Clustering Aggregation

ARISTIDES GIONIS
Yahoo! Research Labs, Barcelona
HEIKKI MANNILA
University of Helsinki and Helsinki University of Technology
and
PANAYIOTIS TSAPARAS
Microsoft Search Labs

We consider the following problem: given a set of clusterings, find a single clustering that agrees as much as possible with the input clusterings. This problem, *clustering aggregation*, appears naturally in various contexts. For example, clustering categorical data is an instance of the clustering aggregation problem; each categorical attribute can be viewed as a clustering of the input rows where rows are grouped together if they take the same value on that attribute. Clustering aggregation can also be used as a metaclustering method to improve the robustness of clustering by combining the output of multiple algorithms. Furthermore, the problem formulation does not require a priori information about the number of clusters; it is naturally determined by the optimization function.

In this article, we give a formal statement of the clustering aggregation problem, and we propose a number of algorithms. Our algorithms make use of the connection between clustering aggregation and the problem of *correlation clustering*. Although the problems we consider are NP-hard, for several of our methods, we provide theoretical guarantees on the quality of the solutions. Our work provides the best deterministic approximation algorithm for the variation of the correlation clustering problem we consider. We also show how sampling can be used to scale the algorithms for large datasets. We give an extensive empirical evaluation demonstrating the usefulness of the problem and of the solutions.

Categories and Subject Descriptors: H.2.8 [**Database Management**]: Database Applications—*Data mining*; F.2.2 [**Analysis of Algorithms and Problem Complexity**]: Nonnumerical Algorithms and Problems

General Terms: Algorithms

Additional Key Words and Phrases: Data clustering, clustering categorical data, clustering aggregation, correlation clustering

A shorter version of this article proceedings of appeared in the International Conference on Data Engineering (ICDE) 2005.

Author's address: A. Gionis, Yahoo! Research Labs, Barcelona, Spain; email: gionis@yahoo-inc.com.

ACM Reference Format:
Gionis, A., Mannila, H., and Tsaparas, P. 2007. Clustering aggregation. ACM Trans. Knowl. Discov. Data, 1, 1, Article 4 (March 2007), 30 pages. DOI = 10.1145/1217299.1217303 http://doi.acm.org/10.1145/1217299.1217303

1. INTRODUCTION

Clustering is an important step in the process of data analysis and has applications to numerous fields. Informally, *clustering* is defined as the problem of partitioning data objects into groups (clusters) such that objects in the same group are similar, while objects in different groups are dissimilar. This definition assumes that there is some well-defined *quality measure* that captures intracluster similarity and/or intercluster dissimilarity. Clustering then becomes the problem of grouping together data objects so that the quality measure is optimized. There is an extensive body of literature on clustering methods, see, for instance, Jain and Dubes [1987]; Hand et al. [2001]; Han and Kamber [2001].

In this article, we consider an approach to clustering that is based on the concept of *aggregation*. We assume that given a set of data objects, we can obtain some information on how these objects should be clustered. This information comes in the form of m clusterings $\mathcal{C}_1, \ldots, \mathcal{C}_m$. The objective is to produce a single clustering $\mathcal{C}$ that agrees as much as possible with the m input clusterings. We define a disagreement between two clusterings $\mathcal{C}$ and $\mathcal{C}'$ as a pair of objects (v, u) such that $\mathcal{C}$ places them in the same cluster, while $\mathcal{C}'$ places them in different clusters or vice versa. If $d(\mathcal{C}, \mathcal{C}')$ denotes the number of disagreements between $\mathcal{C}$ and $\mathcal{C}'$, then the task is to find a clustering $\mathcal{C}$ that minimizes $\sum_{i=1}^{m} d(\mathcal{C}_i, \mathcal{C})$.

As an example, consider the dataset $V = \{v_1, v_2, v_3, v_4, v_5, v_6\}$ that consists of six objects, and let $\mathcal{C}_1 = \{\{v_1, v_2\}, \{v_3, v_4\}, \{v_5, v_6\}\}$, $\mathcal{C}_2 = \{\{v_1, v_3\}, \{v_2, v_4\}, \{v_5\}, \{v_6\}\}$, and $\mathcal{C}_3 = \{\{v_1, v_3\}, \{v_2, v_4\}, \{v_5, v_6\}\}$ be three clusterings of V. Figure 1 shows the three clusterings where each column corresponds to a clustering, and a value i denotes that the tuple in that row belongs in the i-th cluster of the clustering in that column. The right-most column is the clustering $\mathcal{C} = \{\{v_1, v_3\}, \{v_2, v_4\}, \{v_5, v_6\}\}$ that minimizes the total number of disagreements with the clusterings $\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3$. In this example, the total number of disagreements is 5 one with the clustering $\mathcal{C}_2$ for the pair (v_5, v_6), and four with the clustering $\mathcal{C}_1$ for the pairs $(v_1, v_2), (v_1, v_3), (v_2, v_4), (v_3, v_4)$. It is not hard to see that this is the minimum number of disagreements possible for any partition of the dataset V.

We define *clustering aggregation* as the optimization problem where, given a set of m clusterings, we want to find the clustering that minimizes the total number of disagreements with the m clusterings. Clustering aggregation provides a general framework for dealing with a variety of problems related to clustering: (i) it gives a natural clustering algorithm for categorical data; (ii) it handles heterogeneous data where tuples are defined over incomparable attributes; (iii) it determines the appropriate number of clusters and it detects outliers; (iv) it provides a method for improving the clustering robustness by combining the results of many clustering algorithms; and (v) it allows for clustering of data that is vertically partitioned in order to preserve privacy. We elaborate on the properties and the applications of clustering aggregation in Section 2.

	$\mathcal{C}_1$	$\mathcal{C}_2$	$\mathcal{C}_3$	$\mathcal{C}$
v_1	1	1	1	1
v_2	1	2	2	2
v_3	2	1	1	1
v_4	2	2	2	2
v_5	3	3	3	3
v_6	3	4	3	3

Fig. 1. An example of clustering aggregation. $\mathcal{C}_1$, $\mathcal{C}_2$, and $\mathcal{C}_3$ are the input clusterings, and $v_1, \ldots, v_6$ are the objects to be clustered. A value k in the entry $(v_i, \mathcal{C}_j)$ means that object v_i belongs to cluster k of the clustering C_j. Column $\mathcal{C}$ is the clustering that minimizes the disagreements with clusterings $\mathcal{C}_1, \mathcal{C}_2$, and $\mathcal{C}_3$.

The algorithms we propose for the problem of clustering aggregation take advantage of a related problem which is known as *correlation clustering* [Bansal et al. 2004]. In the version of the problem we consider, we are given a graph with edge weights X_{uv} satisfying the triangle inequality. We map clustering aggregation to correlation clustering by considering the tuples of the dataset as vertices of a graph and summarizing the information provided by the m input clusterings by weights on the edges of the graph. The weight X_{uv} of the edge (u, v) is the fraction of clusterings that place u and v in different clusters. For example, the correlation clustering instance for the dataset in Figure 1 is shown in Figure 2. Note that if the weight of the edge (u, v) is less than 1/2, then the majority of the clusterings place u and v together, while if the weight is greater than 1/2, the majority places u and v in different clusters. Ideally, we would like to cut all edges with weight more than 1/2 and not cut all edges with weight less than 1/2. The goal in correlation clustering is to find a partition of the vertices of the graph that cuts as few as possible of the edges with low weight (less than 1/2) and as many as possible of the edges with high weight (more than 1/2). In Figure 2, clustering $\mathcal{C} = \{\{v_1, v_3\}, \{v_2, v_4\}, \{v_5, v_6\}\}$ is obviously the optimal clustering since it only cuts edges with weight greater than 1/2.

Clustering aggregation has been previously considered under a variety of names (consensus clustering, clustering ensemble, clustering combination) in a variety of different areas such as machine learning [Strehl and Ghosh 2002; Fern and Brodley 2003], pattern recognition [Fred and Jain 2002], bioinformatics [Filkov and Skiena 2004], and data mining [Topchy et al. 2004; Boulis and Ostendorf 2004]. The problem of correlation clustering is interesting in its own right, and it has recently attracted a lot of attention in the theoretical computer science community [Bansal et al. 2004; Charikar et al. 2003; Demaine et al. 2006; Swamy 2004]. We review some of the related literature on both clustering aggregation and correlation clustering in Section 3.

Our contributions can be summarized as follows.

- We formally define the problem of clustering aggregation, and we demonstrate the connection between clustering aggregation and correlation clustering.
- We present a number of algorithms for clustering aggregation and correlation clustering. We also propose a sampling mechanism that allows our algorithms to handle large datasets. The problems we consider are NP-hard, yet we are

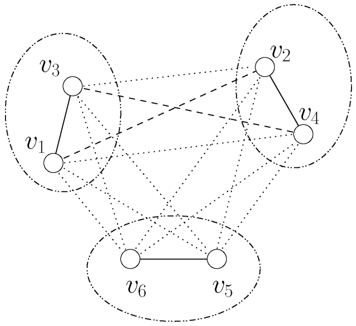


Fig. 2. A correlation clustering instance for the dataset in Figure 1. Solid edges indicate distances of 1/3, dashed edges indicate distances of 2/3, and dotted edges indicate distances of 1. The circles depict the clusters of clustering $\mathcal{C}$ that minimizes the number of disagreements.

still able to provide approximation guarantees for many of the algorithms we propose. For the formulation of correlation clustering we consider, we give a combinatorial deterministic 3-approximation algorithm, which is an improvement over the previously best known deterministic 9-approximation algorithm.

—We present an extensive experimental study wherein we demonstrate the benefits of our approach. Furthermore, we show that our sampling technique reduces the running time of the algorithms without sacrificing the quality of the clustering.

The rest of this article is structured as follows. In Section 2, we discuss the various applications of the clustering aggregation framework. Section 3 contains a review of the related work. The problem statements we consider in this article are formally defined in Section 4. In Section 5, we describe in detail the proposed algorithms for clustering aggregation and correlation clustering as well as the sampling-based algorithm that allows us to handle large datasets. Our experiments on synthetic and real datasets are presented in Section 7. Finally, Section 8 is a short conclusion.

2. APPLICATIONS OF CLUSTERING AGGREGATION

Clustering aggregation can be applied in various settings. We will now present some of the main applications and features of our framework.

Clustering categorical data. An important application of clustering aggregation is that it provides a very natural method for clustering categorical data. Consider a dataset with tuples $t_1, \ldots, t_n$ over a set of categorical attributes $A_1, \ldots, A_m$. The idea is to view each attribute A_j as a way of producing a simple clustering of the data, that is, if A_j contains k_j distinct values, then A_j partitions the data in k_j clusters, one cluster for each value. Then, clustering aggregation considers all those m clusterings produced by the m attributes and tries to find a clustering that agrees as much as possible with all of them.

For example, consider a `Movie` database. Each tuple in the database corresponds to a movie that is defined over a set of attributes such as `Director`, `Actor`, `Actress`, `Genre`, `Year`, etc, some of which take categorical values. Note that each of the categorical attributes naturally defines a clustering. For example, the `Movie.Genre` attribute groups the movies according to their genre, while the `Movie.Director` according to who directed the movie. The objective is to combine all these clusterings into a single clustering.

Methods for aggregating clusterings can also be extended to incorporate domain knowledge when available. For example, if some attributes are more important than others, then we can increase their influence on the aggregate solution by including multiple copies of the specific attributes in the clustering aggregation. Similarly, if we have some prior knowledge about the relationships between the values of a specific attribute (e.g., a hierarchy that renders `thriller` closer to `horror` than to `comedy`) we can incorporate it by adding clusterings that place similar values together. In the movie database example, if the `Movie.Genre` takes the values `thriller`, `horror`, and `comedy`, we can add to the clustering aggregation a clustering that places all movies with genre `thriller` and `horror` together; this will bias the aggregation algorithm towards merging these two values. In this way, we can incorporate the available domain knowledge in the final result in a very natural way.

Clustering heterogeneous data. The clustering aggregation method can be particularly effective in cases where the data are defined over heterogeneous attributes that contain incomparable values. Consider for example the case that there are many numerical attributes whose units are incomparable (say, `Movie.Budget` and `Movie.Year`) and so it does not make sense to compare numerical vectors directly using an L_p-type distance measure. A similar situation arises in the case where the data contains a mix of categorical and numerical values. In such cases, the data can be partitioned vertically into sets of homogeneous attributes, obtain a clustering for each of these sets by applying the appropriate clustering algorithm, and then aggregate the individual clusterings into a single clustering.

Identifying the correct number of clusters. One of the most important features of the formulation of clustering aggregation is that there is no need to specify the number of clusters in the result. The automatic identification of the appropriate number of clusters is a deep research problem that has attracted significant attention (see, e.g., Schwarz [1978]; Hamerly and Elkan [2003]; Smyth [2000]). For most clustering approaches, the quality of the solution (likelihood, sum of distances to cluster centers, etc.) improves as the number of clusters is increased. Thus, the trivial solution of all singleton clusters is the optimal. There are two ways of handling the problem. The first is to have a hard constraint on the number of clusters or on their quality. For example, in agglomerative algorithms, one can either fix in advance the number of clusters in the final clustering or impose a bound on the distance beyond which no pair of clusters will be merged. The second approach uses model selection methods, for example, Bayesian information criterion (BIC) [Schwarz 1978], or cross-validated likelihood [Smyth 2000] to compare models with different numbers of clusters.

The formulation of clustering aggregation gives one way of automatically selecting the number of clusters. If many input clusterings place two objects in the same cluster, then it will not be beneficial for a clustering-aggregation solution to split these two objects. Thus, the solution of all singleton clusters is not a trivial solution for our objective function. Furthermore, if there are k subsets of data objects in the dataset such that, for each subset, the majority of the input clusterings places its elements together and separates them from the rest, then the clustering aggregation algorithm will correctly identify the k clusters, without any prior knowledge of k. A simple instance is the example in Figures 1 and 2 where the optimal solution $\mathcal{C}$ naturally discovers a set of 3 clusters in the dataset.

Indeed, the structure of the objective function ensures that the clustering aggregation algorithms will naturally settle to the appropriate number of clusters. As we will show in our experimental section, our algorithms take advantage of this feature, and for all our datasets, they generate clusterings with a very reasonable number of clusters. On the other hand, if the user insists on a predefined number of clusters, most of our algorithms can be easily modified to return that specific number of clusters. For example, the agglomerative algorithm described in Section 5 can be modified to continue merging clusters until the predefined number is reached.

Detecting outliers. The ability to detect outliers is closely related to the ability to identify the correct number of clusters. If a node is not close to any other nodes, then from the point of view of the objective function, it would be beneficial to assign that node in a singleton cluster. In the case of categorical data clustering, the scenarios for detecting outliers are very intuitive. If a tuple contains many uncommon values, it does not participate in clusters with other tuples, and it is likely that it will be identified as an outlier. Another scenario where it pays off to consider a tuple as an outlier is when the tuple contains common values (and therefore it participates in big clusters in the individual input clusterings), but there is no consensus on a common cluster (e.g., a `horror` movie featuring actress `Julia.Roberts` and directed by the independent director `Lars.vonTrier`).

Improving clustering robustness. Different clustering algorithms have different qualities and different shortcomings. Some algorithms might perform well in specific datasets but not in others, or they might be very sensitive to parameter settings. For example, the single-linkage algorithm is good at identifying elongated regions, but it is sensitive to clusters connected with narrow strips of points. The k-means algorithm is a widely-used technique, but it favors spherical clusters, it is sensitive to clusters of uneven size, and it can get stuck in local optima.

We suggest that by aggregating the results of different clustering algorithms, we can significantly improve the robustness and quality of the final clustering. The idea is that different algorithms make different types of mistakes that can be canceled out in the final aggregation. Furthermore, for objects that are outliers or noise, it is most likely that there will be no consensus on how they should be clustered, and thus it will be more beneficial for the aggregation algorithm to single them out. The intuition is similar to performing rank aggregation for improving the results of Web searches [Dwork et al. 2001]. Our experiments indicate that clustering aggregation can significantly improve the results of individual algorithms.

Privacy-preserving clustering. Consider a situation where a database table is vertically split and different attributes are maintained in different sites. Such a situation might arise in cases where different companies or governmental administrations maintain various sets of data about a common population of individuals. For such cases, our method offers a natural model for clustering the data maintained in all sites as a whole in a privacy-preserving manner, that is, without the need for the different sites to reveal their data to each other and without the need to rely on a trusted authority. Each site clusters its own data independently, and then all resulting clusterings are aggregated. The only information revealed is which tuples are clustered together; no information is revealed about data values of any individual tuples.

3. RELATED WORK

A source of motivation for our work is the literature on comparing and merging multiple rankings [Dwork et al. 2001; Fagin et al. 2003]. Dwork et al. [2001] demonstrated that combining multiple rankings in a metasearch engine for the Web yields improved results and removes noise (spam). The intuition behind our work is similar. By combining multiple clusterings we improve the clustering quality and remove noise (outliers).

The problem of clustering aggregation has been previously considered in the machine learning community, under the name *clustering ensemble* and *consensus clustering*. Strehl and Ghosh [2002] consider various formulations for the problem, most of which reduce the problem to a hypergraph partitioning problem. In one of their formulations, they consider the same graph as in the correlation clustering problem. The solution they propose is to compute the best k-partition of the graph, which does not take into account the penalty for merging two nodes that are far apart. All of their formulations assume that the correct number of clusters is given as a parameter to the algorithm.

Fern and Brodley [2003] apply the clustering aggregation idea to a collection of soft clusterings they obtain by random projections. They use an agglomerative algorithm similar to ours, but again they do not penalize for merging dissimilar nodes. Fred and Jain [2002] propose to use a single linkage algorithm to combine multiple runs of the k-means algorithm.Cristofor and Simovici [2001] observe the connection between clustering aggregation and clustering of categorical data. They propose information theoretic distance measures, and they propose genetic algorithms for finding the best aggregation solution. Boulis and Ostendorf [2004] use linear programming to discover a correspondence between the labels of the individual clusterings and those of an optimal metaclustering. Topchy et al. [2004] define clustering aggregation as a maximum likelihood estimation problem, and they propose an EM algorithm for finding the consensus clustering. Filkov and Skiena [2004] consider the same distance measure between clusterings as ours. They propose a simulating annealing algorithm for finding an aggregate solution and a local search algorithm similar to ours. They consider the application of clustering aggregation to the analysis of microarray data. Recently, Mielikäinen et al. [2006] considered the problem of aggregation for segmentations of sequential data. A segmentation can be thought of as an order-preserving clustering. They showed that the problem can be solved optimally using dynamic programming.

There is an extensive literature in the field of theoretical computer science for the problem of correlation clustering. The problem was first defined by Bansal et al. [2004]. In their definition, the input is a complete graph with $+1$ and -1 weights on the edges. The objective is to partition the nodes of the graph so as to minimize the number of positive edges that are cut, and the number of negative edges that are not cut. The best known approximation algorithm for this problem is by Charikar et al. [2003] who give an LP-based algorithm that achieves an approximation factor of 4. The LP-based algorithm of Charikar et al. is very similar to the BALLS algorithm proposed in this article. One difference is that in Charikar et al., the algorithm works with the edge weights obtained by the LP solution, while in our case, the algorithm works with the input edge weights. The algorithm of Charikar et al. combined with a reduction in Bansal et al. [2004] (Theorem 23) provides a deterministic 9-approximation algorithm for correlation clustering when the edge weights satisfy the *probability condition* (i.e., for every edge (i, j), the cost for taking that edge is $X_{ij} \in [0, 1]$, while the cost for splitting an edge is $1 - X_{ij}$), even if they do not satisfy the triangle inequality.

When the edge weights are arbitrary, the problem is equivalent to the multicut problem as shown in Demaine, Emanuel, Fiat, and Immorlica [2006], and there is a $O(\log n)$-approximation bound. If one considers the corresponding maximization problem, that is, maximize the agreements rather than minimize disagreements, then the situation is much better. Even in the case of graphs with arbitrary edge weights, there is a 0.76-approximation algorithm using semidefinite programming [Charikar et al. 2003; Swamy 2004].

Recently, Ailon et al. [2005] considered a variety of correlation clustering problems. They proposed an algorithm very similar to the BALLS algorithm, and they showed that if the weights obey the probability condition, then their algorithm achieves expected approximation ratio 5. If the weights X_{ij} also obey the triangle inequality, then the algorithm achieves expected approximation ratio 2. For the clustering aggregation problem, they show that choosing the best solution between their algorithm and the best of the input clusterings (the BESTCLUSTERING algorithm) yields a solution with expected approximation ratio 11/7.

One difference of our work with the work of Ailon et al. [2005] is that our algorithm is deterministic, while their algorithm is probabilistic. Furthermore, we investigate experimentally the performance of our algorithms for various applications such as clustering of categorical data, clustering robustness, and finding the correct number of clusters. For the problem of categorical clustering, we compare our algorithms with various existing algorithms to demonstrate the benefits of our approach.

4. DESCRIPTION OF THE FRAMEWORK

We begin our discussion of the clustering aggregation framework by introducing our notation. Consider a set of n objects $V = \{v_1, \ldots, v_n\}$. A clustering $\mathcal{C}$ of V is a *partition* of V into k disjoint sets $C_1, \ldots, C_k$, that is, $\bigcup_i^k C_i = V$ and $C_i \cap C_j = \emptyset$ for all $i \neq j$. The k sets $C_1, \ldots, C_k$ are the clusters of $\mathcal{C}$. For each $v \in V$, we use $\mathcal{C}(v)$ to denote the label of the cluster to which the object v belongs, that is, $\mathcal{C}(v) = j$ if and only if $v \in C_j$. In the following, we consider m clusterings: we write $\mathcal{C}_i$ to denote the ith clustering, and k_i for the number of clusters of $\mathcal{C}_i$. In some cases, it is possible that there are weights associated with the clusterings provided by the user, so that weight w_i captures the user's belief on the quality of clustering $\mathcal{C}_i$. Most of our algorithms can handle a weighted version of the clustering problem, however, we are mostly focusing on the case that all the weights are equal, that is, $w_i = 1$.

In the clustering aggregation problem, the task is to find a clustering that minimizes the disagreements with a set of input clusterings. To make the notion more precise, we need to define a measure of disagreement between clusterings. Consider first two objects u and v in V. The following simple 0/1 indicator function checks if two clusterings $\mathcal{C}_1$ and $\mathcal{C}_2$ disagree on the clustering of u and v.

$$d_{u,v}(\mathcal{C}_1, \mathcal{C}_2) = \begin{cases} 1 & \text{if } \mathcal{C}_1(u) = \mathcal{C}_1(v) \text{ and } \mathcal{C}_2(u) \neq \mathcal{C}_2(v), \\ & \text{or } \mathcal{C}_1(u) \neq \mathcal{C}_1(v) \text{ and } \mathcal{C}_2(u) = \mathcal{C}_2(v), \\ 0 & \text{otherwise.} \end{cases}$$

The distance between two clusterings $\mathcal{C}_1$ and $\mathcal{C}_2$ is defined as the number of pairs of objects on which the two clusterings disagree, that is,

$$d_V(\mathcal{C}_1, \mathcal{C}_2) = \sum_{(u,v) \in V \times V} d_{u,v}(\mathcal{C}_1, \mathcal{C}_2).$$

The clustering aggregation problem can now be formalized as follows.

Problem 1 (*Clustering Aggregation*). Given a set of objects V and m clusterings $\mathcal{C}_1, \ldots, \mathcal{C}_m$ on V, compute a new clustering $\mathcal{C}$ that minimizes the total number of disagreements with all the given clusterings, that is, it minimizes

$$D(\mathcal{C}) = \sum_{i=1}^{m} d_V(\mathcal{C}_i, \mathcal{C}).$$

If weights are provided for the clusterings, then the objective function becomes $D(\mathcal{C}) = \sum_{i=1}^{m} w_i \cdot d_V(\mathcal{C}_i, \mathcal{C})$. Note that when all weights w_i are integers, then we can solve the weighted clustering aggregation problem by adding w_i copies of each clustering $\mathcal{C}_i$ to the input, and then solving the simple aggregation problem. The Clustering Aggregation problem was shown to be NP-hard by Filkov and Skiena [2004], using the results of Barthelemy and Leclerc [1995].

It is easy to show that the distance measure $d_V(\cdot, \cdot)$ satisfies the *triangle inequality* on the space of clusterings.

OBSERVATION 1. *Given a set of objects V, and clusterings $\mathcal{C}_1$, $\mathcal{C}_2$, $\mathcal{C}_3$ on V, we have $d_V(\mathcal{C}_1, \mathcal{C}_3) \leq d_V(\mathcal{C}_1, \mathcal{C}_2) + d_V(\mathcal{C}_2, \mathcal{C}_3)$.*

PROOF. It is sufficient to show that, for each pair (u, v), we have $d_{u,v}(\mathcal{C}_1, \mathcal{C}_3) \leq d_{u,v}(\mathcal{C}_1, \mathcal{C}_2) + d_{u,v}(\mathcal{C}_2, \mathcal{C}_3)$, and the lemma follows from the definition of d_V. Since $d_{u,v}$ takes 0/1 values, the only case in which the triangle inequality could be violated is if $d_{u,v}(\mathcal{C}_1, \mathcal{C}_3) = 1$ and $d_{u,v}(\mathcal{C}_1, \mathcal{C}_2) = d_{u,v}(\mathcal{C}_2, \mathcal{C}_3) = 0$. However, $d_{u,v}(\mathcal{C}_1, \mathcal{C}_3) = 1$ implies that either $\mathcal{C}_1(u) = \mathcal{C}_1(v)$ and $\mathcal{C}_3(u) \neq \mathcal{C}_3(v)$ or that $\mathcal{C}_1(u) \neq \mathcal{C}_1(v)$ and $\mathcal{C}_3(u) = \mathcal{C}_3(v)$. Assume that $\mathcal{C}_1(u) = \mathcal{C}_1(v)$ and $\mathcal{C}_3(u) \neq \mathcal{C}_3(v)$. Then $d_{u,v}(\mathcal{C}_1, \mathcal{C}_2) = 0$ implies that the clusterings $\mathcal{C}_1$ and $\mathcal{C}_2$ agree on the clustering of u and v; therefore, $\mathcal{C}_2(u) = \mathcal{C}_2(v)$. However, since $d_{u,v}(\mathcal{C}_2, \mathcal{C}_3) = 0$, $\mathcal{C}_2$ and $\mathcal{C}_3$ are also in agreement, and thus $\mathcal{C}_3(u) = \mathcal{C}_3(v)$, which contradicts our assumption that $\mathcal{C}_3(u) \neq \mathcal{C}_3(v)$. The case where $\mathcal{C}_1(u) \neq \mathcal{C}_1(v)$ and $\mathcal{C}_3(u) = \mathcal{C}_3(v)$ is treated symmetrically. □

The algorithms we propose for the problem of clustering aggregation take advantage of a related formulation which is a version of the problem known as *correlation clustering* [Bansal et al. 2004]. The version of correlation clustering we consider is different than the original version proposed in Bansal et al. [2004], which coresponds to having binary edge weights, and it is more restricted than the general weighted version since in our case the edge weights satisfy the triangle inequality. Formally, we define correlation clustering as follows.

Problem 2 (*Correlation Clustering*). Given a set of objects V, and distances $X_{uv} \in [0, 1]$ for all pairs $u, v \in V$, find a partition $\mathcal{C}$ for the objects in V that minimizes the score function

$$d(\mathcal{C}) = \sum_{\substack{(u,v)\\ \mathcal{C}(u)=\mathcal{C}(v)}} X_{uv} + \sum_{\substack{(u,v)\\ \mathcal{C}(u)\neq\mathcal{C}(v)}} (1 - X_{uv}).$$

Correlation clustering is a generalization of clustering aggregation. Given the m clusterings $\mathcal{C}_1, \ldots, \mathcal{C}_m$ as input, one can construct an instance of the correlation clustering problem by defining the distances X_{uv} appropriately. Let $X_{uv} = \frac{1}{m} \cdot |\{i \mid 1 \leq i \leq m \text{ and } \mathcal{C}_i(u) \neq \mathcal{C}_i(v)\}|$ be the fraction of clusterings that assign the pair (u, v) into different clusters. For a candidate solution $\mathcal{C}$ of correlation clustering, if $\mathcal{C}$ places u, v in the same cluster, it will disagree with mX_{uv} of the original clusterings, while if $\mathcal{C}$ places u, v in different clusters, it will disagree with the remaining $m(1 - X_{uv})$ clusterings. Thus, for any clustering $\mathcal{C}$, we have $m \cdot d(\mathcal{C}) = \sum_{i=1}^{m} d_V(\mathcal{C}, \mathcal{C}_i) = D(\mathcal{C})$, showing that clustering aggregation reduces to correlation clustering. We note that an instance of correlation clustering produced by an instance of clustering aggregation is a restricted version of the correlation clustering problem.

It is easy to show that the values X_{uv} obey the triangle inequality.

OBSERVATION 2. *For all u, v and w in V, we have that* $X_{uw} \leq X_{uv} + X_{vw}$.

PROOF. Define the indicator function X_{uv}^i, such that $X_{uv}^i = 1$ if $\mathcal{C}_i(u) \neq \mathcal{C}_i(v)$ and zero otherwise. Then $X_{uv} = \frac{1}{m} \sum_{i=1}^{m} X_{uv}^i$. Therefore, it suffices to show that $X_{uw}^i \leq X_{uv}^i + X_{vw}^i$. The only way that this inequality can be violated is if $X_{uw}^i = 1$ and $X_{uv}^i = X_{vw}^i = 0$. However, the latter equality suggests that u, v, w are all placed in the same cluster, thus reaching a contradiction. □

In the weighted case, the distance on the edge (u, v) would be

$$X_{uv} = \frac{\sum_i^m w_i \cdot X_{uv}^i}{\sum_i^m w_i}.$$

Our expositions hold for the weighted case as well. However, for simplicity in the following, we consider only the case where all weights are equal.

Bansal et al. [2004] already showed that correlation clustering is NP-hard. The previous reduction from clustering aggregation to correlation clustering shows that even our version of correlation clustering, which corresponds to the weighted correlation clustering problem where the $+$ and $-$ edge weights sum to 1 and satisfy the triangle inequality, is NP-hard. Since both problems we consider are NP-hard, it is natural to seek algorithms with provable approximation guarantees. For the clustering aggregation problem, it is easy to obtain a 2-approximation solution. The idea is to take advantage of the triangle inequality property of the distance measure $d_V(\cdot, \cdot)$. Assume that we are given m objects in a metric space and we want to find a new object that minimizes the sum of distances from the given objects. Then it is a well-known fact that selecting the best among the m original objects yields a factor $2(1-1/m)$ approximate solution. For our problem, this method suggests taking as the solution to clustering aggregation the clustering $\mathcal{C}_i$ that minimizes $D(\mathcal{C}_i)$. Despite the small approximation factor, this solution is nonintuitive, and we observed that it does not work well in practice. Furthermore, the previous algorithm cannot be used for the problem of correlation clustering; there are no input clusterings to choose from. In general, the correlation clustering problem we consider is not equivalent to the clustering aggregation problem.

5. ALGORITHMS

5.1 Description of the Algorithms

In this section, we present several algorithms for clustering aggregation. Most of our algorithms approach the problem through the correlation clustering problem and most of the algorithms are parameter-free.

The BESTCLUSTERING *Algorithm.* This is the simple algorithm that was mentioned in the previous section. Given m clusterings $\mathcal{C}_1, \ldots, \mathcal{C}_m$, BESTCLUSTERING finds the input clustering $\mathcal{C}_i$ that minimizes the total number of disagreements $D(\mathcal{C}_i)$. Using the data structures described in Barthelemy and Leclerc [1995] or techniques similar to those described in Mielikäinen et al. [2006] the best clustering can be found in time $O(m^2 n)$. As discussed, this algorithm yields a solution with an approximation ratio at most $2(1-1/m)$. In Section 6, we show that this bound is tight, that is, there exists an instance of the clustering aggregation problem where the algorithm BESTCLUSTERING produces a solution of cost exactly $2(1-1/m)$ times the cost of the optimal solution.

The algorithm is specific to clustering aggregation—it cannot be used for correlation clustering. In fact, it is not always possible to construct a clustering aggregation instance that gives rise to the given correlation clustering instance. Any metric X_{uv} that arises out of clustering aggregation is a convex combination of cut metrics and is, therefore, an L_1 metric (see Deza and Laurent [1997]). Thus, a metric X_{uv} that is not an L_1 metric cannot be represented by a clustering aggregation instance.

The BALLS *Algorithm.* The BALLS algorithm is inspired by the algorithm in Charikar et al. [2003], and it works on the correlation clustering problem. It takes as input the matrix of pairwise distances X_{uv}. Equivalently, we view the input as a graph whose vertices are the tuples of a dataset, and the edges are weighted by the distances X_{uv}. The algorithm is defined with an input parameter α, and it is the only algorithm that requires an input parameter. Following the theoretical analysis in Section 6, we can set α to a constant that guarantees a constant approximation ratio. However, different values of α can lead to better solutions in practice.

The intuition of the algorithm is to find a set of vertices that are close to each other and far from other vertices. Given such a set, we consider it to be a cluster, we remove it from the graph, and we proceed with the rest of the vertices. The difficulty lies in finding such a set since in principle any subset of the vertices can be a candidate. We overcome the difficulty by resorting again to the triangle inequality, this time for the distances X_{uv}. In order to find a good cluster, we take all vertices that are close (within a ball) to a vertex u. The triangle inequality guarantees that if two vertices are close to u, then they are also relatively close to each other. We also note that for the correlation clustering problem, it is intuitive that good clusters should be ball-shaped: since our cost function penalizes for long edges that are not cut, we do not expect to have elongated clusters in the optimal solution.

More formally the algorithm is described as follows. It first sorts the vertices in increasing order of the total weight of the edges incident on each vertex. This is a heuristic that we observed to work well in practice. The ordering does not affect the approximation guarantee of the algorithm. At every step, the algorithm picks the first unclustered node u in that ordering. It then finds the set of nodes B that are at a distance of at most 1/2 from the node u, and it computes the average distance $d(u, B)$ of the nodes in B to node u. If $d(u, B) \leq \alpha$, then the nodes in $B \cup \{u\}$ are considered to form a cluster; otherwise, node u forms a singleton cluster.

We can prove that, when setting $\alpha = \frac{1}{4}$, the cost of a solution produced by the BALLS algorithm is guaranteed to be at most 3 times the cost of the optimal clustering. The proof appears in Section 6. In our experiments, we have observed that the value $\frac{1}{4}$ tends to be small as it creates many singleton clusters. For many of our real datasets, we have found that $\alpha = \frac{2}{5}$ leads to better solutions. The complexity of the algorithm is $O(mn^2)$ for generating the table and $O(n^2)$ for running the algorithm.

The AGGLOMERATIVE *Algorithm.* The AGGLOMERATIVE algorithm is a standard bottom-up procedure for the correlation clustering problem. It starts by placing every node into a singleton cluster. It then proceeds by considering the pair of clusters with the smallest average distance. The average distance between two clusters is defined as the average weight of the edges between the two clusters. If the average distance of the closest pair of clusters is less than 1/2, then the two clusters are merged into a single cluster. If there are no two clusters with average distance smaller than 1/2, then no merging of current clusters can lead to a solution with improved cost $d(\mathcal{C})$. Thus, the algorithm stops, and it outputs the clusters it has created so far.

The AGGLOMERATIVE algorithm has the desirable feature that it creates clusters where the average distance of any pair of nodes is at most 1/2. The intuition is that the opinion of the majority is respected on average. Using this property, we are able to prove that when $m = 3$, the AGGLOMERATIVE algorithm produces a solution with cost at most 2 times that of the optimal solution. The proof appears in Section 6. The complexity of the algorithm is $O(mn^2)$ for creating the matrix plus $O(n^2 \log n)$ for running the algorithm.

The FURTHEST *Algorithm.* The FURTHEST algorithm is a top-down algorithm that works on the correlation clustering problem. It is inspired by the furthest-first traversal algorithm, for which Hochbaum and Shmoys [1985] showed that it achieves a 2-approximation for the clustering formulation of p-centers. As the BALLS algorithm uses a notion of a center to find clusters and repeatedly remove them from the graph, the FURTHEST algorithm uses centers to partition the graph in a top-down fashion.

The algorithm starts by placing all nodes into a single cluster. Then it finds the pair of nodes that are furthest apart and places them into different clusters. These two nodes become the centers of the clusters. The remaining nodes are assigned to the center that incurs the least cost. This procedure is repeated iteratively: at each step, a new center is generated that is the furthest from the existing centers, and the nodes are assigned to the center that incurs the least cost. At the end of each step, the cost of the new solution is computed. If it is lower than that of the previous step, then the algorithm continues. Otherwise, the algorithm outputs the solution computed in the previous step. The complexity of the algorithm is $O(mn^2)$ for creating the matrix and $O(k^2n)$ for running the algorithm where k is the number of clusters created.

The LOCALSEARCH *Algorithm.* The LOCALSEARCH algorithm is an application of a local search heuristic to the problem of correlation clustering. The algorithm starts with some clustering of the nodes. This clustering could be a random partition of the data or it could be obtained by running one of the algorithms we have already described. The algorithm then goes through the nodes, and it considers placing them into a different cluster or creating a new singleton cluster with this node. The node is placed in the cluster that yields the minimum cost. The process is iterated until there is no move that can improve the cost. The LOCALSEARCH can be used as a clustering algorithm, but also as a postprocessing step to improve upon an existing solution.

When considering a node v, the cost $d(v, C_i)$ of assigning a node v to a cluster C_i is computed as follows.

$$d(v, C_i) = \sum_{u \in C_i} X_{vu} + \sum_{u \in \overline{C_i}} (1 - X_{vu}).$$

The first term is the cost of merging v in C_i, while the second term is the cost of not merging node v with the nodes not in C_i. We compute $d(v, C_i)$

efficiently as follows. For every cluster C_i, we compute and store the cost $M(v, C_i) = \sum_{u \in C_i} X_{vu}$ and the size of the cluster $|C_i|$. Then the distance of v to C_i is

$$d(v, C_i) = M(v, C_i) + \sum_{j \neq i} (|C_j| - M(v, C_j)).$$

The cost of assigning node v to a singleton cluster is $\sum_j (|C_j| - M(v, C_j))$.

The running time of the LocalSearch algorithm, given the distance matrix X_{uv}, is $O(Tn^2)$, where T is the number of local search iterations until the algorithm converges to a solution for which no better move can be found. Our experiments showed that the LocalSearch algorithm is quite effective, and it improves the solutions found by the previous algorithms significantly. Unfortunately, the number of iterations tends to be large, and thus the algorithm is not scalable to large datasets.

5.2 Handling Large Datasets

The algorithms we described in Section 5.1 take as input the distance matrix so their complexity is quadratic in the number of data objects in the dataset. The quadratic complexity is inherent in the correlation clustering problem since the input to the problem is a complete graph. Given a node, the decision of placing the node to a cluster has to take into account not only the cost of merging the node to the cluster, but also the cost of not placing the node to the other clusters. Furthermore, the definition of the cost function does not allow for an easy summarization of the clusters, a technique that is commonly used in many clustering algorithms. However, the quadratic complexity makes the algorithms inapplicable to large datasets. We will now describe the algorithm Sampling, which uses sampling to reduce the running time of the algorithms.

The Sampling algorithm is run on top of the algorithms we described Section 5. The algorithm performs a preprocessing and postprocessing step that is linear in the size of the dataset. In the preprocessing step, the algorithm samples a set of nodes, S, uniformly at random from the dataset. These nodes are given as input to one of the clustering aggregation algorithms. The output is a set of ℓ clusters $\{C_1, ..., C_\ell\}$ of the nodes in S. In the postprocessing step, the algorithm goes through the nodes in the dataset not in S. For every node, it decides whether or not to place it in one of the existing clusters or to create a singleton cluster. In order to perform this step efficiently, we use the same technique as for the LocalSearch algorithm. We observed experimentally that at the end of the assignment phase there are too many singleton clusters. Therefore, we collect all singleton clusters, and we run the clustering aggregation again on this subset of nodes.

The size of the sample S is determined so that for all large clusters in the dataset the sample will contain at least one node from each such larger cluster with high probability. Large cluster means a cluster that contains a constant fraction of the nodes in the dataset. Using the Chernoff bounds, we can prove that sampling $O(\log n)$ nodes is sufficient to ensure that we will select at least one of the nodes in a large cluster with high probability. Note that although nodes in small clusters may not be selected, these will be assigned in singleton clusters in the postprocessing step. When clustering the singletons, they are likely to be clustered together. Since the size of these clusters is small, this does not incur a significant overhead in the cost of the algorithm.

6. THEORETICAL ANALYSIS

In this section, we consider some of the algorithms described in Section 5, and we prove guarantees for the cost of the solution they produce with respect to the cost of the optimal solution. For any algorithm $\textsc{Alg}$, let $\textsc{Alg}(I)$ denote the cost of the algorithm $\textsc{Alg}$ on input I. Also let $\textsc{Opt}(I)$ denote the cost of the optimal solution on input I. Let $|I|$ denote the length of the input. Define $\mathcal{I}$ to be the set of all possible inputs to $\textsc{Alg}$. We say that the algorithm $\textsc{Alg}$ has approximation ratio $R(\textsc{Alg}, |I|)$ if, for all $I \in \mathcal{I}$, it holds that

$$\textsc{Alg}(I) \leq R(\textsc{Alg}, |I|) \cdot \textsc{Opt}(I).$$

For simplicity, we will usually use $R(\textsc{Alg})$ to denote the approximation ratio of $\textsc{Alg}$. We are interested in bounding $R(\textsc{Alg})$ for the different algorithms.

6.1 The BestClustering Algorithm

The BestClustering algorithm is an approximation algorithm for Problem 1, the clustering aggregation problem. The input I is a set of m clusterings of n points. The cost function $D(\mathcal{C})$ is the number of disagreements of the output clustering $\mathcal{C}$ with the input clusterings. We know that $R(\textsc{BestClustering}) \leq 2(1 - \frac{1}{m})$. We will now prove that this bound is tight.

Theorem 1. *The* BestClustering *algorithm has approximation ratio* $R(\textsc{BestClustering}) \geq 2(1 - \frac{1}{m})$ *for Problem 1.*

Proof. In order to prove the lower bound to the approximation ratio of BestClustering it suffices to construct an instance I of the clustering aggregation problem such that $\frac{\textsc{BestClustering}(I)}{\textsc{Opt}(I)} = 2(1 - \frac{1}{m})$.

Let V be the set of objects and let n denote the size of the set V. We take $n = km$, where $k \geq 2$ is an integer, and we construct m clusterings $\mathcal{C}_1, \ldots, \mathcal{C}_m$ on V as follows. We partition (arbitrarily) the set V into m subsets $V_1, V_2, \ldots, V_m$ of equal size. The clustering $\mathcal{C}_i$ assigns the elements of V_i into singleton clusters, while it groups the elements of each set V_j, $j \neq i$, into a single cluster. Formally, the clustering $\mathcal{C}_i$ assigns distinct labels to all elements of the subset V_i, that is, $\mathcal{C}_i(u) \neq \mathcal{C}_i(v)$, for all $u, v \in V_i$. It assigns the same label to all elements in subset V_j, for all $j \neq i$, that is, $\mathcal{C}_i(u) = \mathcal{C}_i(v)$ for all $u, v \in V_j$. Furthermore, for all $j \neq k$, $\mathcal{C}_i(u) \neq \mathcal{C}_i(v)$ for all $u \in V_j$ and $v \in V_k$.

Due to the symmetry in the definition of subsets $V_1, V_2, \ldots, V_m$ and the clusterings $\mathcal{C}_1, \ldots, \mathcal{C}_m$, selecting any clustering $\mathcal{C}_i$ gives the same number of disagreements $D(\mathcal{C}_i)$. Specifically,

$$D(\mathcal{C}_i) = (m-1)\binom{n/m}{2} + (m-1)\binom{n/m}{2} = 2(m-1)\binom{n/m}{2}.$$

The first $(m-1)\binom{n/m}{2}$ term is due to the elements of the set V_i. The clustering $\mathcal{C}_i$ assigns a different label to each element in V_i, while each of the other $m - 1$

clusterings assigns the same label to all elements in V_i. There are $\binom{n/m}{2}$ pairs, and each of them contributes a disagreement between cluster $\mathcal{C}_i$ and each of the $m-1$ other clusters.

The second $(m-1)\binom{n/m}{2}$ term appears due to the remaining $m-1$ subsets. For each such subset V_j, the clustering $\mathcal{C}_i$ assigns the same label to all elements in V_j. All other clusterings, except for clustering C_j, do exactly the same, in agreement with C_i. Clustering $\mathcal{C}_j$ assigns distinct labels to all the elements of V_j, generating one disagreement for each pair of elements.

Let $\mathcal{C}^*$ denote the clustering produced by the optimal algorithm. Clustering $\mathcal{C}^*$ creates a cluster for each subset V_i. The total number of disagreements is $D(\mathcal{C}^*) = m\binom{n/m}{2}$. Therefore, $D(\mathcal{C}_i) = 2(1-\frac{1}{m})D(\mathcal{C}^*)$, and $\frac{\text{BESTCLUSTERING}(I)}{\text{OPT}(I)} = 2(1-\frac{1}{m})$. □

6.2 The BALLS Algorithm

The BALLS algorithm is an approximation algorithm for Problem 2, the correlation clustering problem. The input I to the problem is a set of n points and the pairwise distances X_{uv}. The cost function is $d(\mathcal{C})$ defined in Section 4. We will prove that the approximation ratio of the algorithm is bounded by a constant.

We first prove the following general lemma.

LEMMA 1. *For any algorithm* ALG *and any pair of objects u and v, if*

(a) $X_{uv} \leq c$ *and* ALG *assigns u and v in the same cluster, or*
(b) $X_{uv} \geq 1-c$ *and* ALG *assigns u and v in different clusters,*

then the cost paid by ALG *on edge (u, v) is at most $\frac{c}{1-c}$ times the cost of the optimal algorithm for (u, v).*

PROOF. In both case (a) and (b), the algorithm ALG pays at most c. If the optimal takes the same decision as ALG, then it pays the same cost. If the optimal takes the opposite decision, then it pays at least $1-c$, hence the ratio $\frac{c}{1-c}$. □

As an obvious corollary, if $X_{uv} \leq 1/2$ and an algorithm assigns u and v to the same cluster, or if $X_{uv} \geq 1/2$ and an algorithm assigns u and v to different clusters, then the algorithm cannot do worse than the optimal on (u, v).

We are now ready to prove the following theorem. Our proof follows along the lines of the analysis in Charikar et al. [2003].

THEOREM 2. *The* BALLS *algorithm has approximation ratio* $\max\{\frac{1-\alpha}{\alpha}, \frac{1+2\alpha}{1-2\alpha}, \frac{2-2\alpha}{1-2\alpha}\}$.
For $\alpha = \frac{1}{4}$, the algorithm achieves an approximation ratio of 3.

PROOF. We analyze the algorithm by bounding the cost that the algorithm pays for each edge in terms of the cost that the optimal algorithm pays for the same edge. Consider an iteration of the BALLS algorithm, and let u be the node selected to be the center of the ball. We now consider the following cases.

Singleton clusters. First, we consider the case that $C = \{u\}$ is selected to be a singleton cluster. Recall that B is the set of nodes that are within distance

$1/2$ from u and that in this case the average distance $d(u, B)$ of the nodes in B to u is more than α. For all edges (u, i) with $i \notin B$, we have $X_{ui} \geq 1/2$. Since the algorithm separates u from i, the cost of the optimal cannot be less on each (u, i). The algorithm also splits all edges (u, i) with $i \in B$, so the cost of the algorithm is

$$\sum_{i \in B}(1 - X_{ui}) = |B| - \sum_{i \in B} X_{ui} \leq (1 - \alpha)|B|,$$

where the fact $\sum_{i \in B} X_{ui} \geq \alpha|B|$ follows from the fact that the algorithm chose $\{u\}$ to be a singleton cluster. On the other hand, the optimal algorithm might choose to place u in the same cluster with some vertices $M \subseteq B$. Thus the cost of the optimal for the edges from u to the set B is

$$\sum_{i \in M} X_{ui} + \sum_{i \in B \setminus M} (1 - X_{ui}) \geq \sum_{i \in B} X_{ui} \geq \alpha|B|,$$

where we used the fact that since $i \in B$, we have that $X_{ui} \leq 1/2$, and thus $1 - X_{ui} \geq X_{ui}$. As a result, the approximation ratio achieved on edges incident to the singleton clusters is at most $R_1 = \frac{(1-\alpha)|B|}{\alpha|B|} = \frac{1-\alpha}{\alpha}$.

Next, we analyze the case where the BALLS algorithm creates the cluster $C = B \cup \{u\}$. Such a cluster C is created when $d(u, B) \leq \alpha$.

Edges within clusters. For the edges of type (u, i) with $i \in B$ that the algorithm places in the cluster C, we have $X_{ui} \leq 1/2$, so the optimal cannot improve the cost by splitting those edges.

The other type of edges within the cluster $C = B \cup \{u\}$ are edges (i, j) with $i, j \in B$. We order the vertices $i \in B$ in order of increasing distance X_{ui} from the node u. For a fixed j, we will bound the cost of the edges (i, j) for $i < j$.

If $X_{uj} \leq \beta$ for a constant $\beta < 1/2$ to be specified later, by the triangle inequality, for all $i < j$, we have that $X_{ij} \leq X_{ui} + X_{uj} \leq 2\beta$. Therefore, by Lemma 1, the approximation ratio for those edges is at most $R_2 = \frac{2\beta}{1-2\beta}$.

If $X_{uj} > \beta$, let C_j be the set of vertices i with $i < j$. Notice that since the average distance from u to the vertices in B is less than α, the average distance from u to vertices in C_j is also less than α since C_j contains a prefix from the list of vertices ordered in ascending order of their distance X_{ui} from node u. The cost of the algorithm for the edges (i, j) where i is in C_j is

$$A_j = \sum_{i \in C_j} X_{ij} \leq \sum_{i \in C_j} X_{uj} + \sum_{i \in C_j} X_{ui} \leq \left(\frac{1}{2} + \alpha\right)|C_j|.$$

On the other hand, assume that the optimal algorithm places some vertices $i \in M_j \subseteq C_j$ in the same cluster with j, and the rest of the vertices $i \in S_j = C_j \setminus M_j$ in different clusters than j; thus $|C_j| = |M_j| + |S_j|$. The cost of the optimal algorithm for the edges (i, j) where i is in C_j can now be written as

$$\begin{aligned} OPT_j &= \sum_{i \in M_j} X_{ij} + \sum_{i \in S_j}(1 - X_{ij}) \\ &\geq \sum_{i \in M_j}(X_{uj} - X_{ui}) + \sum_{i \in S_j}(1 - X_{uj} - X_{ui}) \end{aligned}$$

$$
\begin{aligned}
&= (|M_j| - |S_j|)X_{uj} + |S_j| - \sum_{i \in C_j} X_{ui} \\
&\geq (|M_j| - |S_j|)X_{uj} + |S_j| - \alpha|C_j| \\
&= (|M_j| - |S_j|)X_{uj} + |S_j| - \alpha(|M_j| + |S_j|)
\end{aligned}
$$

We now have two cases.

—If $|M_j| < |S_j|$, we use the fact that $X_{uj} \leq 1/2$ or equivalently $(|M_j| - |S_j|)X_{uj} \geq (|M_j| - |S_j|)/2$ and so the cost of the optimal is $OPT_j \geq (|M_j| - |S_j|)/2 + |S_j| - \alpha(|M_j| + |S_j|) = (\frac{1}{2} - \alpha)(|M_j| + |S_j|) = (\frac{1}{2} - \alpha)|C_j|$. In this case, the approximation factor is at most $R_3 = \frac{\frac{1}{2}+\alpha}{\frac{1}{2}-\alpha} = \frac{1+2\alpha}{1-2\alpha}$.

—If $|M_j| \geq |S_j|$, we use the fact that $X_{uj} \geq \beta$, implying that the cost of the optimal is $OPT_j \geq \beta(|M_j| - |S_j|) + |S_j| - \alpha(|M_j| + |S_j|) = (\beta - \alpha)|M_j| + (1 - \beta - \alpha)|S_j|$. Selecting $\beta \geq \alpha$, we have that $OPT_j \geq (1 - 2\alpha)|S_j|$.

We now consider difference $A_j - OPT_j$. We have that

$$
\begin{aligned}
A_j - OPT_j &= \sum_{i \in C_j} X_{ij} - \left(\sum_{i \in M_j} X_{ij} + \sum_{i \in S_j} (1 - X_{ij}) \right) \\
&= \sum_{i \in S_j} X_{ij} - \sum_{i \in S_j} (1 - X_{ij}) = 2 \sum_{i \in S_j} X_{ij} - |S_j| \\
&\leq 2 \sum_{i \in S_j} 1 - |S_j| = |S_j|,
\end{aligned}
$$

where the last inequality follows from the fact that $X_{ij} \leq 1$ for all edges (i, j). We now look at the ratio $\frac{A_j - OPT_j}{OPT_j}$. We have that

$$
\frac{A_j - OPT_j}{OPT_j} \leq \frac{|S_j|}{(1 - 2\alpha)|S_j|} = \frac{1}{(1 - 2\alpha)},
$$

and therefore,

$$
\frac{A_j}{OPT_j} \leq \frac{2 - 2\alpha}{1 - 2\alpha}.
$$

In this case, the approximation factor is at most $R_4 = \frac{2-2\alpha}{1-2\alpha}$.

Note that R_2 is an increasing function of β. Since $\beta \geq \alpha$, it takes its minimum value for $\beta = \alpha$ which is $R_2 = \frac{2\alpha}{1-2\alpha}$. We also have that $R_2 \leq R_3$, and $R_2 \leq R_4$ for all $\alpha \in (0, 1/2)$. Therefore, the approximation ratio for the edges within a cluster is at most $\max\{R_3, R_4\}$.

Edges across clusters. Finally, we have to bound the cost of edges going from inside C to clusters outside C. For edges of the type (u, i) with $i \notin C$, we have that $X_{ui} > 1/2$ and the algorithm splits those edges so the optimal cannot perform better on any one of those edges. Therefore, we concentrate on edges of the type (i, j) with $i \in C$ and $j \notin C$. In particular, $X_{ui} \leq 1/2$ and $X_{uj} > 1/2$. If $X_{uj} \geq \gamma$ for a constant $\gamma > 1/2$ to be specified later, we have that $X_{ij} \geq X_{uj} - X_{ui} \geq \gamma - 1/2$, so, from Lemma 1, the approximation ratio on those edges will be at most $R_5 = \frac{1-(\gamma-1/2)}{\gamma-1/2} = \frac{3/2-\gamma}{\gamma-1/2}$.

In the remaining case $1/2 < X_{uj} < \gamma$, we proceed by fixing j and bounding the cost of all edges (i, j) for $i \in C$. For some fixed j, assume that the optimal algorithm places some vertices $i \in M_j \subseteq C$ in the same cluster with j, and the rest of the vertices $i \in S_j = C \setminus M_j$ in different clusters than j. Again $|C| = |M_j| + |S_j|$. The cost of the algorithm for all edges (i, j) with $i \in C$ is

$$A_j = \sum_{i\in C}(1-X_{ij}) \leq \sum_{i\in C}(1-(X_{uj}-X_{ui})) \leq \sum_{i\in C}(1-X_{uj})+\sum_{i\in C}X_{ui} \leq \left(\frac{1}{2}+\alpha\right)|C|.$$

The cost of the optimal is bounded from below exactly as in the previous case, that is, $OPT_j \geq (|M_j| - |S_j|)X_{uj} + |S_j| - \alpha(|M_j| + |S_j|)$. If $|M_j| \geq |S_j|$. We use the fact that $X_{uj} > 1/2$, so the cost of the optimal is $OPT_j \geq (\frac{1}{2} - \alpha)|C|$, and the approximation ratio is again R_3.

If $|M_j| < |S_j|$, we use the fact that $X_{uj} < \gamma$, and therefore $OPT_j \geq \gamma(|M_j| - |S_j|) + |S_j| - \alpha(|M_j| + |S_j|) = (\gamma - \alpha)|M_j| + (1 - \gamma - \alpha)|S_j|$. Selecting $\gamma \leq 1 - \alpha$, we have that $OPT_j \geq (1 - 2\alpha)|M_j|$. We consider again the difference $A_j - OPT_j$. We have that

$$\begin{aligned}
A_j - OPT_j &= \sum_{i\in C}(1 - X_{ij}) - \left(\sum_{i\in M_j} X_{ij} + \sum_{i\in S_j}(1 - X_{ij})\right)\\
&= \sum_{i\in M_j}(1 - X_{ij}) - \sum_{i\in M_j} X_{ij} = \sum_{i\in M_j}(1 - 2X_{ij})\\
&\leq |M_j| - 2\sum_{i\in M_j}(X_{uj} - X_{ui})\\
&\leq |M_j| - 2\sum_{i\in M_j} X_{uj} + 2\sum_{i\in M_j} X_{ui}\\
&\leq |M_j| - |M_j| + |M_j| = |M_j|,
\end{aligned}$$

where the last inequality follows from the fact that $X_{ui} \leq 1/2$ and $X_{uj} > 1/2$. Similar to before, we obtain that

$$\frac{A_j}{OPT_j} \leq \frac{2 - 2\alpha}{1 - 2\alpha}.$$

Therefore, the approximation ration in this case is again at most R_4.

We note that the ratio R_5 is a decreasing function of γ. Since we select $\gamma \leq 1 - \alpha$, R_5 takes its minimum value for $\gamma = 1 - \alpha$, which is $R_5 = \frac{1/2+\alpha}{1/2-\alpha} = R_3$.

Bringing it all together. The overall approximation ratio of the BALLS algorithm is $R(\text{BALLS}) \leq \max\{R_1, R_3, R_4\}$. The ratios R_1, R_3, and R_4 are functions of the parameter α. We have that $0 \leq \alpha \leq \frac{1}{2}$, and that R_1 is a decreasing function of α, while R_3 and R_4 are increasing functions of α. For $\alpha = \frac{1}{4}$, the values of all three ratios agree to the value 3. Therefore, we conclude that the approximation ratio of the BALLS algorithm is at most 3. □

There are special cases where the BALLS algorithm can perform better. Consider an instance of the correlation clustering problem that is derived when we consider the aggregation of three clusterings. In this case, the weights X_{uv}

take values in the set $\{0, \frac{1}{3}, \frac{2}{3}, 1\}$. Then, for $\frac{1}{3} \leq \alpha \leq \frac{1}{2}$, the BALLS algorithm achieves approximation ratio 2. This is due to the fact that in a ball B of radius $\frac{1}{2}$ centered at some node u, there are only nodes that are at distance 0 or $\frac{1}{3}$ from node u. Selecting α such that $\frac{1}{3} \leq \alpha \leq \frac{1}{2}$, we force the algorithm to always create a cluster with the nodes in B. Therefore, by definition, the algorithm takes all edges with weight 0 and breaks all edges with weight 1. For the remaining edges, Lemma 1 guarantees that the approximation ratio is at most 2.

However, this is not so interesting since even the simple algorithm that merges only the edges of weight zero achieves the same approximation ratio. In general, assume that the X_{uv} satisfy the following property: there exists a value $\frac{1}{2} \leq c \leq 1$ such that if $X_{uv} \notin \{0, 1\}$, then $X_{uv} \in (1-c, c)$. We say that the values X_{uv} are *symmetrically bounded* by c. In this case, Lemma 1 guarantees that the approximation ratio of any algorithm that merges all the zero weight edges and none of the weight one edges is at at most $c/(1-c)$. For $c < \frac{3}{4}$, we guarantee that the approximation ratio is strictly less than 3. Note that for correlation clustering problems derived from clustering aggregation instances, the distance values are always symmetrically bounded by $(m-1)/m$, yielding an approximation ratio $m-1$.

6.3 The AGGLOMERATIVE Algorithm

For the AGGLOMERATIVE algorithm, we can prove that for all input instances such that the X_{uv} values are symmetrically bounded by c, the algorithm achieves approximation ratio $c/(1-c)$. Although, as we noted before, this is achieved by any algorithm that merges all edges of weight zero and splits all edges of weight one, it is not clear that the AGGLOMERATIVE algorithm satisfies this requirement since we cannot guarantee that the algorithm will not merge any edge of weight one. We can prove the following theorem.

THEOREM 3. *Assume that for all input instances $I \in \mathcal{I}$ for the correlation clustering problem, the values X_{uv} are symmetrically bounded by c for $\frac{1}{2} \leq c < 1$. Then the* AGGLOMERATIVE *algorithm has approximation ratio* $R(\text{AGGLOMERATIVE}) \leq \frac{c}{1-c}$.

PROOF. We are going to bound the approximation ratio of AGGLOMERATIVE by bounding the approximation ratio achieved at each step of the algorithm. First, we note that in the first step, the AGGLOMERATIVE algorithm will merge all edges of weight zero. This is due to the fact that edges with zero weight appear always in cliques (once again due to triangle inequality). Obviously no approximation error is induced for these merges, thus we can examine the behavior of the algorithm after these merges have been completed. Therefore, we can assume that all subsequent merges involve only edges with weight greater than zero.

Consider now one step of the algorithm after all zero-weight edges have been merged, and let k be the number of edges merged by the algorithm at that step. Let $\{c_1, c_2, \ldots, c_p\}$ denote the set of all distinct weights that these edges take ($p \leq k$) in decreasing order. Assume that $c_1 = 1$, otherwise Lemma 1 provides the upper bound to the approximation ratio.[1] We have that $c_p > 0$, since we have assumed that all edges of weight zero have already been merged. Let k_i denote the number of edges of weight c_i. The cost paid by the agglomerative algorithm is $A = k_1 + k_2c_2 + \cdots + k_pc_p$. From the definition of the algorithm, we have that

$$\frac{k_1 + k_2c_2 + \cdots + k_pc_p}{k_1 + \cdots + k_p} \leq \frac{1}{2}.$$

Solving for k_1, we obtain

$$k_1 \leq (1 - 2c_2)k_2 + (1 - 2c_3)k_3 + \cdots + (1 - 2c_p)k_p$$

Therefore,

$$A \leq (1 - c_2)k_2 + (1 - c_3)k_3 + \cdots + (1 - c_p)k_p \leq c(k_2 + k_3 + \cdots + k_p).$$

since for all weights $1 - c \leq c_i \leq c$, and thus $1 - c \leq 1 - c_i \leq c$.

Now let c_q denote the smallest of the weights such that $c_q > 1/2$. The cost of the optimal solution for these edges is at least

$$O = (1 - c_2)k_2 + \cdots (1 - c_q)k_q + c_{q+1}k_{q+1} + \cdots c_pk_p \geq (1 - c)(k_2 + k_3 + \cdots + k_p).$$

Therefore, the approximation ratio is $A/O \leq c/(1 - c)$ which concludes the proof. □

For correlation clustering problems that arise from clustering aggregation problem instances, Theorem 3 guarantees that, when merging m clusterings, the AGGLOMERATIVE algorithm has an approximation ratio of at most $m - 1$.

7. EXPERIMENTAL EVALUATION

We have conducted extensive experiments to test the quality of the clusterings produced by our algorithms on a varied collection of synthetic and real datasets. Furthermore, for our SAMPLING algorithm, we have experimented with the quality vs. efficiency trade-off.

7.1 Improving Clustering Robustness

The goal in this set of experiments is to show how clustering aggregation can be used to improve the quality and robustness of widely-used vanilla clustering algorithms. For the two experiments we are describing, we used synthetic datasets of two-dimensional points.

The first dataset is shown in Figure 3. An intuitively good clustering for this dataset consists of the seven perceptually distinct groups of points. We ran five different clustering algorithms implemented in MATLAB: single linkage, complete linkage, average linkage, Ward's clustering, and k-means algorithm. The first three algorithms are agglomerative bottom-up algorithms that merge pairs of clusters, based on their minimum, maximum, and average distance, respectively. Ward's clustering algorithm is also an agglomerative bottom-up

[1] There are cases where a better approximation ratio may be proven when $c_p < 1$, for example, when X_{uv} takes values from the set $\{1/3, 2/3\}$.

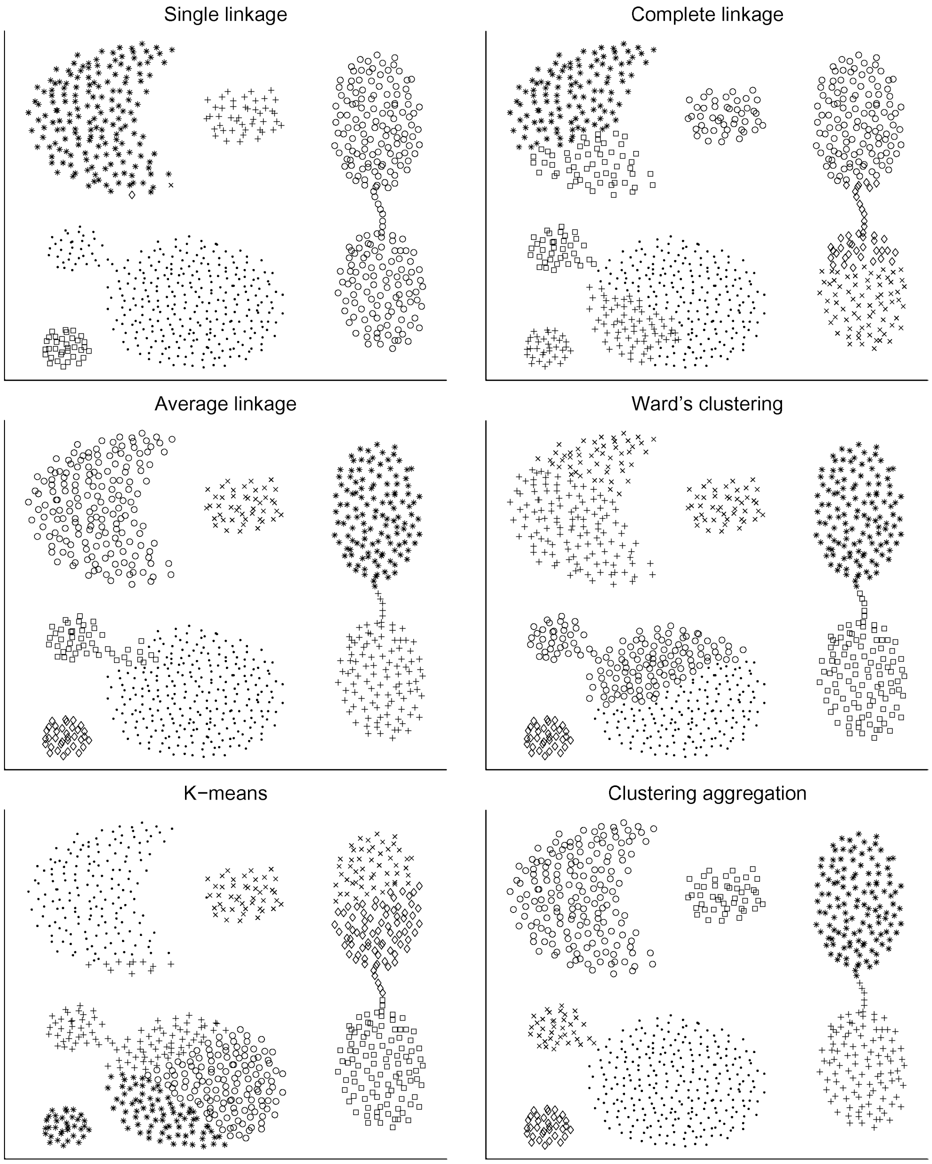


Fig. 3. Clustering aggregation on five different input clusterings. To obtain the last plot, which is the result of aggregating the previous five plots, the AGGLOMERATIVE algorithm was used.

algorithm whose merging criterion is to select the pair of clusters that minimize the sum of the square of distances from each point to the mean of the two clusters. Finally, the *k*-means algorithm is the popular iterative clustering method which is also known as Lloyd's algorithm.

For all of the clusterings, we set the number of clusters to be 7, and for the other parameters, if any, we used MATLAB's defaults. The results for the five clusterings are shown in the first five panels of Figure 3. One sees that all clusterings are imperfect. In fact, the dataset contains features that are known to create difficulties for the selected algorithms such as, narrow bridges between clusters, uneven-sized clusters, etc. The last panel in Figure 3 shows the results of aggregating the five previous clusterings. The aggregated clustering is better than any of the input clusterings (although average linkage comes very close), and it confirms our intuition of how mistakes in the input clusterings can be canceled out.

In our second experiment, the goal is to show how clustering aggregation can be used to identify the correct clusters as well as outliers. Three datasets were created as follows: k^* cluster centers were selected uniformly at random in the unit square, and 100 points were generated from the normal distribution around each cluster center. For the three datasets, we used $k^* = 3$, 5, and 7, respectively. An additional 20% of the total number of points were generated uniformly from the unit square and they were added in the datasets. For each of the three datasets, we ran the k-means algorithm with $k = 2, 3, \ldots, 10$, and we aggregated the resulting clusterings, that is, in each dataset, we performed clustering aggregation on 9 input clusterings. For lack of space, the input clusterings are not shown; however, most are imperfect. Obviously, when k is too small, some clusters get merged, and, when k is too large, some clusters get split. The results of clustering aggregation for the three datasets are shown in Figure 4. We see that the main clusters identified are precisely the correct clusters. Some small additional clusters are also found that contain only points from the background noise, and they can be clearly characterized as outliers.

7.2 Clustering Categorical Data

In this section, we use the ideas we discussed in Section 2 for performing clustering of categorical datasets. We used three datasets from the UCI Repository of machine learning databases [Blake and Merz 1998]. The first dataset, `Votes`, contains voting information for 435 people. For each person, there are votes on 16 issues (`yes`/`no` vote viewed as categorical values) and a class label classifying a person as `republican` or `democrat`. There are a total of 288 missing values. The second dataset, `Mushrooms`, contains information on physical characteristics of mushrooms. There are 8,124 instances of mushrooms, each described by 22 categorical attributes such as shape, color, odor, etc. There is a class label describing if a mushroom is `poisonous` or `edible`, and there are 2,480 missing values in total. Finally, the third dataset, `Census`, has been extracted from the census bureau database, and it contains demographic information on 32,561 people in the US. There are 8 categorical attributes (such as education, occupation, marital status, etc.) and 6 numerical attributes (such as age, capital gain, etc.). Each person is classified according to whether they receive an annual salary of more than $50K or less.

For treating the missing values, we assume that, given a pair of tuples for which an attribute contains at least one missing value, the attribute tosses a random coin and, with probability $\frac{1}{2}$, it reports the tuples as being clustered together, while with probability $\frac{1}{2}$, it reports them as being in separate clusters. Each pair of tuples is treated independently. Essentially we are then

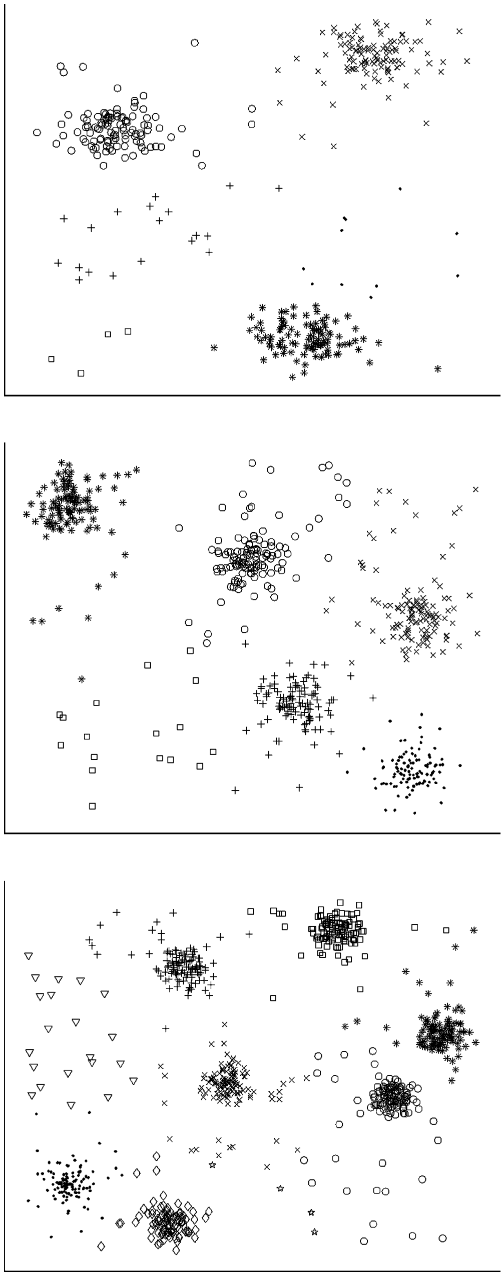


Fig. 4. Finding the correct clusters and outliers. The three figures show datasets with $k^* = 3, 5,$ and 7 clusters, and background noise. The shapes with which the data points are drawn indicate the different clusters found by the clustering aggregation method.

Table I. Results on Votes Dataset
(k is the number of clusters, I is the impurity index, and E_D is the disagreement error. The lower bound on E_D is computed by considering an algorithm that merges all edges with weight less than $\frac{1}{2}$ and splits all edges with weight greater than $\frac{1}{2}$)

Algorithm	k	$I(\%)$	E_D
Class labels	2	0	34,184
Lower bound			28,805
BestClustering	3	15.1	31,211
Agglomerative	2	14.7	30,408
Furthest	2	13.3	30,259
Balls$_{\alpha=0.4}$	2	13.3	30,181
LocalSearch	2	11.9	29,967
ROCK$_{k=2,\theta=0.73}$	2	11	32,486
LIMBO$_{k=2,\phi=0.0}$	2	11	30,147

interested in minimizing the expected number of disagreements between the clusterings.

For each of the datasets, we perform clustering based on the categorical attributes, and we evaluate the clustering using the class labels of the datasets. The intuition is that clusterings with pure clusters, that is, clusters in which all objects have the same class label, are preferable. Thus, if a clustering contains k clusters with sizes $s_1, \ldots, s_k$, and the sizes of the majority class in each cluster are $m_1, \ldots, m_k$, respectively, then we measure the quality of the clustering by an *impurity index* measure, defined as

$$I = \frac{\sum_{i=1}^{k}(s_i - m_i)}{\sum_{i=1}^{k} s_i} = \frac{\sum_{i=1}^{k}(s_i - m_i)}{n}.$$

If a clustering has I value equal to 0, it means that it contains only pure clusters. Notice that clusterings with many clusters tend to have smaller I values—in the extreme case, if $k = n$, then $I = 0$ since singleton clusters are pure. We remark that this measure is only indicative of the cluster quality. It is not clear that the best clusters in the dataset correspond to the existing classes. Depending on the application, one may be interested in discovering different clusters.

We also run comparative experiments with the categorical clustering algorithm ROCK [Guha et al. 2000] and the much more recent algorithm LIMBO [Andritsos et al. 2004]. ROCK uses the *Jaccard coefficient* to measure tuple similarity, and places a link between two tuples whose similarity exceeds a threshold θ. For our experiments, we used values of θ suggested by Guha et al. [2000] in the original ROCK paper. LIMBO uses information theoretic concepts to define clustering quality. It clusters together tuples so that the conditional entropy of the attribute values within a cluster is low. For the parameter ϕ of LIMBO, we again used values suggested in Andritsos et al. [2004]. For both algorithms, we adopt the convention of the LIMBO algorithm, and we treat missing values as separate attribute values.

The results for the Votes and Mushrooms datasets are shown in Tables I and II, respectively. In addition to the impurity index (I), we also show the

Table II. Results on Mushrooms Dataset

Algorithm	k	$I(\%)$	$E_D(\times 10^6)$
Class labels	2	0	13.537
Lower bound			8.388
BestClustering	5	35.4	8.542
Agglomerative	7	11.1	9.990
Furthest	9	10.4	10.169
Balls$_{\alpha=0.4}$	10	14.2	11.448
LocalSearch	10	10.7	9.929
ROCK$_{k=2,\theta=0.8}$	2	48.2	16.777
ROCK$_{k=7,\theta=0.8}$	7	25.9	10.568
ROCK$_{k=9,\theta=0.8}$	9	9.9	10.312
LIMBO$_{k=2,\phi=0.3}$	2	10.9	13.011
LIMBO$_{k=7,\phi=0.3}$	7	4.2	10.505
LIMBO$_{k=9,\phi=0.3}$	9	4.2	10.360

number of clusters of each clustering (k) and the disagreement error (E_D), that is, the objective function in Problem 2 optimized by our algorithms. This is the measure explicitly optimized by our algorithms. Since the clustering aggregation algorithms make their own decisions for the resulting number of clusters, we have run the other two algorithms for the same values of k so that we ensure fairness. Overall the impurity indices are comparable with the exception of LIMBO's impressive performance on Mushrooms for $k = 7$ and $k = 9$. Our algorithms achieve low distance error, with LocalSearch always having the lowest distance error. The distance error for LocalSearch is close to the theoretical lower bound that is computed by considering an idealized algorithm that merges all edges with weight less than $\frac{1}{2}$, and splits all edges with weight more than $\frac{1}{2}$. Furthermore, the attractiveness of the algorithms Agglomerative, Furthest, and LocalSearch lies in the fact that they are completely parameter-free. Neither a threshold nor the number of clusters need to be specified. The number of clusters discovered by our algorithms seem to be very reasonable choices: for the Votes dataset, most people vote according to the official position of their political parties so having two clusters is natural; for the Mushrooms dataset, notice that both ROCK and LIMBO achieve much better quality for the suggested values $k = 7$ and $k = 9$ so it is quite likely that the correct number of clusters is around these values. Indicatively, in Table III, we present the confusion matrix for the clustering produced by the Agglomerative algorithm on the Mushrooms dataset.

For the Census dataset, clustering aggregation algorithms report about 50-60 clusters. To run clustering aggregation on the Census dataset, we need to resort to the Sampling algorithm. As an indicative result, when the Sampling uses the Furthest algorithm to cluster a sample of 4,000 persons, we obtain 54 clusters and the impurity index is 24%. ROCK does not scale for a dataset of this size, while LIMBO with parameters $k = 2$ and $\phi = 1.0$ gives impurity index 27.6%. For contrasting these numbers, we mention that supervised classification algorithms (like decision trees and Bayes classifiers) yield classification error between 14% and 21%—but again, we note that clustering is a conceptually different task than classification. We visually inspected some of the smallest of the 54 different clusters, and many corresponded to distinct social groups, for

Table III. Confusion Matrix for Class Labels and Clusters Found by the Agglomerative Algorithm on Mushrooms Dataset (The column c_j of the Matrix Gives the Number of Mushrooms Found in Cluster C_j Having Labels Poisonous and Edible.)

	c_1	c_2	c_3	c_4	c_5	c_6	c_7
Poisonous	808	0	1296	1768	0	36	8
Edible	2864	1056	0	96	192	0	0

Table IV. An Example of a Small Cluster Representing a Distinct Social Group (The attributes of the table are as follows: WC: Workclass, Edu: Education, MS: Mar-status, Occ: Occupation, Rel: Relationship, Race: Race, Gen: Gender, and NC: Nat-country. The values in the table are as follows: ?: missing value, SelfEmp: Self-emp-not-inc, MarSpAb: Married-spouse-absent, AsAcdm: Assoc-acdm, AsVoc: Assoc-voc, NevMar: Never-married, Div: Divorced, FarFish: Farming-fishing, CrRep: Craft-repair, Unmar: Unmarried, OthRel: Other-relative, OwnCh: Own-child, NiFam: Not-in-family, AIEsk: Amer-Indian-Eskimo, W: White, M: Male, F: Female, AdmCl: Adm-clerical.)

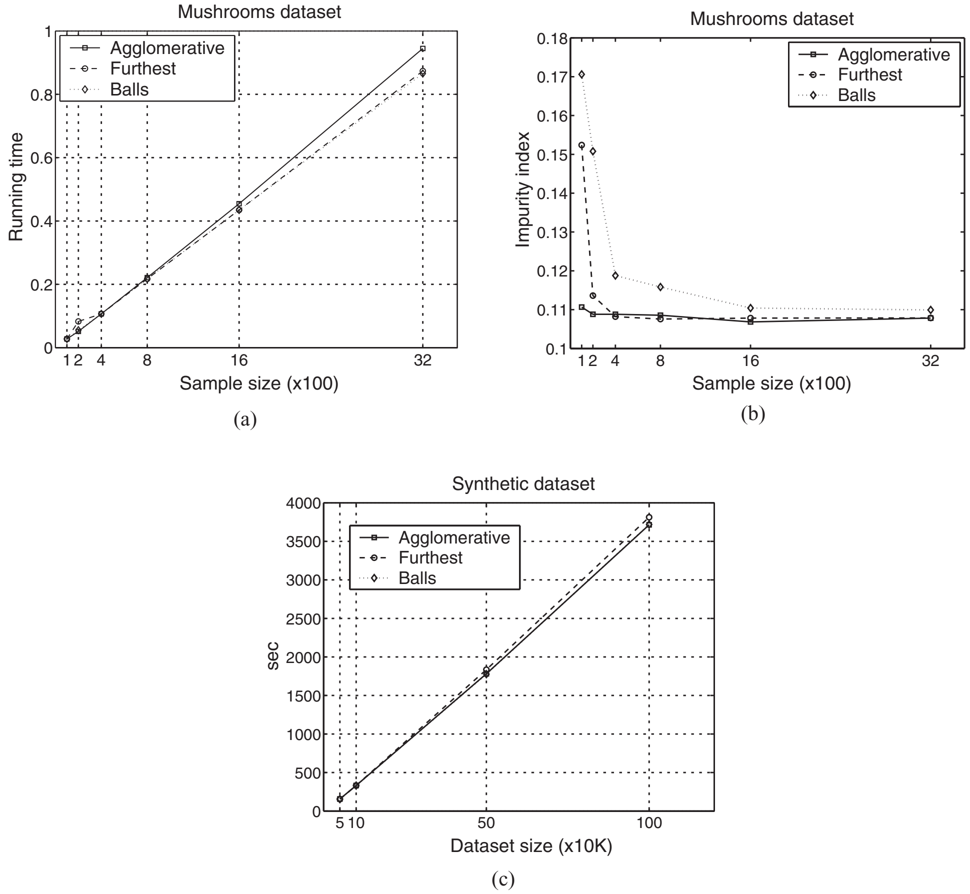
WC	Edu	MS	Occ	Rel	Race	Gen	NC
?	1st-4th	MarSpAb	?	Unmar	AIEsk	M	US
Private	10th	NevMar	FarFish	Unmar	AIEsk	M	US
SelfEmp	10th	MarSpAb	AdmCl	Unmar	AIEsk	F	US
SelfEmp	7th-8th	NevMar	FarFish	Unmar	W	M	US
SelfEmp	AsAcdm	Div	CrRep	OwnCh	AIEsk	M	US
SelfEmp	AsAcdm	MarSpAb	FarFish	Unamr	AIEsk	M	US
SelfEmp	AsAcdm	NevMar	FarFish	OthRel	W	M	US
SelfEmp	AsVoc	NevMar	FarFish	NiFam	AIEsk	M	US

example, male Eskimos occupied with farming-fishing, married Asian-Pacific islander females, unmarried executive-manager females with high-education degrees, etc. An example of such a small cluster is shown in Table IV.

7.3 Handling Large Datasets

In this section, we describe our experiments with the Sampling algorithm that allows us to apply clustering aggregation to large datasets. First we use the Mushrooms dataset to experiment with the behavior of our algorithms as a function of the sample size. As we saw in Table II, the number of clusters found with the nonsampling algorithms is around 10. When sampling is used, the number of clusters found in the sample remains close to 10. For small sample size, clustering the sample is relatively fast compared to the postprocessing phase of assigning the nonsampled points to the best cluster, and the overall running time of the Sampling algorithm is linear. In Figure 5(a), we plot the running time of the Sampling algorithm as a fraction of the running time of the nonsampling algorithm, and we show how it changes as we increase the sample size. For a sample of size 1,600, we achieve more than 50% reduction in the running time. At the same time, the impurity index of the algorithm converges very fast to the value achieved by the nonsampling algorithms. This is shown in Figure 5(b). For sample size 1,600, we have almost the same impurity index with only half of the running time.

We also measured the running time of the Sampling algorithm for large synthetic datasets. We repeated the configuration of the experiments shown in



(a)

(b)

(c)

Fig. 5. Scalability experiments for the Sampling algorithm. (a) The running time as a fraction of the time for the whole dataset plotted against the sample size. (b) The impurity index as a function of the sample size. (c) The running time as a function of the dataset size.

Figure 4 but on a larger scale. Each dataset consists of points generated from clusters normally distributed around five centers plus an additional 20% of uniformly distributed points. We generate datasets of sizes 50K, 100K, 500K, and 1M points. We then cluster the points using MATLAB's k-means implementation for $k = 2, \ldots, 10$, and we run Sampling clustering aggregation on the resulting 9 clusterings. The results are shown in Figure 5(c). These results are for sample size equal to 1,000. Once again, the five correct clusters were identified in the sample, and the running time is dominated by the time it takes to assign the nonsampled points in the clusters of the sample, resulting to the linear behavior shown in the figure.

8. CONCLUDING REMARKS

In this article we considered the problem of clustering aggregation. Simply stated, the idea is to cluster a set of objects by trying to find a clustering that agrees as much as possible with a number of preexisting clusterings. We motivated the problem by describing in detail various applications of clustering aggregation including clustering categorical data, dealing with heterogeneous data, improving clustering robustness, and detecting outliers. We formally defined the problem, and we showed its connection with the problem of correlation clustering. We proposed various algorithms for both the clustering aggregation and the correlation clustering problem including a sampling algorithm that allows us to handle large datasets with no significant loss in the quality of the solutions. We also analyzed the algorithms theoretically, providing approximation guarantees whenever possible. Finally, we verified the intuitive appeal of the proposed approach, and we studied the behavior of our algorithms with experiments on real and synthetic datasets.

ACKNOWLEDGMENTS

We thank Alexander Hinneburg for fruitful discussions, and Periklis Andritsos for pointers to code, data, and his help with the experiments. We also thank the anonymous reviewers for many useful comments.

REFERENCES

AILON, N., CHARIKAR, M., AND NEWMAN, A. 2005. Aggregating inconsistent information: Ranking and clustering. In *Proceedings of the ACM Symposium on Theory of Computing (STOC)*. 684–693.

ANDRITSOS, P., TSAPARAS, P., MILLER, R. J., AND SEVCIK, K. C. 2004. LIMBO: Scalable clustering of categorical data. In *Proceedings of the International Conference on Extending Database Technology (EDBT)*. 123–146.

BANSAL, N., BLUM, A., AND CHAWLA, S. 2004. Correlation clustering. *Machine Learn. 56*, 1–3, 89–113.

BARTHELEMY, J.-P. AND LECLERC, B. 1995. The median procedure for partitions. *DIMACS Series in Discrete Mathematics*, 3–34.

BLAKE, C. L. AND MERZ, C. J. 1998. UCI repository of machine learning databases.

BOULIS, C. AND OSTENDORF, M. 2004. Combining multiple clustering systems. In *Proceedings of the European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)*. 63–74.

CHARIKAR, M., GURUSWAMI, V., AND WIRTH, A. 2003. Clustering with qualitative information. In *Proceedings of the IEEE Symposium on Foundations of Computer Science (FOCS)*. 524–533.

CRISTOFOR, D. AND SIMOVICI, D. A. 2001. An information-theoretical approach to genetic algorithms for clustering. Tech. rep. TR-01-02, UMass, Boston, MA.

DEMAINE, E. D., EMANUEL, D., FIAT, A., AND IMMORLICA, N. 2006. Correlation clustering in general weighted graphs. *Theoret. Comput. Science 361*, 2–3, 172–187.

DEZA, M. AND LAURENT, M. 1997. *Geometry of Cuts and Metrics*. Springer-Verlag.

DWORK, C., KUMAR, R., NAOR, M., AND SIVAKUMAR, D. 2001. Rank aggregation methods for the Web. In *Proceedings of the International World Wide Web Conference*. 613–622.

FAGIN, R., KUMAR, R., AND SIVAKUMAR, D. 2003. Comparing top k lists. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms (SODA)*. 28–36.

FERN, X. Z. AND BRODLEY, C. E. 2003. Random projection for high dimensional data clustering: A cluster ensemble approach. In *Proceedings of the International Conference on Machine Learning (ICML)*. 186–193.

FILKOV, V. AND SKIENA, S. 2004. Integrating microarray data by consensus clustering. *Int. J. AI Tools 13*, 4, 863–880.

FRED, A. AND JAIN, A. K. 2002. Data clustering using evidence accumulation. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*. 276–280.

Guha, S., Rastogi, R., and Shim, K. 2000. ROCK: A robust clustering algorithm for categorical attributes. *Inform. Syst. 25*, 5, 345–366.

Hamerly, G. and Elkan, C. 2003. Learning the k in k-means. In *Advances in Neural Information Processing Systems (NIPS)*.

Han, J. and Kamber, M. 2001. *Data Mining: Concepts and Techniques*. Morgan Kaufmann.

Hand, D., Mannila, H., and Smyth, P. 2001. *Principles of Data Mining*. The MIT Press, Cambridge, MA.

Hochbaum, D. and Shmoys, D. 1985. A best possible heuristic for the k-center problem. *Mathem. Operat. Resea.*, 180–184.

Jain, A. K. and Dubes, R. C. 1988. *Algorithms for Clustering Data*. Prentice-Hall.

Mielikäinen, T., Terzi, E., and Tsaparas, P. 2006. Aggregating time partitions. In *Proceedings of the International ACM SIGKDD Conference on Knowledge Discovery in Data Mining (KDD)*. 347–356.

Schwarz, G. 1978. Estimating Dimension of a Model. *Ann. Statis. 6*, 461–464.

Smyth, P. 2000. Model selection for probabilistic clustering using cross-validated likelihood. *Statist. Comput. 10*, 1, 63–72.

Strehl, A. and Ghosh, J. 2002. Cluster ensembles—A knowledge reuse framework for combining multiple partitions. *J. Machine Learn. Resear. 3*, 583–617.

Swamy, C. 2004. Correlation clustering: maximizing agreements via semidefinite programming. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms (SODA)*. 526–527.

Topchy, A., Jain, A. K., and Punch, W. 2004. A mixture model of clustering ensembles. In *Proceedings of the SIAM International Conference on Data Mining (SDM)*. 379–390.

Received December 2005; revised December 2006; accepted January 2007