligned}
&\frac{|E(\hat{U}(S_{\mathrm{prev}}), \hat{W}(S_{\mathrm{prev}}))\backslash E(\hat{U}(S_{\mathrm{new}}), \hat{W}(S_{\mathrm{new}}))|}{|E(\hat{U}(S_{\mathrm{prev}}), \hat{W}(S_{\mathrm{prev}})) \cup E(\hat{U}(S_{\mathrm{new}}), \hat{W}(S_{\mathrm{new}}))|} \\
&\quad \geq \frac{|E(\hat{U}(S_{\mathrm{new}}), \hat{W}(S_{\mathrm{new}}))\backslash E(\hat{U}(S_{\mathrm{prev}}), \hat{W}(S_{\mathrm{prev}}))|}{|E(\hat{U}(S_{\mathrm{prev}}), \hat{W}(S_{\mathrm{prev}})) \cup E(\hat{U}(S_{\mathrm{new}}), \hat{W}(S_{\mathrm{new}}))|} \\
&\quad > (\delta + 2\epsilon) > \delta. \qquad (31)
\end{aligned}$$

In the above equation we used the fact that for two sets A and B if $|A| \geq |B|$ then $|A \backslash B| \geq |B \backslash A|$. However, while we cannot be certain that $\hat{G}(S_{\text{prev}})$ contains at least as many edges as $\hat{G}(S_{\text{new}})$, we can show that the number of edges in $\hat{G}(S_{\text{prev}})$ cannot be much smaller than the number of edges in $\hat{G}(S_{\text{new}})$. This is due to the fact that the bisubgraphs are sorted according to a related size measure in Step 5. Specifically, using Eq. (16) and the fact that $\hat{G}(S_{\text{new}})$ and $\hat{G}(S_{\text{prev}})$ are ϵ-bicliques, we have

$$
\begin{aligned}
&|E(\hat{U}(S_{\text{prev}}), \hat{W}(S_{\text{prev}}))| \\
&\quad \geq (1-\epsilon) \cdot |\hat{U}(S_{\text{prev}})| \cdot |\hat{W}(S_{\text{prev}})| && (32)\\
&\quad \geq \frac{1-\epsilon}{1+\epsilon/4} \cdot \frac{|U|}{t} \cdot |\hat{T}(S_{\text{prev}})| \cdot |\hat{W}(S_{\text{prev}})| && (33)\\
&\quad \geq \frac{1-\epsilon}{1+\epsilon/4} \cdot \frac{|U|}{t} \cdot |\hat{T}(S_{\text{new}})| \cdot |\hat{W}(S_{\text{new}})| && (34)\\
&\quad \geq \frac{(1-\epsilon)(1-\epsilon/4)}{1+\epsilon/4} \cdot |\hat{U}(S_{\text{new}})| \cdot |\hat{W}(S_{\text{new}})| && (35)\\
&\quad \geq (1-2\epsilon) \cdot |E(\hat{U}(S_{\text{new}}), \hat{W}(S_{\text{new}}))|. && (36)
\end{aligned}
$$

In this case, we can modify Eq. (31) using the fact that for any two sets, A and B, such that $|A| \geq |B|(1-2\epsilon)$ and $|B \backslash A|/|A \cup B| > \delta + 2\epsilon$, we have $|A \backslash B|/|A \cup B| > \delta$ (proved in the appendix in Lemma 13). We will then have that

$$
\frac{|E(\hat{U}(S_{\text{prev}}), \hat{W}(S_{\text{prev}})) \backslash E(\hat{U}(S_{\text{new}}), \hat{W}(S_{\text{new}}))|}{|E(\hat{U}(S_{\text{prev}}), \hat{W}(S_{\text{prev}})) \cup E(\hat{U}(S_{\text{new}}), \hat{W}(S_{\text{new}}))|} > \delta. \tag{37}
$$

Consequently the bisubgraph $\hat{G}(S_{\text{new}})$ does not δ-dominate $\hat{G}(S_{\text{prev}})$, implying that $\tilde{\mathcal{C}}$ is δ-diverse, as required.

We next show that $\tilde{\mathcal{C}}$ $((2\delta + 4\epsilon), 3\epsilon)$-swamps every biclique in $\mathcal{B}(\rho_{\text{U}}, \rho_{\text{W}})$. Let $\mathcal{C}^* = \{(U_i^*, W_i^*)\}_{i=1}^k$ be a collection of k bicliques that $(\delta, 0)$-swamps every biclique in $\mathcal{B}(\rho_{\text{U}}, \rho_{\text{W}})$. The existence of such a collection follows from a greedy procedure that constructs such a collection in a manner that is similar to the last step of our algorithm. Namely, it orders the bicliques in $\mathcal{B}(\rho_{\text{U}}, \rho_{\text{W}})$ according to their size, and in each of the k iterations, it takes into $\mathcal{C}^*$ the next biclique that is not δ-dominated by any previously added biclique. (If this process ends before k bicliques are selected, then this collection actually has the stronger property that it δ-dominates every biclique in $\mathcal{B}(\rho_{\text{U}}, \rho_{\text{W}})$, and the rest of our argument is still applicable.)

Let $\hat{\mathcal{C}} = \{(\hat{U}(S), \hat{W}(S)) : \hat{W}(S) \in \hat{\mathcal{W}}\}$ be the collection of all ϵ-bicliques that may be considered in Step 5 of the algorithm. By Lemma 7, with high probability, for each $(U_i^*, W_i^*) \in \mathcal{C}^*$, there exists a subset $S_i^* \subset X$ for which $\hat{W}(S_i^*) \in \hat{\mathcal{W}}$ and such that $(\hat{U}(S_i^*), \hat{W}(S_i^*))$ $(\epsilon/4)$-dominates (U_i^*, W_i^*) and $|E(\hat{U}(S_i^*), \hat{W}(S_i^*))| \geq |E(U_i^*, W_i^*)|$.

We need to consider two cases. In the first case, there exists a subset $\tilde{S}_i \subset X$ such that $\hat{G}_T(\tilde{S}_i)$ $(\delta + 2\epsilon)$-dominates $\hat{G}_T(S_i^*)$. By Corollary 9 we have that $\hat{G}(\tilde{S}_i)$ $(\delta + 3\epsilon)$-dominates $\hat{G}(S_i^*)$. By applying a triangle-like inequality (proved in the appendix in Lemma 14), we get that $\hat{G}(\tilde{S}_i)$ $(\delta + 4\epsilon)$-dominates (U_i^*, W_i^*).

In the second case, $|\hat{T}(S_i^*)| \cdot |\hat{W}(S_i^*)| \leq |\hat{T}(\tilde{S})| \cdot |\hat{W}(\tilde{S})|$ for all $\tilde{S}$ such that $\hat{W}(\tilde{S}) \in \tilde{\mathcal{W}}$. Hence

$$|E(U_i^*, W_i^*)| \leq |E(\hat{U}(S_i^*), \hat{W}(S_i^*))| \tag{38}$$

$$\leq |U_i^*| \cdot |W_i^*| \tag{39}$$

$$\leq \frac{1+\epsilon/4}{(1-\epsilon/4)(1-\epsilon)} \cdot |E(\hat{U}(\tilde{S}), \hat{W}(\tilde{S}))| \tag{40}$$

$$\leq (1+3\epsilon) \cdot |E(\hat{U}(\tilde{S}), \hat{W}(\tilde{S}))|, \tag{41}$$

where we have used Eq. (16) and the fact that the bisubgraphs considered are ϵ-bicliques. Recall that $\mathcal{C}^*$ $(\delta, 0)$-swamps every biclique in $\mathcal{B}(\rho_{\mathrm{U}}, \rho_{\mathrm{W}})$. Hence, by applying Lemma 14 once again, we get that $\tilde{\mathcal{C}}$ $((2\delta + 4\epsilon), 3\epsilon)$-swamps every such biclique.

We next show that the collection $\tilde{\mathcal{C}}$ is either of size k or, for some δ-diverse collection $\mathcal{C}^*$ of k true bicliques (if such exists), each $G_i^* = (U_i^*, W_i^*)$ in $\mathcal{C}^*$ is $(\delta + 4\epsilon)$-dominated by some ϵ-biclique in $\tilde{\mathcal{C}}$. Consider any such fixed δ-diverse collection $\mathcal{C}^* = \{(U_i^*, W_i^*)\}_{i=1}^k$. (In particular, this may be the same collection considered above, but not necessarily). Once again let $\hat{\mathcal{C}} = \{(\hat{U}(S), \hat{W}(S)) : \hat{W}(S) \in \hat{\mathcal{W}}\}$ be the collection of all ϵ-bicliques that may be considered in Step 5 of the algorithm. By Lemma 7, with high probability, for each $(U_i^*, W_i^*) \in \mathcal{C}^*$, there exists a subset $S_i^* \subset X$ for which $\hat{W}(S_i^*) \in \hat{\mathcal{W}}$ and such that $(\hat{U}(S_i^*), \hat{W}(S_i^*))$ $(\epsilon/4)$-dominates (U_i^*, W_i^*).

Suppose that the algorithm selects less then k subsets from $\hat{\mathcal{W}}$. Then it must be the case that for each S_i^* as above, either $\hat{W}(S_i^*) \in \tilde{\mathcal{W}}$, or there exists some subset $\tilde{S}_i \subset X$ such that $\hat{G}_T(\tilde{S}_i)$ $(\delta + 2\epsilon)$-dominates $\hat{G}_T(S_i^*)$. By Corollary 9 we have that $\hat{G}(\tilde{S}_i)$ $(\delta + 3\epsilon)$-dominates $\hat{G}(S_i^*)$. By applying a triangle-like inequality (proved in the appendix in Lemma 14), we get that $\hat{G}(\tilde{S}_i)$ $(\delta + 4\epsilon)$-dominates (U_i^*, W_i^*).

The lower bound on the size of $\hat{W}$ for each $\hat{W} \in \hat{\mathcal{W}}$ follows by definition of the algorithm, and the lower bound on the size of $\hat{U} = \Gamma_\epsilon(\hat{W})$ holds because the algorithm considers only good subsets (with high probability). □

In the case that there is no δ-diverse collection of k bicliques in $\mathcal{B}(\rho_{\mathrm{U}}, \rho_{\mathrm{W}})$ and $|\tilde{\mathcal{C}}| < k$ then the proof of Theorem 2 implies that (with high confidence) that *every* biclique in $\mathcal{B}(\rho_{\mathrm{U}}, \rho_{\mathrm{W}})$ is $(2\delta + 4\epsilon)$-dominated by some ϵ-biclique in $\tilde{\mathcal{C}}$.

5. Finding approximate ϵ-bicliques

In Section 4.1, we showed that if the graph contains a large biclique (U^*, W^*), then we can find a subset $\hat{W}$ such that $|E(\Gamma_\epsilon(\hat{W}), \hat{W})| \geq (1 - 2\epsilon) \cdot |U^*| \cdot |W^*|$. Suppose we know that there exists an ϵ-biclique $(U^\star, W^\star)$ such that $|U^\star| \geq \rho_{\mathrm{U}} \cdot |U|$ and $|W^\star| \geq \rho_{\mathrm{W}} \cdot |U|$, but there is not necessarily such a large biclique. We next show how the Approximate Maximum Biclique algorithm can be modified so as to obtain a $4\epsilon^{1/3}$-biclique $(\hat{U}, \hat{W})$ such that $|E(\hat{U}, \hat{W})| \geq (1 - 5\epsilon^{1/3})|E(U^\star, W^\star)|$. The extension of finding a collection of large $O(\epsilon^{1/3})$-bicliques can be done analogously to what is described in Section 4.2.

Let $\hat{m} = c_1 \cdot \frac{1}{\epsilon^2} \cdot \log \frac{1}{\rho_W \cdot \epsilon}$, $m = c_2 \cdot \frac{1}{\rho_U} \cdot \hat{m}$, and $t = c_3 \cdot \frac{1}{\rho_U \cdot \epsilon^2} \cdot m$. Here c_1, c_2, and c_3 are constants that can be determined from the analysis, and we assume that $\rho_U|U|$ and $\rho_W|W|$ are lower bounds on the sizes of $U^\star$ and $W^\star$, respectively.

Algorithm Approximate Maximum ϵ-Biclique

1. Draw a sample X of m vertices uniformly and independently from U.
2. Draw an additional sample T of t vertices uniformly and independently from U.
3. For each subset S of X that has size $\hat{m}$ do:
   (a) $\hat{W}(S) \leftarrow \Gamma_{2\epsilon^{2/3}}(S)$
   (b) $\hat{T}(S) \leftarrow T \cap \Gamma_{4\epsilon^{1/3}}(\hat{W}(S))$.
4. Among all subsets S considered by the algorithm for which $|\hat{W}(S)| \geq (1-2\epsilon^{1/3})\rho_W|W|$ and $|\hat{T}(S)| \geq (1-3\epsilon^{1/3})\rho_U t$, let S^+ be such that $|\hat{T}(S^+)| \cdot |\hat{W}(S^+)|$ is maximized. Output $\hat{W}(S^+)$.

Theorem 3. *Let $\tilde{\rho} \cdot |U| \cdot |W|$ be the size of a maximum ϵ-biclique whose sides are of size at least $\rho_U|U|$ and $\rho_W|W|$, respectively. With probability at least $2/3$, Algorithm Approximate Maximum ϵ-biclique outputs a subset $\hat{W} = \hat{W}(S^+)$ such that*

$$|E(\Gamma_{4\epsilon^{1/3}}(\hat{W}), \hat{W})| \geq (1 - 5\epsilon^{1/3}) \cdot \tilde{\rho} \cdot |U| \cdot |W|,$$

where $|\hat{W}| \geq (1 - 2\epsilon^{1/3})\rho_W|W|$ and $|\Gamma_{4\epsilon^{1/3}}(\hat{W})| \geq (1 - 4\epsilon^{1/3})\rho_U|U|$.

In order to prove the theorem, we first modify the definition of high-degree and low-degree vertices and the definition of a good seed. Here $(U^\star, W^\star)$ is a fixed maximum ϵ-biclique (satisfying $|U^\star| \geq \rho_U \cdot |U|$ and $|W^\star| \geq \rho_W \cdot |W|$). Recall that for a subset S, $\hat{W}(S) = \Gamma_{2\epsilon^{2/3}}(S)$.

Definition 8. We say that a vertex $w \in W$ has high degree with respect to $U^\star$ if $\frac{|\Gamma(w) \cap U^\star|}{|U^\star|} \geq 1 - 3\epsilon^{2/3}$. Otherwise it has low degree with respect to $U^\star$.

Using a calculation similar to the one in Eq. (4), it can be verified that since $(U^\star, W^\star)$ is an ϵ-biclique, all but at most an $\epsilon^{1/3}$-fraction of the vertices in $W^\star$ neighbor at least $(1 - \epsilon^{2/3})$ of the vertices in $U^\star$, and hence have high degree with respect to $U^\star$.

Definition 9. We say that a subset $S \subseteq U^\star$ is a good seed of $U^\star$ if the following two conditions hold: (1) The number of vertices w in $W^\star$ such that $w \notin \hat{W}(S)$ is at most $2\epsilon^{1/3}|W^\star|$, and (2) The number of vertices in $\hat{W}(S)$ that have low degree with respect to $U^\star$ is at most $\epsilon^{1/3}|W^\star|$.

By the above definition, S is a good seed of $U^\star$ if $\hat{W}(S)$ contains almost all vertices from $W^\star$ and almost no vertices that have low degree with respect to $U^\star$.

Lemma 10. *With probability at least* 19/20, *a uniformly selected sample of $\hat{m}$ vertices from $U^\star$ is a good seed of $U^\star$.*

Lemma 10 is proved similarly to Lemma 2. As in the proof of Lemma 2, we need to upper bound the number of low-degree vertices that have relatively many neighbors in S, and hence, are included in $\hat{W}(S)$. In addition we need to upper bound the number of vertices in $W^\star$ that neighbor at least $(1-\epsilon^{2/3})$ of the vertices in $U^\star$—but have relatively few neighbors in S—and hence, are not included in $\hat{W}(S)$.

Given the size of the sample m we obtain as a corollary:

Corollary 11. *With probability at least* 2/3, *the sample selected in Step* 1 *of the algorithm contains a subset S that is a good seed of $U^\star$.*

Lemma 12. *If S is a good seed of $U^\star$ and $\hat{W}(S) = \Gamma_{2\epsilon^{2/3}}(S)$, then*

$$|E(\Gamma_{4\epsilon^{1/3}}(\hat{W}(S)), \hat{W}(S))| \geq (1-4\epsilon^{1/3})|E(U^\star, W^\star)|.$$

Proof: The proof is similar to the proof of Lemma 3, but is actually simpler since we prove a somewhat weaker claim (which cannot be significantly strengthened). For the sake of simplicity, let $\hat{W} = \hat{W}(S)$, and let $\hat{U}$ denote $\Gamma_{4\epsilon^{1/3}}(\hat{W}(S))$. The set $\hat{W}$ consists of three subsets: (1) the subset of vertices in $W^\star \cap \hat{W}$; (2) the subset of vertices, denoted H, that do not belong to $W^\star$ and have high degree with respect to $U^\star$; and (3) the subset of vertices, denoted L, that do not belong to $W^\star$ and have low degree with respect to $U^\star$. By the premise of the lemma, that S is a good seed, we know that $|W^\star \cap \hat{W}| \geq (1-2\epsilon^{1/3})|W^\star|$ and $|L| \leq \epsilon^{1/3}|W^\star|$.

Let $Q^\star \subseteq U^\star$ denote the subset of vertices in $U^\star$ that have at least $(1-\sqrt{3}\epsilon^{1/3})|H|$ neighbors in H. It can be verified, using a calculation similar to the one in Eq. (4), that by definition of H, $|Q^\star| \geq (1-\sqrt{3}\epsilon^{1/3})|U^\star|$.

Since $(U^\star, W^\star)$ is an ϵ-biclique, every vertex in $Q^\star \subseteq U^\star$ has at least $(1-\epsilon)|W^\star|$ neighbors in $W^\star$. Since $\hat{W}$ contains all but at most $2\epsilon^{1/3}|W^\star|$ of the vertices in $W^\star$, every vertex in $Q^\star$ has at least $(1-(\epsilon+2\epsilon^{1/3}))|W^\star| \geq (1-3\epsilon^{1/3})|W^\star \cap \hat{W}|$ neighbors in $W^\star \cap \hat{W}$. Combining this with what we have shown above concerning the number of neighbors that every vertex in $Q^\star$ has in H, we get that every vertex in $Q^\star$ has at least

$$(1-3\epsilon^{1/3})|W^\star \cap \hat{W}| + (1-\sqrt{3}\epsilon^{1/3})|H|$$

neighbors in $\hat{W}$.

Since $|L| \leq \epsilon^{1/3}|W^\star|$, $|W^\star \cap \hat{W}| \geq (1-2\epsilon^{1/3})|W^\star|$ and $|\hat{W}| = |L|+|H|+|W^\star \cap \hat{W}|$, it can be verified that every vertex in $Q^\star$ has at least $(1-4\epsilon^{1/3})|\hat{W}|$ neighbors in $\hat{W}$; it follows that $Q^\star \subseteq \hat{U}$. We have shown that $|Q^\star| \geq (1-\sqrt{3}\epsilon^{1/3})|U^\star|$ (where $Q^\star \subseteq U^\star$), and we know that $\hat{W}$ contains all but at most $2\epsilon^{1/3}$ of the vertices in $W^\star$ (since S is a good seed). Therefore, $E(\hat{U}, \hat{W})$ contains all edges in $E(U^\star, W^\star)$ but at most $(\sqrt{3}\epsilon^{1/3} + 2\epsilon^{1/3}) \cdot |U^\star| \cdot |W^\star|$. The lemma follows. □

Theorem 3 follows from Lemma 12 in a similar fashion to the way Theorem 1 follows from Lemma 3.

6. An algorithm that is sublinear in $|U|$ and in $|W|$

In this section we present an algorithm that given ρ_U and ρ_W for which $\mathcal{B}(\rho_U, \rho_W)$ is non-empty, outputs an *implicit representation* of a bisubgraph $(\hat{U}, \hat{W})$ such that with high probability, $|E(\hat{U}, \hat{W})| \geq (1 - 2\epsilon) \cdot |\hat{U}| \cdot |\hat{W}|$, and $|E(\hat{U}, \hat{W})|$ is almost as large as the size of a maximum biclique in $\mathcal{B}(\rho_U, \rho_W)$. Note that $(\hat{U}, \hat{W})$ is not necessarily an ϵ-biclique, or even a 2ϵ-biclique, but rather it is a very dense bisubgraph. By an implicit representation of a bisubgraph $(\hat{U}, \hat{W})$, we mean a pair (Z, Y), where $Z \subseteq U$, $Y \subseteq \Gamma(Z) \subseteq W$, such that $\hat{W} = \Gamma(Z)$ and $\hat{U} = \Gamma_{2\epsilon}(Y)$. Both subsets Z and Y are of size polynomial in $1/\epsilon$, $1/\rho_U$, $1/\rho_W$, and independent of $|U|$ and $|W|$. The total running time of the algorithm that finds (Z, Y) is logarithmic in both $|U|$ and $|W|$.

This algorithm is a variant of the Approximate Maximum Biclique algorithm. The only difference is that in order to reduce the dependency that algorithm has on $|W|$, we sample from W instead of considering all of W (or, more precisely, all the neighbors that vertices in S have in W) in the third step of the algorithm. Analgous variants can be obtained for the other algorithms as well.

We note that reducing the dependence on $|W|$ comes at two costs. First, the pair $(\hat{U}, \hat{W})$ that is implicitly determined by the output of the algorithm is not necessary on $O(\epsilon)$-biclique, that is, it is not ensured that every vertex in $\hat{U}$ neighbors almost all of $\hat{W}$. Rather, $(\hat{U}, \hat{W})$ is a very dense subgraph, so that almost every vertex in $\hat{U}$ neighbors almost all of $\hat{W}$. Secondly, in the context of clustering, the output of the algorithm is less natural than the one in the Maximum Biclique Algorithm since no conjunctive description is output. Nonetheless, in the more general context of sublinear algorithms for graph theoretic problems, we obtain a more efficient algorithm.

Let $\hat{m}$ be as defined in the Good Seed Algorithm. Let $m = \frac{c_1}{\rho_U} \cdot \hat{m}$, $q = \frac{c_2 \cdot \log \frac{1}{\rho_U \rho_W \epsilon}}{\epsilon^2 \cdot \rho_W} \cdot m$ and $t = \frac{c_3}{\rho_U \cdot \rho_W \cdot \epsilon^2} \cdot m$. Here c_1, c_2, and c_3 are constants that can be determined from the analysis.

Algorithm Approximate Maximum Implicit Biclique

1. Draw the following samples:
   - X and T are samples of m and t vertices, respectively, selected uniformly and independently from U.
   - Y is a sample of q vertices selected uniformly and independently from W.
2. For each subset S of X that has size $\hat{m}$ do:
   (a) $\hat{Y}(S) \leftarrow Y \cap \Gamma(S)$.
   (b) $\hat{T}(S, Y) \leftarrow T \cap \Gamma_{2\epsilon}(\hat{Y}(S))$.
3. Among all subsets S, let Z be such that $|\hat{Y}(Z)| \geq (1 - \epsilon)\rho_W|Y|$, $\hat{T}(S, Y) \geq (1 - \epsilon)\rho_U|T|$, and $|\hat{T}(Z, Y)| \cdot |\hat{Y}(Z)|$ is maximized. Output the pair $(Z, \hat{Y}(Z))$.

Theorem 4. *Let $\rho^* \cdot |U| \cdot |W|$ be the size of a maximum biclique in $\mathcal{B}(\rho_U, \rho_W)$. With probability at least $2/3$, Algorithm Approximate Maximum Implicit Biclique outputs a pair of subsets $(Z, \hat{Y}(Z))$ so that*

$$|E(\Gamma_{2\epsilon}(\hat{Y}(Z)), \Gamma(Z))| \geq (1 - 4\epsilon) \cdot |\Gamma_{2\epsilon}(\hat{Y}(Z))| \cdot |\Gamma(Z)|$$

and

$$|E(\Gamma_{2\epsilon}(\hat{Y}(Z)), \Gamma(Z))| \geq (1 - 6\epsilon) \cdot \rho^* \cdot |U| \cdot |W|,$$

where $|\Gamma(Z)| \geq (1 - 2\epsilon)\rho_W|W|$ and $|\Gamma_{2\epsilon}(\hat{Y}(Z))| \geq (1 - 2\epsilon)\rho_U|U|$. The running time of the algorithm is exponential in poly$(1/\epsilon)$, *quasi-polynomial in $\frac{1}{\rho_U}$ and $\frac{1}{\rho_W}$, and logarithmic in $|U|$ and $|W|$.*

Proof: For the sake of simplicity, in all that follows we use the term "with high probability" to mean with probability at least $1 - \gamma$ for some sufficiently small constant γ. By selecting the constants c_1, c_2, and c_3 that determine the sample sizes to be sufficiently large, we can ensure that the sum of all the γ's is at most $1/3$ as required.

Let (U^*, W^*) be a maximum biclique in $\mathcal{B}(\rho_U, \rho_W)$. As shown in the proof of Theorem 1, with high probability, there is a subset S of the sample X (having size $\hat{m}$) for which the ϵ-biclique $(\Gamma_\epsilon(\Gamma(S)), \Gamma(S))$ has at least as many edges as the maximum biclique (U^*, W^*). Let us denote this subset by S^*.

For any given subset S of X, let $\Gamma(S)$ be denoted by $\hat{W}(S)$, let $\Gamma_\epsilon(\hat{W}(S))$ be denoted by $\hat{U}(S)$, and let $T \cap \Gamma_\epsilon(\hat{W}(S))$ be denoted by $\hat{T}(S)$, so that all notations are as in the original Approximate Maximum Biclique algorithm and Theorem 1. By definition, $\hat{G}(S) = (\hat{U}(S), \hat{W}(S))$ is an ϵ-biclique. Recall, in the proof of Theorem 1, we showed that with high probability over the choice of T, the subset $Z \subset X$ that maximizes $|\hat{T}(Z)| \cdot |\hat{W}(Z)|$ is such that $|E(\hat{U}(Z), \hat{W}(Z))| \geq (1 - 2\epsilon) \cdot |E(\hat{U}(S^*), \hat{W}(S^*))|$. Combining this with the lemma stated in the previous paragraph, we have that for this subset Z, $|E(\hat{U}(Z), \hat{W}(Z))| \geq (1 - 2\epsilon) \cdot |U^*| \cdot |W^*|$.

In the current algorithm, however, we are interested in a possibly different subset that maximizes another quantity. For any subset $S \subset X$ ($|S| = \hat{m}$), let $\hat{U}(S, Y)$ denote the set $\Gamma_{2\epsilon}(\hat{Y}(S))$, and recall that $\hat{T}(S, Y) = T \cap \Gamma_{2\epsilon}(\hat{Y}(S))$. Then, we are interested in the bisubgraph $(\hat{U}(Z, Y), \hat{W}(Z))$ where Z is the subset that maximizes $|\hat{T}(Z, Y)| \cdot |\hat{Y}(Z)|$ (and for which $|\hat{Y}(Z)|$ is sufficiently large). First, it can be verified that by our choice of $q = |Y|$ (using a multiplicative Chernoff bound and a union bound) with high probability over the choice of Y:

1. For every S such that $|\hat{W}(S)| \geq (1-2\epsilon)|W|$, we have that $|\hat{Y}(S)| \geq (1 - (\epsilon/4)) \cdot \frac{|\hat{W}(S)|}{|W|} \cdot |Y|$
2. For every S such that $|\hat{W}(S)| < (1 - 2\epsilon)\rho_w|W|$, we have that $|\hat{Y}(S)| < (1 - \epsilon)\rho_w|Y|$.

Let us assume from now on that these events in fact occur. In particular this implies that no S for which $|\hat{W}(S)| < (1-2\epsilon)\rho_w|W|$ is a candidate output in the third step of the algorithm, and for the subset S^* (which satisfies $|\Gamma(S^*)| \geq \rho_w|W|$ since $\Gamma(S^*) \supseteq W^*$), we have that $|\hat{Y}(S^*)| \geq (1 - (\epsilon/4)) \cdot \rho_w \cdot |Y|$.

Consider from this point on only subsets S such that $|\hat{W}(S)| \geq (1-\epsilon)\rho_{\mathrm{W}}|W|$ (so that in particular $|\hat{Y}(S)| \geq (1-2\epsilon)\rho_{\mathrm{W}}|Y|$). Since $\hat{Y}(S)$ is uniformly distributed in $\Gamma(S)$, we can obtain the following using basic probabilistic arguments (similar to those required for Lemma 2): With high probability over the choice of Y:

1. For every S as stated above, the number of vertices in $\hat{U}(S,Y) = \Gamma_{2\epsilon}(\hat{Y}(S))$ that are not in $\Gamma_{3\epsilon}(\hat{W}(S))$ is at most $(\epsilon/4)\rho_{\mathrm{U}}\rho_{\mathrm{W}}|U|$
2. For the set S^*, the number of vertices that are in $\Gamma_{\epsilon}(\hat{W}(S^*)) = \hat{U}(S^*)$ but not in $\hat{U}(S^*,Y)$ is at most $(\epsilon/4)\rho_{\mathrm{U}}|U|$.

The second item implies that

$$|\hat{U}(S^*,Y)| \geq |\hat{U}(S^*)| - (\epsilon/4)\cdot\rho_{\mathrm{U}}\cdot|U|, \tag{42}$$

and the first item implies that for every candidate S,

$$|E(\hat{U}(S,Y),\hat{W}(S))| \geq (1-3\epsilon)\cdot(|\hat{U}(S,Y)| - (\epsilon/4)\rho_{\mathrm{U}}\rho_{\mathrm{W}}|U|)\cdot|\hat{W}(S)|. \tag{43}$$

Hereafter, assume that these events in fact occur. We also assume that the sample T is such that for every candidate S, $(1-\epsilon/4)\frac{\hat{U}(S,Y)}{|U|} \leq \frac{|\hat{T}(S,Y)|}{t} \leq (1+\epsilon/4)\frac{\hat{U}(S,Y)}{|U|}$.

By Lemma 3, we know that $|\hat{U}(S^*)| \geq (1-\epsilon/4)\cdot|U^*|$. Combining with Eq. (42) we find that

$$|\hat{U}(S^*,Y)| \geq (1-\epsilon/2)\cdot|U^*|. \tag{44}$$

By applying a similar argument to that used in the proof of Theorem 1, we have that with high probability over the choice of T, for the selected subset Z that maximizes $|\hat{T}(Z,Y)|\cdot|\hat{Y}(Z)|$,

$$|\hat{U}(Z,Y)|\cdot|\hat{W}(Z)| \geq (1-\epsilon)\cdot|\hat{U}(S^*,Y)|\cdot|\hat{W}(S^*)|. \tag{45}$$

The lower bounds on $\hat{W}(Z)$ and $\hat{U}(Z,Y)$ are also easily established.

It remains to establish a lower bound on the denseness of the bisubgraph $(\hat{U}(Z,Y)), \hat{W}(Z))$. By Eq. (45) we know that

$$|\hat{U}(Z,Y)| \geq (1-\epsilon)\cdot|\hat{U}(S^*,Y)|\cdot\frac{\hat{W}(S^*)}{\hat{W}(Z)}. \tag{46}$$

Using our bound on $|\hat{U}(S^*,Y)|$ from Eq. (44), and our assumption that $\hat{W}(Z) \geq (\rho_{\mathrm{W}}/2)|W|$ (or otherwise Z would not be selected), we get that

$$|\hat{U}(Z,Y)| \geq (\rho_{\mathrm{U}}\rho_{\mathrm{W}}/4)\cdot|U| \tag{47}$$

Combining this with Eq. (43) we get that

$$|E(\hat{U}(Z,Y),\hat{W}(Z))| \geq (1-4\epsilon)\cdot|\hat{U}(Z,Y))|\cdot|\hat{W}(Z)| \tag{48}$$

as required. Finally, using Eqs. (44) and (45) (and the fact that $|\hat{W}(S^*)| \geq |W^*|$), we get that $|E(\hat{U}(Z, Y)), \hat{W}(Z))|$ is at least $(1 - 6\epsilon)$ times the size of the maximum biclique, as claimed.

The bound on the running time is obtained similarly to the way the bound was obtained in the proof of Theorem 1, with the exception that here we only consider a sample Y from W. □

7. Data streams

So far we have been concerned with clustering a static dataset in time that depends only linearly on the number of attributes and not the number of data points. We now turn to the problem of clustering a dynamic stream where, in contrast, we can read each point in the stream but are memory constrained and thus cannot maintain the entire stream in main memory. A data stream is a more fitting model when a large volume of data is continuously arriving and it is not feasible to store all the data. In the product bundling example, it may not be practical to store every transaction ever made by a customer.

A data stream is a sequence of points, $u_1, \ldots, u_i, \ldots u_{|U|}$, that can be read only once in increasing order of the indices i. We assume that each point $u_i \in \{0, 1\}^d$, although our algorithm works if the points come from any categorical space. Note that the dimensions are fixed ahead of time and thus the stream applies only to the addition of new points and not new dimensions. In addition, we assume that the stream has been partitioned into chunks, $C_1, \ldots, C_J$, where each C_i corresponds to a contiguous section of the stream. In the agglomerative model, the goal is to output a collection of clusters after each of the chunks, C_i, that are approximately as good as the optimum clusters for $C_1 \cup \cdots \cup C_i$. The performance of an algorithm that operates on data streams is measured by the amount of information it stores in main memory and the quality of the solution it finds.

Stream clustering has been previously studied under other definitions of clustering, e.g., k-Center (Charikar et al., 1997) and k-Median (Guha et al., 2003; Charikar, O'Callaghan, & Panigrahy, 2003). Stream clustering has also been studied in the sliding window model for the k-Median clustering objective (Babcock et al., 2003). We discuss how to find one good conjunctive cluster in the agglomerative setting, but the argument can be extended to the case of finding multiple clusters.

Note that the optimum biclique may drift from one chunk of the stream to the next. Let (U_i, W_i) be the optimum bicliques after the chunks $C_1 \cup \cdots \cup C_i$, for $i = 1, \ldots, J$. A straightforward algorithm for identifying the optimum biclique is to consistently maintain separate random samples X and T (as in Algorithm Approximate Maximum Biclique) of the data stream using Vitter's reservoir sampling technique (Vitter, 1985), although with a slightly larger sample to account for the fact that there might be J different optimum bicliques. Such an approach would give guarantees similar to Theorem 1 in that at each chunk the algorithm will have a biclique that has boundably fewer edges than the optimum.

However, since we are allowed to read each point in the stream, we can actually do better. We show that we can identify a relaxed edge biclique with at least as many edges as the optimum, as opposed to boundably fewer. The idea is to apply the Good Seed algorithm by starting from the attributes, W, instead of from the points, U. At any chunk in the stream,

this approach will yield a biclique with at least as many edges as the optimum. Let X be a sample of vertices drawn from W. For each subset S of X of a specified size, and for each w_i in W the stream algorithm maintains the set of attributes S, the number of points $z(S)$ that have streamed by that are in $\Gamma(S)$, and count(w_i, S), the number of points that have streamed by in $\Gamma(S)$ that satisfy the attribute w_i. If at any chunk we wish to output a relaxed edge biclique, we determine $\hat{W}_S$ for each subset S—these are the vertices w_i in W for which count$(w_i, S) \geq (1-\epsilon)z(S)$—and output the biclique corresponding to the subset S with the most edges. Let $\hat{m} = \Theta(\frac{1}{\epsilon^2} \log \frac{J}{\rho_U \epsilon})$ and let $m = O(\frac{\log J}{\rho_W} \hat{m})$.

Stream Approximate Max Edge Biclique Algorithm

1. $X \leftarrow$ sample from W of size m; $\ell \leftarrow 1$
2. For each subset S of X of size $\hat{m}$, initialize $(S, z(S), \{(w_i, \text{count}(w_i, S)) : i = 1, \ldots, |W|\})$ where $z(S) = 0$, and count$(w_i, S) = 0$ for all i.
3. For each point u_j in chunk C_ℓ that streams by
   (a) For each subset S of X
      (i) If u_j satisfies the attributes in the subset S then
         (A) Increment $z(S)$
         (B) If $w_i \in W$ is satisfied by u_j then increment count(w_i, S)
4. Output best biclique: For a given subset S of X let $\hat{W}_S$ be the vertices for which count$(w_i, S) \geq (1-\epsilon)z(S)$. Over all subsets S of X, output the conjunctive description $\hat{W}_S$ with maximum $\sum_{i \in \hat{W}_S}$ count(w_i, S).
5. Proceed to next chunk: Increment ℓ and goto Step 3.

Theorem 5. *Let (U_ℓ, W_ℓ) be the optimum biclique for the chunks $C_1 \cup \cdots \cup C_\ell$ for $\ell = 1, \ldots, J$. Let S_ℓ be the subset of X that yields $\hat{W}_{S_\ell}$ the output of the stream algorithm after the chunks $C_1, \ldots, C_\ell$. With probability at least $\frac{2}{3}$, for all ℓ*

$$|E(\hat{U}_\ell, \hat{W}_\ell)| \geq |E(U_\ell, W_\ell)|.$$

The amount of memory used by the algorithm is quasi-polynomial in $\frac{J}{\rho_U}$ and $\frac{J}{\rho_W}$, exponential in poly$(1/\epsilon)$, *and linear in $|W|$.*

Proof Sketch: For a given prefix of the stream $C_1 \cup \cdots \cup C_\ell$, with probability at least $\frac{2}{3J}$, $|E(\hat{U}_\ell, \hat{W}_\ell)| \geq |E(U_\ell, W_\ell)|$ by an argument similar to that given in Lemma 1 and by the sample sizes m and $\hat{m}$. Thus, by the union bound, with probability at least $\frac{2}{3}$, for all ℓ, $|E(\hat{U}_\ell, \hat{W}_\ell)| \geq |E(U_\ell, W_\ell)|$. □

8. Conclusions and future work

We introduced a new, graph-theoretic formulation of the clustering problem where the goal is to identify a collection of conjunctive cluster descriptions. The formulation differs from

previous approaches in that the clusters discovered overlap and also do not necessarily cluster all the points. In addition, a cluster has a looser interpretation in that a point may be assigned to a cluster description even if it only satisfies most of the attributes in the conjunctive description. A natural correspondence is shown between a conjunctive cluster and a relaxed biclique in a bipartite graph. A simple algorithm is given that discovers a collection of relaxed bicliques that are diverse, and also swamp the large, true bicliques in the graph. The algorithm can be modified to identify such a collection even if the underlying graph possesses no true bicliques, but rather only relaxed bicliques. A key property of the algorithms is their sublinear behavior. Specifically, the algorithms' running time does not depend on the number of points to be clustered and depends only linearly on the number of attributes. Finally, the algorithms can be modified to conjunctively cluster a stream of points, which may be desirable if clusters change over time.

Many interesting problems remain to be solved. In terms of extending our current formulation, it may be worthwhile to consider variants where vertices are weighted. Such an extension might indicate that some variables matter more than others. In the product bundling domain, it might indicate that some products or some customers matter more when constructing clusters. Another extension is to consider variants where edges are weighted.

If U corresponds to some subset of points in $\{0, 1\}^d$ and W corresponds to d variables, then U and W are at different scales since $1 \leq |U| \leq 2^d$ and $1 \leq |W| \leq d$. Thus our quality measure of $|U| \cdot |W|$ implicitly favors clusters with more points over clusters with longer description length. It may be worthwhile to measure the quality of a conjunctive cluster by $|U|^\alpha \cdot |W|^\beta$ for constants α, β as opposed to $|U| \cdot |W|$ to account for the discrepancy in the scale of U and W.

In terms of clustering categorical data, the algorithms given in this paper can be run on a bipartite graph where there is one vertex in W for each attribute/value combination in the original dataset. While such an approach works, our algorithms do not exploit the fact that if a point $u \in U$ is adjacent to a specific attribute/value combination, it will not be adjacent to any of the other values that that attribute can have. We leave open the question of whether there are algorithms that can take advantage of this property.

Our brute force enumeration of all subsets of a fixed size is not so desirable given the ensuing quasi-polynomial dependence on $\frac{1}{\rho_U}$ and $\frac{1}{\rho_W}$. While heuristics of the kind proposed in the Apriori (Agrawal, Imielinski, & Swami, 1993) algorithm may be useful in reducing the number of bicliques considered, it would be interesting to eliminate the quasi-polynomial dependence in the worst case bounds.

In terms of collections of clusters, our definition of "k-best-conjunctive-clusters" is just one option, and it may be interesting to investigate other possibilities. Finally, we have only considered k conjunctive clusters and it may be interesting to consider k disjunctive clusters, or other cluster representations.

Appendix

Lemma 13. *Let A, B be sets such that $|A| \geq (1-\epsilon)|B|$ and $|B \setminus A|/|A \cup B| \geq \delta + 2\epsilon$. Then $|A \setminus B|/|A \cup B| \geq (\delta + \epsilon)$.*

Proof:

$$\begin{aligned}|A\backslash B| = |A| - |A \cap B| &= (|B| - |A \cap B|) + (|A| - |B|) \\ &\geq |B| - |A \cap B| - \epsilon|B| \\ &\geq (\delta + \epsilon)|A \cup B|\end{aligned}$$

□

Lemma 14. *Let I, J, K be bisubgraphs. If I α-dominates J and J β-dominates K then I $(\alpha + \beta)$-dominates K, provided that J has fewer edges than I.*

Proof: For a bisubgraph X, let $|X|$ denote the number of edges in X. For another bisubgraph Y, let $|X\backslash Y|$ denote the number of edges in X that are not in Y, and let $|X \cup Y|$ denote the number of edges in X and Y. We wish to show that if $|J\backslash I|/|I \cup J| = \alpha$ and $|K\backslash J|/|K \cup J| = \beta$ then $|K\backslash I|/|I \cup K| \leq (\alpha + \beta)$.

Note that $|J\backslash I|/|I \cup J| = \alpha$ implies that $|I|/|I \cup J| = (1 - \alpha)$. It thus follows that $|I \cup J| = |I|/(1 - \alpha)$ and consequently that $|J\backslash I| = \alpha(|I|/(1 - \alpha))$. Similarly, $|K\backslash J| = \beta(|J|/(1 - \beta))$.

By the above argument, if we can show that $|K\backslash I| \leq (\alpha + \beta)(|I|/(1 - (\alpha + \beta)))$, then the lemma will be proved.

In order to show this we first need a simple triangle-inequality like property: $|K\backslash I| \leq |K\backslash J| + |J\backslash I|$. In particular, we show that $(K\backslash I) \subseteq (K\backslash J) \cup (J\backslash I)$. Let K' denote the edges that are only in K (and not I or J) and define J' similarly. Then observe that $(K\backslash I) = K' \cup J \cap K$, $(K\backslash J) = K' \cup K \cap I$, and $(J\backslash I) = J' \cup J \cap K$. The triangle-inequality like property follows since $(K' \cup J \cap K) \subseteq (K' \cup K \cap I) \cup (J' \cup J \cap K)$.

Getting back to the lemma, assuming that $|J| \leq |I|$, we have:

$$\begin{aligned}|K\backslash I| &\leq |K\backslash J| + |J\backslash I| \\ &\leq (\beta|J|/(1 - \beta)) + (\alpha|I|/(1 - \alpha)) \\ &\leq (\beta|I|/(1 - (\alpha + \beta))) + (\alpha|I|/(1 - (\alpha + \beta))) \\ &\leq (\alpha + \beta)|I|/(1 - (\alpha + \beta))\end{aligned}$$

□

Acknowledgments

We thank the anonymous reviewers and Dan Oblinger for insightful comments.

Notes

1. In the more general categorical case, there is a vertex w in W for each combination of attribute and value. Furthermore, w is adjacent to a point in U if the point u possesses the attribute/value combination corresponding to w.
2. If the algorithm is not provided with lower bounds ρ_{U} and ρ_{W}, then it can search for them using a standard halving process.

References

Agrawal, R., Gehrke, J. E., Gunopulos, D., & Raghavan, P. (1998). Automatic subspace clustering of high dimensional data for data mining applications. In *Proceedings of the ACM SIGMOD International Conference on Management of Data* (pp. 94–105).

Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. In *Proceedings of the ACM SIGMOD International Conference on Management of Data* (pp. 207–216).

Alon, N., Dar, S., Parnas, M., & Ron, D. (2003). Testing of clustering. *SIAM Journal on Discrete Math*, 393–417.

Alon, N., Fischer, E., Krivelevich, M., & Szegedy, M. (2000). Efficient testing of large graphs. *Combinatorica, 20*, 451–476.

Arora, S., Karger, D., & Karpinski, M. (1995). Polynomial time approximation schemes for dense instances of NP-hard problems. In *Proceedings of the 27th Annual ACM Symposium on the Theory of Computing* (pp. 284–293).

Arya, V., Garg, N., Khandekar, R., Munagala, K., & Pandit, V. (2001). Local search heuristic for *k*-median and facility location problems. In *Proceedings of the 33rd Annual ACM Symposium on the Theory of Computing* (pp. 21–29).

Babcock, B., Datar, M., Motwani, R., & O'Callaghan, L. (2003). Maintaining variance and *k*-medians over data stream windows. In *Proceedings of the 22nd ACM Symposium on Principles of Database Systems* (pp. 234–243).

Bansal, N., Blum, A., & Chawla, S. (2002). Correlation clustering. In *Proceedings of the 43rd Annual Symposium on Foundations of Computer Science* (pp. 238–247).

Charikar, M. (2000). Greedy approximation algorithms for finding dense components in a graph. In *Proceedings of the 3rd International Workshop on Approximation Algorithms for Combinatorial Optimization Problems* (pp. 84–95).

Charikar, M., Chekuri, C., Feder, T., & Motwani, R. (1997). Incremental clustering and dynamic information retrieval. In *Proceedings of the 29th Annual ACM Symposium on Theory of Computing* (pp. 626–635).

Charikar, M., & Guha, S. (1999). Improved combinatorial algorithms for the facility location and *k*-median problems. In *Proceedings of the 40th Annual Symposium on Foundations of Computer Science* (pp. 378–388).

Charikar, M., Guruswami, V., & Wirth, A. (2003). Clustering with qualitative information. In *Proceedings of the 44th Annual Symposium on Foundations of Computer Science* (pp. 524–533).

Charikar, M., O'Callaghan, L., & Panigrahy, R. (2003). Better streaming algorithms for clustering problems. In *Proceedings of the 35th Annual ACM Symposium on the Theory of Computing* (pp. 30–39).

Cheng, Y., & Church, G. (2000). Biclustering of expression data. In *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology* (pp. 93–103).

Fernandez de la Vega, W. (1996). MAX-CUT has a randomized approximation scheme in dense graphs. *Random Structures and Algorithms, 8:3*, 187–198.

Demaine, E., & Immorlica, N. (2003). Correlation clustering with partial information. In *Proceedings of the 6th International Workshop on Approximation Algorithms for Combinatorial Optimization Problems* (pp. 1–13).

Dempster, A., Laird, N., & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society Series B, 39*, 1–38.

Dhillon, I. (2001). Co-clustering documents and words using bipartite spectral graph partitioning. In *Knowledge Discovery and Data Mining* (pp. 269–274).

Dhillon, I., Mallela, S., & Modha, D. (2003). Information-theoretic co-clustering. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 89–98).

Duda, R., Hart, P., & Stork, D. (2000). *Pattern classification*, 2nd edn. (vol. November). New York: John Wily & Sons, Inc.

Emmanuel, D., & Fiat, A. (2003). Correlation clustering—Minimizing disagreements on arbitrary weighted graphs. In *Proceedings of the 11th European Symposium on Algorithms* (pp. 208–220).

Feder, T., & Greene, D. (1988). Optimal algorithms for approximate clustering. In *Proceedings of the 20th Annual ACM Symposium on the Theory of Computing* (pp. 434–444).

Feder, T., & Motwani, R. (1995). Clique partitions, graph compression and speeding-up algorithms. *J. Computer and System Sciences, 51*, 261–272.

Feige, U. (2002). Relations between average case complexity and approximation complexity. In *Proceedings of the 34th Annual ACM Symposium on the Theory of Computing* (pp. 534–543).

Fisher, D., & Langley, P. (1985). Approaches to conceptual clustering. In *Proceedings of the 9th International Joint Conference on Artificial Intelligence*, vol II (pp. 691–697).

Flake, G., Lawrence, S., & Giles, C. L. (2000). Efficient identification of web communities. In *Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 150–160).

Frieze, A., & Kanan, R. (1999). Quick approximation to matrices and applications. *Combinatorica, 19:2*, 175–220.

Ganti, V., Gehrke, J., & Ramakrishnan, R. (1999). CACTUS—Clustering categorical data using summaries. In *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 73–83).

Garey, M. R., & Johnson, D. S. (1979). *Computers and intractability: A guide to the theory of NP-completeness*. W.H. Freeman and Company.

Gibson, D., Kleinberg, J., & Raghavan, P. (1998). Inferring web communities from link topology. In *Proceedings of the 9th ACM Conference on Hypertext and Hypermedia* (pp. 225–234).

Gibson, D., Kleinberg, J., & Raghavan, P. (2000). Clustering categorical data: An approach based on dynamical systems. *VLDB Journal: Very Large Data Bases, 8:3/4*, 222–236.

Goldberg, A. (1984). Finding a maximum density subgraph. *UC Berkeley Tech Report, CSD-84-171*.

Goldreich, O., Goldwasser, S., & Ron, D. (1998). Property testing and its connection to learning and approximation. *Journal of the ACM, 45:4*, 653–750.

Guha, S., Meyerson, A., Mishra, N., Motwani, R., & O'Callaghan, L. (2003). Clustering data streams: Theory and practice. *IEEE Transactions on Knowledge and Data Engineering, 15:3*, 515–528.

Guha, S., Rastogi, R., & Shim, K. (2000). ROCK: A robust clustering algorithm for categorical attributes. *Information Systems, 25:5*, 345–366.

Gunopulos, D., Mannila, H., Khardon, R., & Toivonen, H. (1997). Data mining, hypergraph transversals, and machine learning. In *Proceedings of the Sixteenth ACM SIG-SIGMOD-SIGART Symposium on Principles of Database Systems* (pp. 209–216).

Hartigan, J. (1972). Direct clustering of a data matrix. *Journal of the American Statistical Association, 67:337*, 123–129.

Hochbaum, D. (1998). Approximating clique and biclique problems. *Journal of Algorithms, 29:1*, 174–200.

Hochbaum, D., & Shmoys, D. (1986). A unified approach to approximate algorithms for bottleneck problems. *Journal of the ACM, 33:3*, 533–550.

Indyk, P. (1999). Sublinear time algorithms for metric space problems. In *Proceedings of the 31st Annual ACM Symposium on the Theory of Computing* (pp. 428–434).

Jain, N., & Vazirani, V. (1999). Primal-dual approximation algorithms for metric facility location and k-median problems. In *Proceedings of the 40th Annual Symposium on Foundations of Computer Science* (pp. 2–13).

Kannan, R., Vempala, S., & Vetta, A. (2000). On clusterings—Good, bad and spectral. In *Proceedings of the 41st Annual Symposium on Foundations of Computer Science* (pp. 367–377).

Kanungo, T., Mount, D. M., Netanyahu, N., Piatko, C., Silverman, R., & Wu, A. Y. (2002). A local search approximation algorithm for k-means clustering. In *Proceedings of the 18th Annual ACM Symposium on Computational Geometry* (pp. 10–18).

Kleinberg, J., Papadimitriou, C., & Raghavan, P. (1998). Approximation algorithms for segmentation problems. In *Proceedings of the 30th Annual ACM Symposium on the Theory of Computing* (pp. 473–482).

Kumar, R., Raghavan, P., Rajagopalan, S., & Tomkins, A. (1999). Trawling the web for emerging cyber-communities. *Computer Networks, 31:11–16*, 1481–1493.

Madeira, S., & Oliveira, A. (2004). Biclustering algorithms for biological data analysis: A survey. Instituto de Engenharia de Sistemas e Computadores, INESC-ID Tech. Rep. 1/2004. Lisbon, Portugal.

Michalski, R. (1980). Knowledge acquisition through conceptual clustering: A theoretical framework and an algorithm for partitioning data into conjunctive concepts. Technical Report 1026, Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, Illinois.

Mishra, N., Oblinger, D., & Pitt, L. (2001). Sublinear time approximate clustering. In *Proceedings of the 12th Annual ACM-SIAM Symposium on Discrete Algorithms* (pp. 439–447).

Murali, T., & Kasif, S. (2003). Extracting conserved gene expression motifs from gene expression data. In *Pacific Symposium on Biocomputing* (pp. 77–88).

Peeters, R. (2003). The maximum edge biclique problem is NP-complete. *Discrete Applied Mathematics, 131:3*, 651–654.

Pitt, L., & Reinke, R. E. (1987). Criteria for polynomial-time (conceptual) clustering. *Machine Learning, 2:4*, 371–396.

Procopiuc, C., Jones, M., Agarwal, P., & Murali, T. (2002). A Monte Carlo algorithm for fast projective clustering. In *Proceedings of the ACM SIGMOD International Conference on Management of Data* (pp. 418–427).

Selim, S., & Ismail, M. (1984). K-means type algorithms: A generalized convergence theorem and characterization of local optimality. *IEEE Transansactions on Pattern Analysis and Machine Intelligence, 6:1*, 81–86.

Tanay, A., Sharan, R., & Shamir, R. (2002). Discovering statistically significant biclusters in gene expression data. *Bioinformatics, 18:1*, 136–144.

Thorup, M. (2001). Quick k-median, k-center, and facility location for sparse graphs. In *Proceedings of the Annual International Colloquium on Automata, Languages and Programming* (pp. 249–260).

Peleg, D., Feige, U., & Kortsarz, G. (2001). The dense-k-subgraph problem. *Algorithmica, 29:3*, 410–421.

Vitter, J. (1985). Random sampling with a reservoir. *ACM Transactions on Mathematical Software, 11:1*, 37–57.

Yannakakis, M. (1981). Node-deletion problems on bipartite graphs. *SIAM Journal on Computing, 10:2*, 310–327.

Received February 3, 2003
Revised February 17, 2004
Accepted February 18, 2004
Final manuscript February 18, 2004