Error Bounds for Correlation Clustering

Thorsten Joachims TJ@CS.CORNELL.EDU
Cornell University, Dept. of Computer Science, 4153 Upson Hall, Ithaca, NY 14853 USA

John Hopcroft JEH@CS.CORNELL.EDU
Cornell University, Dept. of Computer Science, 5144 Upson Hall, Ithaca, NY 14853 USA

Abstract

This paper presents a learning theoretical analysis of correlation clustering (Bansal et al., 2002). In particular, we give bounds on the error with which correlation clustering recovers the correct partition in a planted partition model (Condon & Karp, 2001; McSherry, 2001). Using these bounds, we analyze how the accuracy of correlation clustering scales with the number of clusters and the sparsity of the graph. We also propose a statistical test that analyzes the significance of the clustering found by correlation clustering.

1. Introduction

While we have gained a detailed learning theoretical understanding of supervised learning over the last decades, our understanding of unsupervised clustering is still rather limited. For example, how much data is necessary so that a clustering algorithm outputs a reliable clustering? How does the amount of data depend on the distribution of the data? Is the particular clustering produced by some algorithm significant?

This paper addresses these questions for a particular graph-based clustering algorithm, namely correlation clustering (Bansal et al., 2002). Correlation clustering is a particularly attractive clustering method, since its solution can be approximated efficiently (see e.g. (Demaine & Immorlica, 2003; Swamy, 2004)) and it automatically selects the number of clusters. While Bansal et al. (2004) briefly discuss the behavior of their algorithm under noise in the data, no learning theoretic analysis exists yet. To conduct the analysis, we propose a simple probabilistic model over graphs that extends the planted partition model (Condon & Karp, 2001; McSherry, 2001). An advantage of this model is that its simplicity allows a concise analysis, while providing a starting point for exploring more complex models.

While a substantial amount of theoretical work on clustering algorithms exists, much of this work is concerned primarily with computational aspects (e.g. (Dasgupta, 1999; McSherry, 2001)). Probably the most general learning theoretic model of clustering to date is “Empirical Risk Approximation” (Buhmann, 1998; Buhmann & Held, 1999), which applies to clustering algorithms that optimize an objective function. Buhmann (1998) uses uniform convergence arguments to bound the difference between the objective value a clustering achieves on the training data and its objective over the data distribution. Ben-David follows this approach to derive finite sample bounds for k-median clustering (Ben-David, 2004). The k-means or vector quantization problem is probably the best studied clustering problem. Among other results, statistical consistency was proven by Pollard (Pollard, 1981), and lower (Bartlett et al., 1998) and upper bounds (Linder et al., 1994) on the quantization error are know. Our work is substantially different since it considers non-metric clustering problems where the data comes in the form of graphs. Graph-based clustering problems are ubiquitous in WWW search and social network analysis (e.g. (Kleinberg, 1999)). Instead of limiting our analysis to investigating statistical consistency, like the work of von Luxburg et al. (2004) for spectral clustering, we rather use a more restrictive model in which we can prove finite sample bounds for correlation clustering.

Our analysis makes three contribution. First, we define a model in which we derive finite-sample error bounds for correlation clustering. Second, we study the asymptotic behavior of correlation clustering with respect to the density of the graph and the scaling of cluster sizes. And finally, we propose a statistical test for evaluating the significance of a clustering.

Appearing in *Proceedings of the 22^{nd} International Conference on Machine Learning*, Bonn, Germany, 2005.

Figure 1. Example of correlation clustering on graph with 6 vertices. The graph and it's weight matrix W are depicted on the left. Solid edges indicate a weight of $+1$, dotted edges a weight of -1. The correlation clustering is depicted on the right.

2. Correlation Clustering

The correlation clustering of an n vertex weighted graph with edge weights $W_{ij} \in \Re$ is the partition of the vertices that minimizes the sum of positive weights that are cut minus the negative weights that are not cut. An example of an (undirected) graph with six vertices is given in Figure 1. In this example, the matrix of edge weights W contains only three possible values, namely -1, 0, and $+1$. The correlation clustering of W is depicted on the right-hand side of Figure 1. The clustering contains one cluster containing vertices 1,2,3, and 4 and another cluster containing vertices 5 and 6. This clustering cuts 2 (directed) edges with weight $+1$, while it fails to cut 2 (directed) edges with weight -1. This gives this clustering a score of $1 * 2 - (-1) * 2 = 4$, which optimizes the objective function of correlation clustering.

More formally, the correlation clustering $\hat{S}$ of a graph with edge weights W is given by the solution $\hat{Y}$ of the following integer program (Demaine & Immorlica, 2003). The number k of clusters is not fixed by the user, but determined as part of the clustering process. The edge weights W_{ij} enter the optimization problem as follows. W^+ is equal to adjacency matrix W, except that all negative edge weights are replaced by 0. Similarly, W^- is equal to W, except that all positive edge weights are replaced by 0. The optimization is over the $n \times n$ matrix Y with elements $Y_{ij} \in \{0, 1\}$. A value of 1 for Y_{ij} indicates that objects x_i and x_j are in the same cluster. A value of 0 indicates that they are in different clusters.

$$\begin{aligned}
\min_{Y} \quad & \sum_{i=1}^{n} \sum_{j=1}^{n} \left[(1 - Y_{ij}) W_{ij}^{+} - Y_{ij} W_{ij}^{-} \right] & (1) \\
\text{subject to} \quad & \forall i : Y_{ii} = 1 & (2) \\
& \forall i, j : Y_{ij} = Y_{ji} & (3) \\
& \forall i, j, k : Y_{ij} + Y_{jk} \leq Y_{ik} + 1 & (4) \\
& \forall i, j : Y_{ij} \in \{0, 1\} & (5)
\end{aligned}$$

We call Y a *cluster indicator matrix*. The constraints of the optimization problem directly encode the three conditions in the definition of an equivalence relation, namely reflexivity, symmetry, and transitivity. This means that any feasible Y — and therefore also the solution $\hat{Y}$ — directly corresponds to an equivalence relation and it is straightforward to derive a clustering from the solution $\hat{Y}$. We denote the clustering that corresponds to an indicator matrix Y with $S(Y)$. Vice versa, we denote with $Y(S)$ the cluster indicator matrix induced by clustering S on X. Finally, we define the cost of a matrix Y

$$cost_W(Y) = \sum_{i=1}^{n}\sum_{j=1}^{n}\left[(1 - Y_{ij})W_{ij}^{+} - Y_{ij}W_{ij}^{-}\right] \qquad (6)$$

as the value of the objective function for that clustering. For simplicity of notation, we assume that diagonal entries of W are always non-negative, i.e. $W_{ii} \geq 0$.

Note that the formulation of the optimization problem can be simplified. In particular, the reflexivity constraints and the associated variables Y_{ii} can be dropped. Similarly, one can eliminate the symmetry constraints by unifying their variables. While the solution of the optimization problem is known to be NP-complete (Bansal et al., 2002), there are effective approximation algorithms for this problem (e.g. (Bansal et al., 2002; Demaine & Immorlica, 2003; Swamy, 2004)).

3. Generalized Planted Partition Model

In this section we define a probabilistic data model similar to the one in (Condon & Karp, 2001; McSherry, 2001). For data generated according to this model, we will derive results that describe how accurately correlation clustering recovers the correct cluster structure.

In our model we assume that there is an arbitrary "true" partition $S^* = \{S_1^*, ..., S_{k^*}^*\}$ of the vertices X (i.e. $S_1^* \cup ... \cup S_k^* = X$ and $S_i^* \cap S_j^* = \emptyset$). Unlike in the model of Condon and Karp (2001), the number of clusters k^* and the size of each cluster are arbitrary and unknown to the clustering algorithm. To this partition S^* corresponds a probability distribution $P_{S^*}(W)$ over edge weights. We assume that $P_{S^*}(W)$ is the process that generates the data we want to cluster. The goal of the clustering algorithm is to recover the true partition S^* underlying the data generating process $P_{S^*}(W)$ from a single realization of edge weights W. The class of distributions $P_{S^*}(W)$ we consider is defined as follows.

Definition 1 (Gen. Planted Partition Model) *In a graph with n edges, the edge weights are generated by a distribution*

$$P_{S^*}(W|M, a, b) = \prod_{i=1}^{n}\prod_{j=1}^{n} P_{S^*}(W_{ij}|M_{ij}, a, b) \qquad (7)$$

so that each element W_{ij} of W is a bounded independent random variable in the interval $[a, b]$ with mean M_{ij}. Each $P_{S^*}(W_{ij}|M_{ij}, a, b)$ is constrained by the true partitioning S^* as follows. If $Y(S^*)_{ij} = 1$ (vertices i and j are in the same cluster), the mean M_{ij} of W_{ij} must fulfill the constraint that $M_{ij} \geq \mu_+ > 0$. If $Y(S^*)_{ij} = 0$ (vertices i and j are in different clusters), the mean M_{ij} of W_{ij} must fulfill $M_{ij} \leq \mu_- < 0$.

One can think of $P_{S^*}(W|M, a, b)$ (or $P_{S^*}(W)$ for short) as generating edge weights that are a noisy representation of the underlying true partition S^*. Figure 2 gives an example of a true partition S^*, how its structure is reflected in the matrix of means M, and how a particular matrix of edge weights W is drawn from $P_{S^*}(W)$. The matrix of means M controls the amount of noise[1] and we summarize M using the two parameters μ_+ and μ_-. μ_+ is a lower bound on the mean edge weight between any two vertices that are in the same cluster of S^*, while μ_- is an upper bound on the mean edge weight between any two vertices in different clusters.

$$\begin{aligned} \text{if } Y(S^*)_{ij} = 1 \text{ (same cluster):} \quad & E(W_{ij}) \geq \mu_+ > 0 \\ \text{if } Y(S^*)_{ij} = 0 \text{ (different cluster):} \quad & E(W_{ij}) \leq \mu_- < 0 \end{aligned}$$

In the example in Figure 2, μ_+ is 0.2 and μ_- is -0.3. Note that the model requires that the mean weight of between cluster edges be less than zero, and that the mean weight of within cluster edges be greater than zero. In addition, it requires that all weights are bounded[2], i.e. $\forall i, j : a \leq W_{ij} \leq b$.

This class of distributions $P_{S^*}(W)$ can be used to model a variety of clustering applications. Here are three examples:

Pair-Wise Classification This example is the application Bansal et al. (2002) use to motivate correlation clustering. Edge weights W_{ij} are derived from classifications of pairs as e.g. in noun-phrase coreference resolution (Ng & Cardie, 2002). For each pair of objects, a classification rule makes an independent judgment of whether two vertices should be in the same cluster or not. The edge weight is derived from the confidence in the judgment.

---

[1] A natural and straightforward extension is to allow a small random subset of edges to violate their constraints on the mean.

[2] Alternatively, we could assume that all variances are bounded.

Citation Network Analysis Edge weights represent citations in a network of bibliographic references. Each citation edge receives a weight of 1, non-present citations receive a negative weight w_-. We will discuss the value of w_- in Section 4.2. The matrix of means M reflects the probabilities with which documents cite each other dependent on whether they are in the same cluster or not.

Co-Clustering The co-clustering model (Dhillon, 2001), originally proposed for text, simultanously clusters the rows and the columns of a term/document matrix. This leads to a bipartite graph model with terms and documents being two sets of vertices. An edge of weight $W_{ij} = 1$ is present in the graph, if term i occurs in document j, it is equal to some negative value w_- if the word does not occur. Weights between terms and between documents are zero.

We will discuss more detailed results for query clustering in search engines and citation network analysis in Sections 4.2 and 5 to further illustrate the model.

4. Analysing the Error of Correlation Clustering

In this section, we analyze how well correlation clustering can reconstruct the true partition S^* based on a weight matrix W drawn from a probability distribution $P_{S^*}(W)$ that conforms with our generalized planted partition model. Figure 2 illustrates the statistical model in which we evaluate correlation clustering. In this model, correlation clustering is applied to the weight matrix W generated from $P_{S^*}(W)$. The resulting cluster indicator matrix $\hat{Y}$ and the partition $\hat{S}$ it induces are then compared against the true clustering S^*. We measure the error of $\hat{S}$ with respect to S^* using the following pair-wise loss function.

$$d(\hat{S}, S^*) = ||Y(\hat{S}) - Y(S^*)||_F^2 \tag{8}$$

Here $||.||_F$ denotes the Frobenius norm. Intuitively, $d(.,.)$ measures the distance between two clusterings as the number of elements that are different in the corresponding cluster indicator matrices. In the example in Figure 2, this difference is depicted as the shaded region in the right-most panel and has value $d(\hat{S}, S^*) = 10 + 8 = 18$.

In the following, we will first derive upper bounds on the error $d(\hat{S}, S^*)$ of correlation clustering with respect to the number of vertices n and the values of μ_+ and μ_-. After deriving the general results, we will apply them to the example settings mentioned above. Finally, we will discuss the asymptotic behavior of correlation clustering in our model.

4.1. Error Bound for Finite Samples

In our Planted Partition Model there is a true partition S^* of the given set of vertices X. Associated with the partition S^* is a probability distribution $P_{S^*}(W)$ of edge weights so that the mean of each within cluster edge exceeds $\mu_+ > 0$, and so that the mean of each between cluster edge is less than $\mu_- < 0$.

Our argument is structured as follows. First, given two partitions S and S^* with distance $d(S, S^*)$, we bound the probability that a weight matrix W drawn from $P_{S^*}(W)$ has a lower cost for partition S than for the true partition S^*, i.e. $cost_W(Y(S)) \leq cost_W(Y(S^*))$. In a second step, we will bound the probability that a weight matrix W drawn from $P_{S^*}(W)$ has a cost lower than the true partition for any partition S with $d(S, S^*) \geq \delta$, i.e. the probability that $\exists S : d(S, S^*) \geq \delta \wedge cost_W(Y(S)) \leq cost_W(Y(S^*))$. This bounds the probability of drawing a W so that correlation clustering returns a partition $\hat{S}$ which has an error $d(\hat{S}, S^*)$ greater than δ.

There are two components contributing to $d(\hat{S}, S^*)$. Let $d_+(S, S^*) = \delta_+$ be the number of vertex pairs that are clustered together in S^* but not in S. Similarly, let $d_-(S, S^*) = \delta_-$ be the number of vertex pairs that are clustered together in S but not in S^*. This means that $d(S, S^*) = \delta_+ + \delta_- = \delta$. Based on the magnitude of these two types of errors, we can bound the probability of the model generating a W for which the incorrect partition S has a lower cost than the true partition S^*.

Lemma 1 *Given two partitions S^* and S with $d_+(S, S^*) = \delta_+$ and $d_-(S, S^*) = \delta_-$, it holds for W drawn from the generalized planted partition model $P_{S^*}(W)$ with $\mu_- < 0 < \mu_+$ and $a \leq W_{ij} \leq b$ that*

$$P(cost_W(Y(S)) \leq cost_W(Y(S^*)) | S, S^*, \delta_+, \delta_-) \leq e^{-2\frac{(\delta_+ \mu_+ - \delta_- \mu_-)^2}{(\delta_+ + \delta_-)(b-a)^2}}$$

for any $\delta_+ + \delta_- \in [0, n(n-1)]$.

Proof *We can compute the difference of costs $cost_W(Y(S)) - cost_W(Y(S^*))$ of $Y(S)$ and $Y(S^*)$ with respect to W as*

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\left[(1-Y_{ij})W_{ij}^+ - Y_{ij}W_{ij}^-\right] - \sum_{i=1}^{n}\sum_{j=1}^{n}\left[(1-Y_{ij}^*)W_{ij}^+ - Y_{ij}^*W_{ij}^-\right]$$

$$= \sum_{\{(i,j):Y_{ij} \neq Y_{ij}^*\}}\left[(1-Y_{ij})W_{ij}^+ - Y_{ij}W_{ij}^-\right] - \sum_{\{(i,j):Y_{ij} \neq Y_{ij}^*\}}\left[(1-Y_{ij}^*)W_{ij}^+ - Y_{ij}^*W_{ij}^-\right]$$

$$= \sum_{\{(i,j):Y_{ij} \neq Y_{ij}^*\}}\left[Y_{ij}W_{ij}\right] - \sum_{\{(i,j):Y_{ij} \neq Y_{ij}^*\}}\left[Y_{ij}^*W_{ij}\right]$$

More precisely, if the distance between two clusterings S and S^* is $d_+(S,S^*) = \delta_+$ and $d_-(S,S^*) = \delta_-$, then there are exactly $\delta_+ + \delta_-$ elements of $Y(S)$ and $Y^*(S)$ on which $cost_W(Y(S))$ and $cost_W(Y(S^*))$ differ. Denote the corresponding sets of edges as D_+ and D_-. This implies that if $cost_W(S) \leq cost_W(S^*)$, then the following sum must be negative.

$$\sum_{(i,j)\in D_+} W_{ij} - \sum_{(i,j)\in D_-} W_{ij} \leq 0 \tag{9}$$

Since the edge weights in W are drawn independently, we can use Hoeffding's inequality to bound the probability that this sum is negative. Hoeffding's inequality bounds the deviation of a sum of independent and bounded random variables $X_k \in [a_i, b_i]$ from its mean.

$$P\left(\sum X_k - E(\sum X_k) \leq c\right) \leq e^{\frac{-2c^2}{\sum (b_i-a_i)^2}}$$

In our case, we set $\sum X_k = \sum_{(i,j)\in D_+} W_{ij} - \sum_{(i,j)\in D_-} W_{ij}$, $E(\sum X_k) = \sum_{(i,j)\in D_+} M_{ij} - \sum_{(i,j)\in D_-} M_{ij}$, $c = -E(\sum X_k)$, and $\sum (b_i - a_i)^2 = (\delta_+ + \delta_-)(b-a)^2$. We can now apply Hoeffding's inequality to bound the probability $P(\sum X_k \leq 0) = P(\sum X_k - E(\sum X_k) \leq -E(\sum X_k))$.

$$P\left(\sum X_k \leq 0\right) \quad \leq \quad e^{\frac{-2\left(\sum_{(i,j)\in D_+} M_{ij} - \sum_{(i,j)\in D_-} M_{ij}\right)^2}{(\delta_+ + \delta_-)(b-a)^2}}$$

Since the M_{ij} are bounded by μ_+ and μ_- in the planted partition model, it holds that

$$\left(\sum_{(i,j)\in D_+} M_{ij} - \sum_{(i,j)\in D_-} M_{ij}\right)^2 \geq (\delta_+\mu_+ - \delta_-\mu_-)^2 \ ,$$

which completes the proof of the lemma. ■

The lemma bounds for a particular clustering S the probability of drawing a W for which S has a misleadingly good cost. To bound the probability for all clusterings, we need an upper bound on the number of possible clusterings. The exact number of clusterings is know as the Bell number, but the following bound suffices for our purposes.

Lemma 2 *The number $C^{\#}(n)$ of possible clusterings of n points is at most n factorial.*

Proof *By induction over n. For $n = 1$ there is exactly one clustering. Given $C^{\#}(n-1)$, for each clustering of $n-1$ objects the n-th object can either start it's own cluster, or join one of at most $n-1$ existing clusters in the clustering. So, there are at most n ways to extend each of the existing clusterings of n objects. This implies $C^{\#}(n) \leq nC^{\#}(n-1)$.* ■

We can now bound the probability that correlation clustering returns a partition with large error. We state the theorem in terms of the error rate $Err(S, S^*)$, which is the fraction of misclassified edges

$$Err(S,S^*) = \frac{d(S,S^*)}{n(n-1)}$$

Note that the following bound is with respect to the randomness in drawing the cost matrix W. However, note that the bound also holds for cases where the optimization problem in Eqs. (1)-(5) does not have a unique solution.

Theorem 1 *Given the true partition S^* of n points, the probablity that correlation clustering returns a partition $\hat{S}$ with $Err(\hat{S}, S^*) \geq \epsilon$ in the planted partion model with $\mu = min\{\mu_+, -\mu_-\}$ and $a \leq W_{ij} \leq b$ is bounded by*

$$P(Err(\hat{S},S^*) \geq \epsilon) \quad \leq \quad e^{n\ln(n) - 2\,\epsilon\, n(n-1)\frac{\mu^2}{(b-a)^2}} \tag{10}$$

Proof *We bound the probability that any partition with error $d(S,S^*) \geq \delta = \epsilon\, n(n-1)$ has a cost that is better or equal to that of the true partition S^*. The bound follows directly from the union bound and Lemmas 1 and 2.*

$$P(\exists S : d(S,S^*) \geq \delta \wedge cost_W(Y(S)) \leq cost_W(Y(S^*)))$$
$$\leq \ n! \ e^{-2\frac{(\delta_+\mu_+ - \delta_-\mu_-)^2}{(\delta_+ + \delta_-)(b-a)^2}} \ \leq \ e^{n\ln(n) - 2\delta\frac{\mu^2}{(b-a)^2}}$$

This proves that with high probability no partition S with distance $d(S,S^*) \geq \delta$ has a cost that is better than the cost of the true partition S^*. Since correlation clustering returns the partition $\hat{S}$ with the lowest cost, $\hat{S}$ has a distance $d(\hat{S},S^*)$ less than δ with high probablity. ■

Related bounds were derived by Condon and Karp (2001), as well as McSherry (2001). However, Condon and Karp (2001) consider a more restricted setting where all clusters have equal size, and this size is known to the clustering algorithm a priori. Furthermore, both bounds are different from our work, since they do not quantify the error between S^* and an imperfect $\hat{S}$.

The following example illustrates the bound. Assume a planted partition model with $\mu = \mu_+ = -\mu_- = 0.5$ and bounds $a = 1$ and $b = -1$. Let's assume we have $n = 3000$ objects $X = (x_1, ..., x_n)$ and a true partition $S^* = \{S_1, S_2, S_3\}$ with three clusters of size 1000 each. Applying the bound tells us that with 95% confidence, the error rate $Err(\hat{S}, S^*)$ of the partition $\hat{S}$ returned by correlation clustering is at most 2.2%. Furthermore, for true clusters of size k, moving e objects out of the correct cluster leads to a pairwise loss of at least $e\,k - e(e+1)/2$. This minimum pairwise loss is achieved by splitting one of the clusters into two subclusters of size e and $k - e$. Therefore, with 95% confidence at most 215 of the 3000 objects are not clustered correctly.

4.2. Application: Query Clustering

We will now illustrate how the planted partition model can be substantiated with particular parameters according to application settings. We use query clustering in search engines as an example. In query clustering, the goal is to group queries together that are semantically related (e.g. the queries "imdb" and "movie reviews"). To measure the relation between two queries, we make use of the fact that users reformulate queries during their search. We consider a fixed set of n queries as nodes (e.g. n popular queries). Using the query log over some time interval, the adjacency matrix W is constructed by assigning $W_{ij} = 1$ if some user issued query x_i directly followed by query x_j, and $W_{ij} = w_- < 0$ otherwise. We will discuss the choice of w_- below. This representation exploits that two consecutive queries by the same user are likely to be related.

Before applying correlation clustering to W, we define what we mean by a cluster of related queries. We define that queries within the same cluster co-occur in the query log during the time interval with probabiltiy at least p_+, while between cluster co-occurrences have probability $p_- < p_+$. The independence assumption approximately holds in this setting (especially, if one considers only one query pair per user), so that one can apply our results for the planted partition model as follows.

Corollary 1 *Based on the true partition S^* of n nodes, the edges of a directed graph are independently drawn so that within-cluster edges have probability at least p_+, and between-cluster edges have probabiltity less than $p_- < p_+$. From this graph construct W by assigning $W_{ij} = 1$ to each element corresponding to an edge, and $W_{ij} = w_- = \frac{p_+ + p_-}{p_+ + p_- - 2}$ otherwise. The probability that the error rate $Err(\hat{S}, S^*)$ of the correlation clustering $\hat{S}$ of W is greater than ϵ is bounded by*

$$P(Err(\hat{S}, S^*) \geq \epsilon) \;\leq\; e^{n \ln(n) - \frac{1}{2}\,\epsilon\, n(n-1)(p_+ - p_-)^2} \quad (11)$$

We omit the proof for brevity, since it is a direct consequence of Theorem 1. Note that the particular choice of w_- maximizes $\mu = min\{\mu_+, -\mu_-\}$. It is straightforward to derive other (and potentially tighter) versions of the bound by replacing Hoeffding's inequality, but omit their discussion for brevity.

5. Asymptotic Behavior

How does the bound scale if the number of nodes in the graph grows? Growing graphs are natural, for example, in citation network analysis. Clustering in citation networks is used to reveal groups of related publications. Similar to query clustering, one could use correlation clustering to find clusters of papers that reference each other with high frequency. Let W be the adjacency matrix of the citation graph in which $W_{ij} = 1$ if paper x_i cites paper x_j, and $W_{ij} = w_- < 0$ otherwise.

Clustering in citation networks is different from query clustering in at least two respects[3]. First, while it is easy to control the sparsity of the graph by considering shorter or longer query logs in query clustering, the sparsity of the citation graph cannot be manipulated. Second, with a growing number of nodes, the number of clusters grows as well. We discuss both issues in the following.

5.1. How does the Bound Scale with Increasing Sparsity of the Graph?

If the lower bound $\mu = min\{\mu_+, -\mu_-\}$ on the difference of means for between and within cluster edges is a constant independent of n, then in Theorem 1 the probability that the error is greater than any constant fraction ϵ goes to zero since the second term in the exponent of (10) is order n^2. However, μ being a constant independent of n leads to very dense data. In citation network analysis, for example, graphs are usually very sparse with only a constant number of nonzero entries per row. Such a level of sparsity implies that μ is of size $\frac{1}{n}$ and that the second term in the exponent is constant. In this case, the first term dominates giving a meaningless bound of $e^{n \ln(n)}$. Thus, if we wish to have small probability of more than a constant fraction error, we need μ to grow faster than $\sqrt{\frac{\ln(n)}{n}}$ for the second term in the exponent of e to dominate.

---

[3] Furthermore, the independence assumption is likely to be less valid than in query clustering.

5.2. How does the Bound Scale with Cluster Size?

For an increasing number of nodes n, assume that each true partition S^*_n contains a fixed number k of clusters $S^*(n) = \{S^*_1(n), \dots, S^*_k(n)\}$ that each grow proportionally with n. Let $f_i = \frac{|S^*_i(n)|}{n}$ be the constant fraction of nodes in cluster $S^*_i(n)$ and, without loss of generality, let cluster k be the smallest cluster. With increasing n, does correlation clustering eventually recover each of the clusters? Suppose that we want to guarantee with high probability that all but a fraction of $\gamma \le \frac{f_k}{2}$ nodes are clustered correctly. If γn nodes are misclassified by some partition S, the value of the pairwise loss is at least $d(S, S^*(n)) \ge 2n^2\gamma(f_k - \gamma)$. Since $d(S, S^*(n))$ is quadratic in n, the bound from Theorem 1 shows that the probability of misclassifying a constant fraction γ of nodes goes to zero.

If the clusters do not grow proportionally with n but slower, the pairwise loss $d(S, S^*(n))$ does not grow quadratically in n. This happens, for example, when clusters grow at different rates or when the number of clusters grows with n. To ensure convergence of the bound from Theorem 1, we need $d(S, S^*(n))$ to grow faster than $n \ln(n)$. This is ensured if the fraction of nodes in each cluster grows faster than $\frac{\ln(n)}{n}$.

6. Is a Clustering Significant?

In typical applications of correlation clustering we are given a set of data W and we apply correlation clustering to detect potential cluster structure. So far, this paper addressed the question of whether the correlation clustering $\hat{S}$ reveals the *true* underlying structure S^*. We will now turn to the related question of whether the data reveals *any* significant cluster structure. Answering this question is important, since it provides a practitioner with a measure of confidence (or lack thereof) in $\hat{S}$. In the following we use correlation clustering to derive a significance test that let's us reject the null hypothesis that the data was produced by a random process without any underlying structure.

As the null hypothesis, we use a planted partition model where all edge-weight distributions $P(W_{ij})$ have the same mean, i.e. $M_{ij} = M_{kl}$. This null hypothesis captures that there is no structure in our data. For simplicity of presentation, we consider only the setting of citation network analysis, so that all W_{ij} take only two values indicating whether a particular edge is present or not. Let p be the probability that any given edge is present. For correlation clustering, the resulting graph is transformed into a weighted complete graph with weight matrix W by weighting present edges with 1, and inserting an edge with weight $w_- < 0$ whenever there is no edge present.

In this model we can pick a cost threshold ρ and bound the probability that the distribution from the null hypothesis generates a set of data W for which the partition $\hat{S}$ returned by correlation clustering has $cost_W(Y(\hat{S}))$ less than the threshold ρ. If we observe that $cost_W(Y(\hat{S}))$ is less than ρ for our given set of data W, we can use this bound on the probability to reject the null hypothesis with the corresponding confidence. For technical reasons that will be discussed in the proof of Lemma 3, ρ must be less than $(n^2 - \eta)p - (\eta - n)(1 - p)w_-$. The following derivation of the significance test proceeds by first bounding the probability for a single S in Lemma 3, and then by extending the result to hold uniformly for all S in Theorem 2.

Lemma 3 *Given a graph with n nodes and a particular clustering S of the graph for which we denote $||Y(S)||$ as η. Let $w_- \in \Re$, $\rho \in \Re$, and $p \in \Re$ so that $w_- < 0$, $0 \le \rho \le (n^2 - \eta)p - (\eta - n)(1 - p)w_-$, and $0 \le p \le 1$. If we randomly generate weights on the edges of the graph so that edges have weight 1 with probability p and weight w_- otherwise, the probability that clustering S has $cost_W(Y(S)) \le \rho$ is*

$$P(cost_W(Y(S)) \le \rho | S, \eta) \le e^{-2\frac{((n^2-\eta)p-(\eta-n)(1-p)w_- - \rho)^2}{n(n-1)(1-w_-)^2}}.$$

Proof *Since we have $n(n-1)$ random variables (i.e. the off diagonal entries of the cost matrix) that are bounded within $[w_-, 1]$, we can apply Hoeffding's inequality and get*

$$P(cost_W(Y(S)) \le \rho | S, \eta) \le e^{-2\frac{(E(cost_W(Y(S)) - \rho)^2}{n(n-1)(1-w_-)^2}} \quad (12)$$

for $\rho \in [0, (n^2 - \eta)p - (\eta - n)(1 - p)w_-]$. Note that ρ has to be less than $E(cost_W(Y(S))$ for Hoeffding's inequality to apply, thus the restriction to the interval. It remains to determine the expected cost $E(cost_W(Y(S))$. For a partition matrix $Y(S)$ with $(\eta - n)$ off-diagonal entries equal to 1 and $(n^2 - \eta)$ entries equal to 0, the expectation is $E(cost_W(Y(S)) = (n^2 - \eta)p - (\eta - n)(1 - p)w_-$. Substituting this into (12) yields the result. ∎

Theorem 2 *Let $w_- \in \Re$, $\rho \in \Re$, and $p \in \Re$ so that $w_- < 0$, $0 \le \rho \le n(n-1)min\{p, -(1-p)w_-\}$, and $0 \le p \le 1$. For a complete graph with n nodes where edges have weight 1 with probability p and weight w_- otherwise, the probability that the clustering $\hat{S}$ returned by correlation clustering has $cost_W(Y(\hat{S})) \le \rho$ is*

$$P(cost_W(Y(\hat{S})) \le \rho) \le e^{n \ln(n) - 2\frac{n(n-1)\left(min\{p, (p-1)w_-\} - \frac{\rho}{n(n-1)}\right)^2}{(1-w_-)^2}}$$

Proof *We prove a uniform bound in the sense that*

$$P(cost_W(Y(\hat{S})) \leq \rho) \leq P(\exists S : cost_W(Y(S)) \leq \rho)$$

To apply the union bound, we need a bound on $P(cost_W(Y(S)) \leq \rho | S, \eta)$ *that holds independent of* η*. Relaxing the bound from Lemma 3, it holds for every clustering* S *independent of* $\eta = ||Y(S)||$ *that*

$$\begin{aligned} & P(cost_W(Y(S)) \leq \rho | S) \\ \leq \quad & e^{-2\frac{(min_{0 \leq \eta \leq n(n-1)}\{(n^2-\eta)p-(\eta-n)(1-p)w_-\}-\rho)^2}{n(n-1)(1-w_-)^2}} \\ = \quad & e^{-2\frac{n(n-1)(min\{p,(1-p)w_-\}-\frac{\rho}{n(n-1)})^2}{(1-w_-)^2}} \end{aligned}$$

Applying the union bound w.r.t. the upper bound on the number of clusterings from Lemma 2 yields the result. ∎

Note that p is a parameter that needs to be fixed independent of the data. However, for practical purposes one can consider estimating p as the fraction of positive edges in W. Given p, a reasonable choice for w_- is $w_- = -\frac{p}{1-p}$, since it maximizes the numerator in the exponent. For this choice of w_- we can apply the bound from Theorem 2 in a hypothesis test as follows. We decide on a confidence level δ and solve

$$e^{n\ln(n)-2\frac{n(n-1)\left(p-\frac{\rho}{n(n-1)}\right)^2}{(1-w_-)^2}} \leq \delta \tag{13}$$

for the significance threshold ρ as follows.

$$\rho \leq n(n-1)p - n\left(1+\frac{p}{1-p}\right)\sqrt{\frac{n\ln(n)-\ln(\delta)}{2}} \tag{14}$$

If $cost_W(Y(\hat{S}))$ is less or equal to ρ, we can reject the null hypothesis with confidence δ.

7. Conclusions and Future Work

We presented a simple probabilistic graph model in which we analyze correlation clustering. The model allows us to derive finite sample bounds on the error with which correlation clustering recovers the graph structure. The results give insight into the behavior of correlation clustering with respect to the number of nodes, the density of the edges, and the number of clusters. Furthermore, we derive a test which can be applied to validate the significance of a given clustering.

While the planted partition model is an interesting starting point for analyzing clustering algorithms, there is need for generalizing the model and removing its assumptions. Clearly, the biggest assumption in the model is that edge weights are independently distributed. It is an interesting question whether this assumption can be relaxed without making the bounds too lose for any practical relevance.

This work was funded in part under NSF awards IIS-0412894, IIS-0312910, and the KD-D grant.

References

Bansal, N., Blum, A., & Chawla, S. (2002). Correlation clustering. *IEEE Symposium on Foundations of Computer Science (FOCS).*

Bansal, N., Blum, A., & Chawla, S. (2004). Correlation clustering. *Machine Learning*, *56*.

Bartlett, P. L., Linder, T., & Lugosi, G. (1998). The minimax distortion redundancy in empirical quantizer design. *IEEE Transactions on Information Theory*, *44*, 1802–1813.

Ben-David, S. (2004). A framework for statistical clustering with a constant time approximation algorithms for k-median clustering. *Conference on Learning Theory (COLT).*

Buhmann, J. (1998). *Empirical risk approximation: An induction principle for unsupervised learning* (Technical Report IAI-TR-98-3). Universitaet Bonn.

Buhmann, J., & Held, M. (1999). Model selection in clustering by uniform convergence bounds. *Neural Information Processing Systems (NIPS)* (pp. 216–222).

Condon, & Karp (2001). Algorithms for graph partitioning on the planted partition model. *Random Structures & Algorithms*, *18*, 116–140.

Dasgupta, S. (1999). Learning mixtures of Gaussians. *IEEE Symposium on Foundations of Computer Science (FOCS)* (pp. 634–644).

Demaine, & Immorlica (2003). Correlation clustering with partial information. *International Workshop on Approximation Algorithms for Combinatorial Optimization (APPROX).*

Dhillon, I. (2001). Co-clustering documents and words using bipartite spectral graph partitioning. *ACM SIGKDD Conference.*

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM*, *46*, 604–632.

Linder, T., Lugosi, T., & Zeger, K. (1994). Rates of convergence in the source coding theorem, in empirical quantizer design, and in universal lossy source coding. *IEEE Transactions on Information Theory*, *40*, 1728–1740.

McSherry, F. (2001). Spectral partitioning of random graphs. *IEEE Symposium on Foundations of Computer Science (FOCS)*.

Ng, V., & Cardie, C. (2002). Improving machine learning approaches to coreference resolution. *Annual Meeting of the Assoc. for Comp. Linguistics (ACL)*.

Pollard, D. (1981). Strong consistency of k-means clustering. *The Annals of Statistics*, *9*, 135–140.

Swamy, C. (2004). Correlation clustering: Maximizing agreements via semidefinite programming. *Symposium on Discrete Algorithms (SODA)*.

von Luxburg, U., Bousquet, O., & Belkin, M. (2004). On the convergence of spectral clustering on random samples: The normalized case. *Conference on Learning Theory (COLT)* (pp. 457–471).

Figure 2. Illustration of planted partition model and the inference process.