Weighted Graph Clustering with Non-Uniform Uncertainties

Yudong Chen YUDONG.CHEN@EECS.BERKELEY.EDU
University of California, Berkeley. Berkeley, CA 94720, USA

Shiau Hong Lim MPELSH@NUS.EDU.SG
National University of Singapore, Singapore 117575

Huan Xu MPEXUH@NUS.EDU.SG
National University of Singapore, Singapore 117575

Abstract

We study the graph clustering problem where each observation (edge or no-edge between a pair of nodes) may have a different level of confidence/uncertainty. We propose a clustering algorithm that is based on optimizing an appropriate *weighted* objective, where larger weights are given to observations with lower uncertainty. Our approach leads to a convex optimization problem that is efficiently solvable. We analyze our approach under a natural generative model, and establish theoretical guarantees for recovering the underlying clusters. Our main result is a general theorem that applies to any given weight and distribution for the uncertainty. By optimizing over the weights, we derive a provably optimal weighting scheme, which matches the information theoretic lower bound up to logarithmic factors and leads to strong performance bounds in several specific settings. By optimizing over the uncertainty distribution, we show that non-uniform uncertainties can actually help. In particular, if the graph is built by spending a limited amount of resource to take measurement on each node pair, then it is beneficial to allocate the resource in a non-uniform fashion to obtain accurate measurements on a few pairs of nodes, rather than obtaining inaccurate measurements on many pairs. We provide simulation results that validate our theoretical findings.

1. Introduction

Graph clustering concerns with finding densely connected groups of nodes in a graph. Here an edge between two nodes usually indicates certain underlying similarity or affinity between nodes, whereas the absence of an edge indicates dissimilarity and distance. Therefore, the goal of graph clustering is to infer groups of closely related nodes given the (often noisy) similarity/dissimilarity observations encoded in the graph. Graph clustering is an important subroutine in many applications, such as community detection, user profiling and VLSI network partitioning (Mishra et al., 2007; Yahoo!-Inc, 2009; Krishnamurthy, 1984).

In many of these applications, however, the edge/non-edge between each node pair may represent very different levels of confidence of the similarity of the nodes. In some cases, the observation of an edge (the absence of an edge, resp.) may be generated by accurate measurements and thus is a strong indicator that the two nodes should be assigned the same (different, resp.) clusters. In other circumstances, the observations may be very uncertain and thus an edge or the absence of it provides little information about the cluster structure. As an extreme case, the observations between some node pairs may carry no information at all, so these pairs are effectively unobserved. An example with non-uniform uncertainties is crowd-clustering (Gomes et al., 2011; Yi et al., 2012), where a number of users are asked whether or not they think a pair of nodes (e.g., movies or images) are similar, and the final graph is obtained by aggregating the users' answers, for instance by taking a majority vote. The confidence level are naturally different across pairs: a pair receiving a large number of unanimous votes has a higher confidence level than those receiving few votes or divergent votes; in particular, pairs receiving no votes are completely uncertain.

In such a non-uniform setting, each pair of nodes should be treated differently according to the level of uncertainty

Proceedings of the 31^{st} *International Conference on Machine Learning*, Beijing, China, 2014. JMLR: W&CP volume 32. Copyright 2014 by the author(s).

between them. In many cases, a priori knowledge is available for the uncertainty levels since the graph is built from a known or controlled process. Intuitively, taking advantage of such knowledge should improve clustering performance.

Our contributions:

In this paper, we exploit the above intuition, and propose a new approach for clustering graphs with non-uniform edge uncertainties. Our approach is based on finding the clusters that optimizes an appropriate *weighted* objective, where larger weights are given to node pairs with lower levels of uncertainties. Doing so leads to an intractable combinatorial optimization problem, and our algorithm is a convex relaxation of it. To study the performance of our approach, we consider a natural probabilistic model for generating a random graph from an unknown set of clusters with non-uniform uncertainties. We provide theoretical guarantees for when the solution to the convex relaxation coincides with the combinatorial problem and exactly recovers the underlying clusters. Our main result is a general theorem that applies to *any given weights* and *any uncertainty distribution*. The theorem leads to a principled way of choosing the weights *optimally* based on knowledge of the uncertainties. By specializing this general result to different settings for the uncertainties, we recover the best known theoretic guarantees for clustering partially observed graphs, and obtain new results for more general non-uniform settings. In particular, we show that a weighted approach using the knowledge of the non-uniform uncertainties have order-wise better guarantees than an unweighted approach. We call this the "power of knowledge".

We use the above results to obtain theoretical insights to a *resource allocation* problem in graph clustering. As a corollary of our main theorem, it can be shown that the clustering problem becomes *easier* when the uncertainties are more non-uniform across the node pairs (provided that the knowledge of the uncertainty levels is appropriately exploited). Therefore, if the uncertainty level of the observation on each node pair depends on the resource devoted to it, then it is often more beneficial, sometimes significantly, to allocate the resource in a non-uniform way, with most of the resource spent on a small number of nodes so that they have low uncertainty levels.

We provide simulation results to corroborate our theoretical findings. The results demonstrate that the weighted approach and the optimal weights outperform other methods, and non-uniform resource allocation lead to performance gain.

1.1. Related Work

Planted partition model/Stochastic block model: The setup in this paper is closely related to the classical *planted partition model* (Condon & Karp, 2001), also known as the *stochastic block model* (Rohe et al., 2011). In these models, n nodes are partitioned into several groups called the underlying clusters, and a random graph is generated by placing an edge between each pair of nodes independently with probability p or q (with $p > q$) depending on whether the nodes belong to the same cluster. The goal is to recover the underlying clusters given the graph. Various approaches have been proposed for this problem, including spectral clustering algorithms (McSherry, 2001; Giesen & Mitsche, 2005; Chaudhuri et al., 2012; Rohe et al., 2011) and randomized combinatorial methods (Shamir & Tsur, 2007). Most related to us are the convex optimization approaches in Chen et al. (2012; 2011); Ames & Vavasis (2011); Oymak & Hassibi (2011); Jalali et al. (2011); Mathieu & Schudy (2010), which are similar to an unweighted version of our method. Except for a few exceptions detailed below, most previous work focused on the uniform uncertainty case.

Graph clustering with non-uniform uncertainty: Chaudhuri et al. (2012) explicitly consider non-uniform uncertainty under the planted partition model. They study the setup where each *node* is associated with a confidence d_i, and the probability of placing an edge between nodes i and j is $d_i p d_j$ ($d_i q d_j$ resp.) if i and j belong to the same cluster (different clusters resp.). A spectral clustering approach is proposed to tackle the problem. In our model the non-uniformity is *pair-wise* and need not have a product form as in their work.

As we explained earlier, clustering with partial observations can be considered as a special case of non-uniform uncertainties. Here, an edge or the absence of an edge is observed for a subset of the nodes pairs. For the other node pairs only a "?" is observed, meaning no information is available, which corresponds to maximum uncertainty in our non-uniform setting. A line of work explicitly addresses this setup. One natural approach, taken by Oymak & Hassibi (2011), is to convert the problem into one with uniform uncertainty by imputing the missing observations (either with no-edge or random choices), and then apply standard (unweighted) graph clustering methods. A more refined approach deals with the partial observations directly (Shamir & Tishby, 2011; Jalali et al., 2011). The work by Jalali et al. (2011) has the best known guarantees for the planted partition model with randomly missing observations. Their formulation is a special case of the more general method in this paper, and our theoretic results subsume theirs. There exists other work on clustering with partial observations (e.g., Balcan & Gupta, 2010; Voevodski et al., 2010; Krishnamurthy et al., 2012), but under rather different settings; it is also unclear how to generalize these methods to more general non-uniform uncertainty setting.

Another related line of work is *correlation clustering* (Bansal et al., 2004). There the goal is to find a set of clusters that minimize the total number of disagreements between the clusters and the observed graph. One may also consider minimizing a weighted sum of the disagreements (Demaine et al., 2006), leading to a weighted objective similar to ours. Work on correlation clustering focuses on establishing NP-hardness of the optimization problem and developing approximation schemes (Bansal et al., 2004; Giotis & Guruswami, 2006; Charikar et al., 2005). In contrast, we take a statistical approach akin to the planted partition model, and study conditions under which the underlying clusters can be recovered with high probability. Therefore, the theoretical results for correlation clustering are not directly comparable to ours.

Recovering sparse signals and low-rank matrices with non-uniform priors: The problem of matrix decomposition (Candès et al., 2011; Chandrasekaran et al., 2011) concerns with separating a low-rank matrix from sparse entry-wise errors of arbitrary magnitude. A standard approach is based on convex optimization using the trace norm (a.k.a. nuclear norm) as a convex surrogate of the rank function. Chen et al. (2013); Li (2013) consider extensions of the problem with unobserved entries. A similar approach has been applied to graph clustering (Jalali et al., 2011; Chen et al., 2011; Oymak & Hassibi, 2011; Mathieu & Schudy, 2010). Our clustering algorithm can be considered as separating a low rank matrix from sparse errors with a *non-uniform* prior. While our analysis is restricted to the graph clustering, our method and intuition may be relevant to the general matrix decomposition problem. To the best of our knowledge, matrix decomposition with general non-uniform error probability has not been studied in the literature.

A related problem is recovering of a sparse vector signal, for which non-uniform priors have been considered. In a standard setting, it is assumed that each entry of the unknown vector is known to have a different probability of being non-zero. Khajehnejad et al. (2011) propose a weighted ℓ_1 norm minimization approach to address this problem. The analysis of the approach mainly focuses on the special case where the non-zero probability can take one of two possible values (Khajehnejad et al., 2011; Oymak et al., 2011; Krishnaswamy et al., 2012). In this two-value setting, it is shown that the weighted approach is superior to an unweighted approach.

2. Problem Setup and Algorithms

In this section we provide a formal setup of the graph clustering problem with non-uniform uncertainty levels. We consider a probabilistic model for generating the graph and the edge uncertainties based on a set of underlying unknown clusters. We then present our weighted formulation for recovering the underlying clusters, which is efficiently solvable and weighs each node pair differently.

2.1. Model

Suppose there are n nodes which are partitioned into r unknown clusters of size *at least* K. We observe an unweighted graph of the nodes, given as an adjacency matrix $A \in \{0,1\}^{n\times n}$, which is generated as follows: For two nodes i and j in the same cluster, we observe an edge between them (i.e., $A_{ij} = 1$) with probability $1 - P_{ij}$, and no edge otherwise; for i and j in different clusters, we observe an edge between them with probability P_{ij}, and no edge otherwise. Therefore, P_{ij} can be considered as the probability of false observation between i and j, where a false observation is a missing edge between two nodes in the same clusters (a false negative), or an edge between two nodes in different clusters (a false positive).

We are interested in the case where the P_{ij}'s are potentially different across the (i,j)'s, so the uncertainties associate with each observation A_{ij} are non-uniform. In particular, $P_{ij} = 0$ means that A_{ij} is a noiseless observation of the cluster relation between the nodes i and j, whereas $P_{ij} = \frac{1}{2}$ means A_{ij} is purely random and thus the relation between i and j is unobserved. We use $P = (P_{ij})_{i,j=1}^{n} \in \mathbb{R}^{n\times n}$ to denote the matrix of error probabilities.

It is in general an ill-posed problem to recover the clusters for arbitrary error probabilities P. For example, if $P_{ij} = \frac{1}{2}, \forall j$ for some node i, then the graph contains no information about the node i so exact cluster recovery is impossible. To avoid such pathological situation, we assume that P is randomly generated with i.i.d. entries. In particular, for each (i,j) and independent of all others, P_{ij} is a random variable with some distribution $\mathbb{Q}$ supported on $[0, 1/2]$, where Q is either the corresponding probability mass function if P_{ij} takes discrete values or the probability density function if P_{ij} is continuous.

2.2. Our Algorithm

Our algorithm is based on finding a clustering of the nodes that optimizes an appropriate objective. To this end we need the notion of a *cluster matrix*: given a clustering of the n nodes, the associated cluster matrix is an n-by-n $0-1$ matrix Y such that $Y_{ij} = 1$ if and only if the nodes i and j are assigned to the same clusters. Let Y^* be the true cluster matrix associated with the underlying clusters that generate the graph A. Our goal is therefore to recover Y^* from the graph A. A natural approach, akin to *correlation clustering* (Bansal et al., 2004), is to find a clustering Y that maximizes the sum of the total number of edges inside the clusters and the the total number of missing edges across clusters. This can be written as maximizing the quantity

$$\begin{aligned} & \textstyle\sum_{i.j} A_{ij}Y_{ij} + \sum_{i.j}\left(1-A_{ij}\right)\left(1-Y_{ij}\right) \\ = \; & \textstyle\sum_{i.j}\left(2A_{ij}-1\right)Y_{ij} + C \end{aligned}$$

over all cluster matrix Y, where C collects the terms independent of Y. This formulation gives equal weights to each node pair. When the uncertainty levels are non-uniform, we consider instead the following weighted formulation:

$$\begin{aligned} \max_{Y\in\mathbb{R}^{n\times n}} \quad & \sum_{i,j} B_{ij}(2A_{ij}-1)Y_{ij} \\ \text{s.t.} \quad & Y \text{ is a cluster matrix}, \end{aligned} \tag{1}$$

where $B_{ij} \geq 0$ is the weight assigned to the pair (i,j).

The formulation (1) is a hard combinatorial optimization problem because there are exponentially many possible cluster matrices. To obtain a tractable formulation, we relax the constraint "Y is a cluster matrix" with a set of convex constraints. Specifically, we consider the following:

$$\max_{Y\in\mathbb{R}^{n\times n}} \quad \sum_{i,j} B_{ij}\left(2A_{ij}-1\right)Y_{ij} \tag{2}$$

$$\text{s.t.} \quad Y \in \mathcal{S}, \tag{3}$$

$$0 \leq Y_{ij} \leq 1, \qquad \forall i,j, \tag{4}$$

where $\mathcal{S}$ is a convex set that contains Y^*. We may use either one of the following choices:

$$\begin{aligned} \mathcal{S}_{\text{nuclear}} &= \left\{Y \in \mathbb{R}^{n\times n} : \|Y\|_* \leq n\right\}, \\ \mathcal{S}_{\text{psd}} &= \left\{Y \in \mathbb{R}^{n\times n} : \text{trace}(Y) \leq n, Y \succeq 0\right\}; \end{aligned} \tag{5}$$

here $\|Y\|_*$ is the trace norm (a.k.a. nuclear norm) of Y, defined as the sum of the singular values of Y, and $Y \succeq 0$ is the positive semidefinite constraint. Both $\mathcal{S}_{\text{nuclear}}$ and $\mathcal{S}_{\text{psd}}$ are standard convex relaxations for positive semidefinite low-rank matrices and cluster matrices (Mathieu & Schudy, 2010; Jalali et al., 2011). For both choices of $\mathcal{S}$, the formulation (2)–(4) is a semidefinite program (SDP) and can be solved in polynomial-time. Fast first-order solvers can also be used; we describe one such solver in the simulation section.

3. Theoretical Guarantees

In this section, we provide theoretical guarantees for the formulation (2)–(5) in recovering the true clusters Y^*. We first present a general main theorem that applies to any weights $\{B_{ij}\}$ and any distribution $\mathbb{Q}$ for the error probabilities $\{P_{ij}\}$. We next derive a *provably optimal* way of choosing the weights, and characterizes its performance using the general theorem. We then specialize our theorem to different settings of $\mathbb{Q}$ and $\{B_{ij}\}$.

In the sequel, *with high probability* (w.h.p.) means with probability at least $1 - n^{-10}$.

3.1. Main Theorem

We assume that the weights satisfy $B_{ij} = f(P_{ij})$ for some function f, so B_{ij} depends on the value of P_{ij} but not the location (i,j). In this case, the B_{ij}'s are in general random and identically distributed. A constant function $f(\cdot)$ corresponds to uniform weights that are independent of the error probabilities. We have the following general theorem.

Theorem 1. *Suppose there exists $b > 0$ such that $0 \leq B_{ij} \leq b$ almost surely for all (i,j). Then Y^* is the unique optimal solution to the program (2)–(5) with high probability if for all (i,j) and a universal constant c_0,*

$$\mathbb{E}\left[\left(\frac{1}{2} - P_{ij}\right)B_{ij}\right] \geq c_0\left(\frac{b\log n}{K} + \frac{\sqrt{\mathbb{E}\left[B_{ij}^2\right] n\log n}}{K}\right). \tag{6}$$

Here the theorem applies to both choices in (5), and the expectations and probability are w.r.t. the randomness of $\{P_{ij}\}$, $\{B_{ij}\}$ and A.

Remark 1. *The condition (6) is identical for different (i,j) since the (P_{ij}, B_{ij})'s are identically distributed. This remark applies to any expression that involves the expectations or distributions of P_{ij}, B_{ij} and A_{ij}.*

We note that b in the theorem can be any number and is allowed to scale with n and K etc.

3.2. Optimal Weighting and Maximum Likelihood Estimators

We now turn to the question of how to choose the weights $B_{ij} = f(P_{ij})$ or equivalently the function $f(\cdot)$. Under the generative model in Section 2.1, a natural candidate is to consider the Maximum Likelihood objective, which we now show is a special case of our weighted objective (2). Given the graph A and the error probabilities P, the Maximum Likelihood Estimator (MLE) searches for a cluster matrix Y which maximizes the log likelihood $\log\mathbb{P}_Y(A)$ of observing A given Y, where

$$\begin{aligned} & \log\mathbb{P}_Y(A) \\ =& \sum_{i,j}\log\Big[\left(1-P_{ij}\right)^{A_{ij}Y_{ij}}\left(P_{ij}\right)^{(1-A_{ij})Y_{ij}} \\ & \qquad \times \left(P_{ij}\right)^{A_{ij}(1-Y_{ij})}\left(1-P_{ij}\right)^{(1-A_{ij})(1-Y_{ij})}\Big] \\ =& \sum_{i,j}\left(2A_{ij}-1\right)Y_{ij}\log\left(\frac{1-P_{ij}}{P_{ij}}\right) + C', \end{aligned}$$

where C' collects the terms that are independent of Y. Therefore, the MLE objective corresponds to our objective (2) with the weights $B_{ij} = \log\left(\frac{1-P_{ij}}{P_{ij}}\right)$. In fact, we may use any *upper bound* $\bar{P}_{ij} \leq \frac{1}{2}$ of the exact error probability P_{ij}, which provides additional flexibility in the sequel. We refer to this as the *MLE weights*, namely

$$B_{ij}^{\mathrm{MLE}} := \log\left(\frac{1-\bar{P}_{ij}}{\bar{P}_{ij}}\right), \text{ for each } (i,j). \tag{7}$$

Remark 2. *The MLE weight B_{ij}^{MLE} has a natural interpretation. When $\bar{P}_{ij} = P_{ij} = \frac{1}{2}$, the observation A_{ij} on the pair (i,j) is purely random. In this case, $B_{ij}^{MLE} = \log\left((1-\frac{1}{2})/\frac{1}{2}\right) = 0$ so we assign zero weight to the pair (i,j). If $\bar{P}_{ij} = P_{ij} \to 0$, then A_{ij} is noiseless and we have exact knowledge about the cluster relationship between i and j; in particular $A_{ij} = 1$ if and only if i and j are in the same underlying cluster. In this case, the MLE weight satisfies $B_{ij}^{MLE} \to +\infty$, so by maximizing $B_{ij}\left(2A_{ij}-1\right)Y_{ij}$, we force Y_{ij} to equal A_{ij}, agreeing with the exact knowledge we possess.*

Our analysis shows that the weights B_{ij}^{MLE} are order-wise optimal under certain technical conditions, in the sense that its performance is (order-wise) at least as good as any other choice of for the weights B_{ij}. In fact, it is order-wise as good as *any algorithm* for recovering Y^*. To see this, we first derive a guarantee for the MLE weights. The following corollary is proved using the general Theorem 1. Recall that $\bar{P}_{ij} \le \frac{1}{2}$ is any upper bound on P_{ij}.

Corollary 1. *Suppose $\frac{1}{2} \ge \bar{P}_{ij} \ge \epsilon, \forall (i,j)$ almost surely for some $\frac{1}{4} \ge \epsilon > 0$. The weighted formulation (2)–(5) with the MLE weights $B_{ij} = B_{ij}^{MLE}$ has a unique optimal solution equal to Y^* with high probability provided*

$$\mathbb{E}\left[\left(\frac{1}{2}-\bar{P}_{ij}\right)\log\frac{1-\bar{P}_{ij}}{\bar{P}_{ij}}\right] \ge c_1\frac{n}{K^2}\log\left(\frac{1}{\epsilon}\right)\log n, \forall (i,j) \tag{8}$$

for some universal constant c_1. Moreover, the condition (8) is satisfied if we take $\bar{P}_{ij} = \max\left\{P_{ij}, \frac{1}{16}\right\}$ in (7) and

$$\mathbb{E}\left[\left(\frac{1}{2}-P_{ij}\right)^2\right] \ge c_2\frac{n}{K^2}\log n, \quad \forall (i,j). \tag{9}$$

Again we note that ϵ in the corollary may scale with n and K, and need not be bounded away from zero.

We next establish a theorem that characterizes the performance limit of *all* algorithms. The theorem provides a lower bound on the minimax error probability and is proved by an information-theoretic argument. It generalizes similar lower bounds by Chaudhuri et al. (2012); Chen et al. (2011) for the uniform uncertainty setting.

Theorem 2. *Suppose there are $r = 2$ clusters with equal size $K = n/2$. Let $\mathcal{Y} := \{Y : Y \text{ is a cluster matrix}\}$ be the set of all possible cluster matrices. If*

$$\mathbb{E}\left[\left(\frac{1}{2}-P_{ij}\right)^2\right] \le c'\frac{1}{n}, \quad \forall (i,j) \tag{10}$$

for some universal constant c', then we have

$$\inf_{\hat{Y}}\sup_{Y^*\in\mathcal{Y}} \mathbb{P}\left[\hat{Y}\left(A,P\right) \ne Y^*\right] \ge \frac{1}{2},$$

where the infimum is over all measurable functions $\hat{Y}$ that map (A,P) to an element of $\mathcal{Y}$, and the probability $\mathbb{P}\left[\cdot\right]$ is w.r.t. the randomness of A and P.

Theorem 2 shows that in the case with $r = 2$ clusters, any algorithm fails with positive probability if (10) holds. In this setting, Corollary 1 guarantees that the formulation (2)–(4) with the MLE weights succeeds w.h.p. if

$$\mathbb{E}\left[\left(1/2-P_{ij}\right)^2\right] \gtrsim \frac{\log n}{n}, \quad \forall (i,j).$$

This matches the condition (10) up to a logarithmic factor, and thus cannot be substantially improved. This shows that the MLE weights is order-wise optimal in this case. We expect this to be true generally. Indeed, our simulations show that the MLE weights do outperform other weight schemes in a broad range of settings.

3.3. Consequences

3.3.1. The Power of Knowledge

The above results characterize the benefit of utilizing the knowledge of P via a weighted approach as compared to an unweighted approach that ignores P. Suppose $P_{ij} \ge \epsilon_0$ for some universal constant $\epsilon_0 > 0$. If we use uniform weights $B_{ij} \equiv 1$ corresponding to an unweighted formulation, then Theorem 1 shows that the formulation succeeds w.h.p. if

$$\left(\frac{1}{2}-\mathbb{E}\left[P_{ij}\right]\right)^2 \gtrsim \frac{\log^2 n}{K^2} + \frac{n\log n}{K^2}, \quad \forall (i,j). \tag{11}$$

On the other hand, if we have access to knowledge of P_{ij}, we may use the optimal MLE weights $B_{ij} = B_{ij}^{\mathrm{MLE}}$, and Corollary 1 guarantees that the weighted formulation succeeds w.h.p. as long as the condition (9) is satisfied. The RHS of (9) is always no greater than the RHS of (11). More importantly, the LHS of (9) is strictly larger than the LHS of (11) whenever P_{ij} is not equal to a constant almost surely, because $\mathbb{E}\left[P_{ij}^2\right] > \left(\mathbb{E}\left[P_{ij}\right]\right)^2$. The gap is exactly the variance of P_{ij}, and is large precisely when the error probability is far from being uniform.

3.3.2. Two-level Uncertainty and Partial Observations

We consider a more concrete setting where P_{ij} is non-uniform and can take one of two values. In particular, we assume that $P_{ij} = p_1$ with probability q and $P_{ij} = p_2$ with probability $1-q$, where $p_1 < p_2$. If we use uniform weighting $B_{ij} \equiv 1$, then by applying (11) and computing the expectation, we obtain that the unweighted convex formulation succeeds if

$$\left(\frac{1}{2}-p_2\right)^2+2q\left(\frac{1}{2}-p_2\right)(p_2-p_1)+q^2\,(p_2-p_1)^2 \gtrsim \frac{n\log^2 n}{K^2}.$$

If we use the optimal weights B_{ij}^{MLE}, then (9) in Corollary 1 guarantees that the weighted formulation succeeds if

$$\left(\frac{1}{2}-p_2\right)^2+2q\left(\frac{1}{2}-p_2\right)(p_2-p_1)+q\,(p_2-p_1)^2 \gtrsim \frac{n\log^2 n}{K^2}.$$

Note that the left hand side becomes strictly larger as $q(p_2-p_1)^2 > q^2(p_2-p_1)^2$, and hence the weighted formulation succeeds for a wider range of the model parameters.

A special case of the above setting is when $p_2=\frac{1}{2}$. This means that with probability $1-q$, A_{ij} is purely random and effectively unobserved, and with probability q it is observed but contains an error with probability p_1. This coincide with the graph clustering with partial observation problem that has been studied before. In this case the unweighted approach and the weighted approach require

$$q^2\,(1/2-p_1)^2 \gtrsim n\log^2 n/K^2$$

and

$$q\,(1/2-p_1)^2 \gtrsim n\log^2 n/K^2,$$

respectively. Therefore, in the case with $K=\Theta(n)$ and $p_1=\frac{1}{4}$, the weighted approach can handle as few as $qn^2=\Theta\left(n\log^2 n\right)$ observations, whereas the unweighted approach requires $\Theta\left(n\sqrt{n}\log n\right)$ observations, which is order-wise larger. We note that in this case the MLE weights is equivalent to assigning zero weight to unobserved node pairs and uniform positive weight to observe pairs. Our result matches the best existing bounds for the partial observation setting given in Jalali et al. (2011); Chen et al. (2011).

4. Resource Allocation in Graph Clustering

Our results in the previous section show that given a graph with known uncertainties, one could achieve better performance guarantee by employing appropriate weights in solving the optimization problem. Here, we look at a complementary problem: Suppose we have the ability to control the uncertainties by spending some available resource (e.g., performing queries or measurements) to build the graph, how should one allocate this limited amount of resource in order to optimize the performance guarantee for the cluster recovery procedure? We show that our theoretical results can provide a principled solution to this resource allocation problem.

Suppose that we wish to recover the underlying clusters by first assigning the probability distribution P_{ij} for each pair of nodes and then using our proposed weighted scheme. The required amount of resource M_{ij} should naturally be higher if the error probability P_{ij} is small, and vice versa. The exact relationship between M_{ij} and P_{ij} depends on the particular setup being considered. By Corollary 1, in order to maximize the probability of successful recovery, we should aim to maximize $\mathbb{E}[(\frac{1}{2}-P_{ij})^2]$ over all possible distributions on P_{ij}, subject to our resource constraint. We examine several different scenarios for the relationship between P_{ij} and M_{ij}.

Model 1: We first consider a linear model $M_{ij}=\gamma(\frac{1}{2}-P_{ij})$ or equivalently $P_{ij}=\frac{1}{2}-\frac{M_{ij}}{\gamma}$, where $\gamma>0$. Note that $M_{ij}=0$ implies $P_{ij}=\frac{1}{2}$, which is equivalent to an unobserved pair. Let $M=\mathbb{E}\sum_{i<j}M_{ij}$ be the expected total amount of available resource. This implies that in our probabilistic model, the expected P_{ij} for each error probability for each node pair $\mathbb{E}[P_{ij}]=\frac{1}{2}-\frac{2M}{n(n-1)\gamma}=c$ is a constant. Note that the expectation is with respect to the distribution $Q(P_{ij})$ on P_{ij} which is supported on $[0,1/2]$. By Corollary 1, we therefore wish to maximize the LHS of (9) subject to the resource constraint. This is equivalent to solving the following variational problem:

$$\begin{aligned}\max_{Q}\quad & \mathbb{E}_Q[P_{ij}^2]\\ \text{s.t.}\quad & \mathbb{E}_Q[P_{ij}]=c.\end{aligned}$$

(We note again that expressions like the one above are independent of i and j, since the P_{ij}'s are assumed to be identically distributed.) The problem has a simple solution. Due to the convexity of the function P_{ij}^2, it is easy to show that the optimal distribution is to place all the probability mass on $P_{ij}=0$ and $P_{ij}=1/2$ such that the expectation is c, in other words $Q(0)=1-2c$ and $Q(1/2)=2c$. This shows that instead of spending the resource uniformly to obtain many "moderately certain" observations, one should spend all the resource on a small number of accurate observations. We summarize the above with the following corollary:

Corollary 2. *For the resource allocation problem with $M_{ij}=\gamma(\frac{1}{2}-P_{ij})$ and per-pair resource $\mathbb{E}M_{ij}=\alpha$, the order-wise optimal distribution on P_{ij} is $Q(0)=1-2c$ and $Q(1/2)=2c$ where $c=\frac{1}{2}-\frac{\alpha}{\gamma}$.*

Model 2: Another natural model assumes that M_{ij} is inversely proportional to P_{ij}. This model is motivated by the central limit theorem which asserts that the variance of the mean of independent observations decreases with the inverse of the number of observations. More precisely, we assume $M_{ij} = \frac{1}{P_{ij}} - 2$, so $P_{ij} = 1/2$ corresponds to no observation at all and $M_{ij} = 0$ resource is needed. Again, by Corollary 1 we seek to maximize $\mathbb{E}[(\frac{1}{2} - P_{ij})^2]$ subject to the constraint $\mathbb{E}\sum_{i<j} M_{ij} = M$. The general optimal distribution is complicated and we therefore consider a simplified scenario where the resource will be spent uniformly on a random subset S of all $n(n-1)/2$ pairs, such that each pair (i,j) is in S with probability β. Let $\alpha = \frac{2M}{n(n-1)}$ be the amount of per-pair resource. In this case, $M_{ij} = \frac{\alpha}{\beta}$ with probability β and $M_{ij} = 0$ otherwise. Equivalently, $P_{ij} = \frac{\beta}{\alpha + 2\beta}$ with probability β and $P_{ij} = \frac{1}{2}$ otherwise. Given α, the objective is then to maximize

$$\mathbb{E}\left[\left(\frac{1}{2} - P_{ij}\right)^2\right] = \beta\left(\frac{1}{2} - \frac{\beta}{\alpha + 2\beta}\right)^2$$

by choosing the fraction β of selected pairs. A little calculus reveals that the maximum is attained at $\beta = \frac{\alpha}{2}$ when $\alpha \leq 2$ and $\beta = 1$ otherwise. We therefore have the following:

Corollary 3. *For the resource allocation problem with $M_{ij} = \frac{1}{P_{ij}} - 2$, suppose that we select each node pair with probability β and allocate the total resource M uniformly to the selected pairs to achieve the same uncertainty $P_{ij} = \frac{\beta}{\alpha+2\beta}$ where $\alpha = \frac{M}{n^2}$. The optimal performance is achieved with $\beta = \frac{\alpha}{2}$ if $\alpha \leq 2$ and $\beta = 1$ otherwise.*

5. Empirical Results

To empirically evaluate our theoretic findings in Section 3, we follow Chen et al. (2012) to adapt the Augmented Lagrangian Multiplier algorithm by Lin et al. (2009) to solve our weighted program (2)–(5). In particular, we may replace the constraint $\|Y\|_* \leq n$ by a regularization term in the objective function and solve the following program:

$$\begin{aligned} \min_{Y \in \mathbb{R}^{n \times n}} \quad & \|Y\|_* + \lambda \sum_{ij} B_{ij}|A_{ij} - Y_{ij}| \\ \text{s.t.} \quad & 0 \leq Y_{ij} \leq 1, \forall i, j, \end{aligned}$$

where λ is chosen to be small enough (usually around $1/\sqrt{n}$) such that the solution satisfies $\|Y\|_* \leq n$.

In our simulations, we repeat each test 100 times and report the success rate, which is the fraction of attempts where the true clustering is recovered. Error bars show 95% confidence interval. In all our experiments, we choose to report results from parameter regions where the problem is neither too hard nor too easy to solve. We note that qualitatively the results are similar across a wide range of distributions.

We test our theory by comparing three different weighting schemes for graph clustering with non-uniform uncertainties. In particular, we compare the weighted formulation with the MLE weights with the unweighted formulation. The third candidate is a step weighting schemes where the different observation uncertainties are quantized into 3 levels, each with a different weight, with higher weights on the more certain observations.

The ground truth consists of n nodes divided into 4 equal-size clusters. We tested with $n = 200$ and $n = 1000$. The following stochastic model is used: for each pair of nodes, with probability q it is unobserved (i.e., $P_{ij} = 0.5$), otherwise the uncertainties vary uniformly at random between $0.26 - \Delta$ and $0.26 + \Delta$. The graph is then generated according to the model in Section 2.1 with error probability P_{ij}. For $n = 200$ we use $q = 0.6$ while for $n = 1000$ we use $q = 0.7$ since the problem is easier for larger n. For step weights, we split the range $(0.26 - \Delta, 0.26 + \Delta)$ to 3 equal intervals $(0.26 - \Delta, 0.26 - \frac{\Delta}{3})$, $(0.26 - \frac{\Delta}{3}, 0.26 + \frac{\Delta}{3})$ and $(0.26 + \frac{\Delta}{3}, 0.26 + \Delta)$. All P_{ij} in the same interval receives the same weight, which is is based on the MLE weight of the largest P_{ij} in the interval.

Note that the expected total number of errors is the same for all Δ while the range of uncertainties increases with Δ. Figures 1 and 2 show the results for $n = 200$ and 1000.

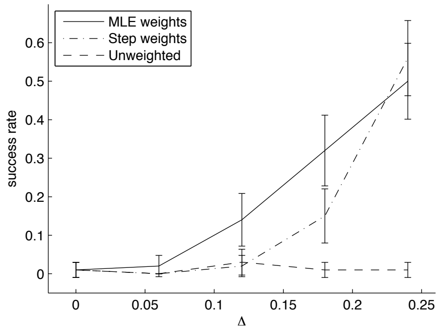

Figure 1. Success rates of different weighting schemes ($n = 200$)

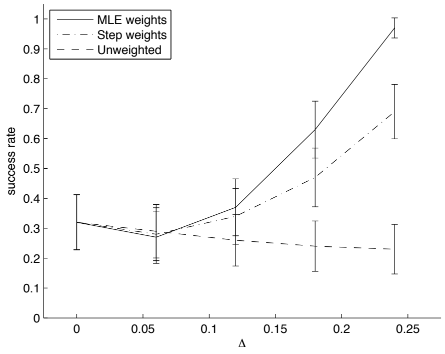

Figure 2. Success rates of different weighting schemes ($n = 1000$)

The results show that the success rate is higher when weighting is used, especially when the uncertainties in the observations are non-uniform. Furthermore, it shows that the MLE weights outperforms the step/quantized weights. This shows that not all weighting schemes are equal, and the largest performance gain is achieved when the uncertainties are most non-uniform. This is consistent with our theoretical prediction.

Also, recall that in the resource allocation problem with the linear model $M_{ij} = \alpha - \beta P_{ij}$, we showed that the optimal choice of P_{ij} subject to the constraint that $\mathbb{E}(P_{ij})$ is a constant would be to maximize the variance. The results in Figures 1 and 2 confirm this choice since $\mathbb{E}(P_{ij})$ is the same for all Δ and the best recovery performance is achieved when Δ is the largest.

To test the clustering performance under various cluster size K, we run another set of experiments on 200 nodes with a fix uncertainty distribution but with different cluster sizes. Figure 3 shows the results of 4 different weighting schemes. The distribution of P_{ij} is such that 20% of the pairs are unobserved, and among the rest, P_{ij} is either 0.1 or 0.4, each with 0.5 probability. The MLE weight is again $B_{ij} = \log \frac{1-P_{ij}}{P_{ij}}$. "Weight A" reduces the weight for $P_{ij} = 0.1$ and increases the weight for $P_{ij} = 0.4$ such that their relative strength is only half compared to that of the MLE weights. On the other hand, "Weight B" increases the relative strength such that it is double that of the MLE weights. The results in Fig. 3 show that the MLE weights indeed achieves the best performance across various cluster sizes while the unweighted solution performs worst.

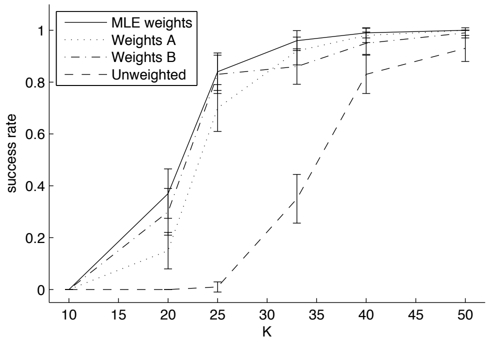

Figure 3. Success rates with respect to changing cluster sizes

Our last experiment concerns the analysis regarding resource allocation under Model 2. Figure 4 shows results for $\alpha = 1.6$ on 100 nodes, with a wide range of β. Indeed, with the same resource constraint, the best success rate in recovering the underlying graph is achieved around $\beta = 0.8$ as predicted by Corollary 3. It is interesting to note that observing 80% of the data with error rate 0.25 (overall error rate 0.3) actually results in a better success rate than observing all the data with error rate 0.28. This result is not obvious a priori and demonstrates the predictive power of our theoretical results.

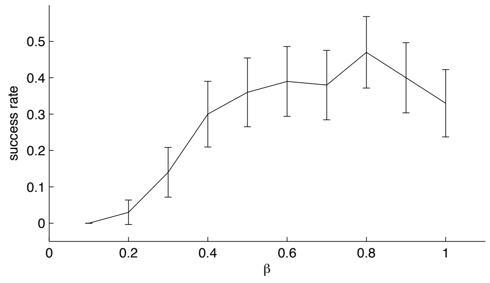

Figure 4. Verifying the predicted optimal resource allocation under Model 2.

6. Conclusion

We studied the graph clustering problem where observations have different levels of confidence. We proposed an (computationally tractable) approach based on optimizing an appropriate weighted objective. Our analysis establishes a general theoretical guarantee for correct recovery of the clustering structure, which applies to any weighting scheme and any uncertainty distribution. The general theorem leads to a provably optimal weighting scheme, and applies to several specific settings including partial observation and resource allocation.

Our approach and analysis highlight the "power of knowledge" and the "concentration gain": using prior knowledge of the uncertainty levels improves performance, and a few accurate measurements are better than many inaccurate measurements.

This paper focuses on graph clustering. The stepping stone of our approach is low-rank-and-sparse matrix decomposition based on nuclear norm relaxations. It is of interest to extend our methods and analysis to general matrix decomposition problems with non-uniform priors.

Acknowledgments

This research is partially supported by the Ministry of Education of Singapore through AcRF Tier Two grant R-265-000-443-112.

References

Ames, B. and Vavasis, S. Nuclear norm minimization for the planted clique and biclique problems. *Mathematical Programming*, 129(1):69–89, 2011.

Balcan, M.F. and Gupta, P. Robust hierarchical clustering. In *The Conference on Learning Theory (COLT)*, 2010.

Bansal, N., Blum, A., and Chawla, S. Correlation clustering. *Machine Learning*, 56(1):89–113, 2004.

Candès, E., Li, X., Ma, Y., and Wright, J. Robust principal component analysis? *J. ACM*, 58:1–37, 2011.

Chandrasekaran, V., Sanghavi, S., Parrilo, S., and Willsky, A. Rank-sparsity incoherence for matrix decomposition. *SIAM J. on Optimization*, 21(2):572–596, 2011.

Charikar, Moses, Guruswami, Venkatesan, and Wirth, Anthony. Clustering with qualitative information. *J. Comput. Syst. Sci.*, 71(3):360–383, 2005.

Chaudhuri, K., Chung, F., and Tsiatas, A. Spectral clustering of graphs with general degrees in the extended planted partition model. *COLT*, 2012.

Chen, Yudong, Jalali, Ali, Sanghavi, Sujay, and Xu, Huan. Clustering partially observed graphs via convex optimization. *Arxiv preprint arXiv:1104.4803*, 2011.

Chen, Yudong, Sanghavi, Sujay, and Xu, Huan. Clustering sparse graphs. In *NIPS 2012. Available on arXiv:1210.3335*, 2012.

Chen, Yudong, Jalali, Ali, Sanghavi, Suj, and Caramanis, Constantine. Low-rank matrix recovery from errors and erasures. *IEEE Trans. Information Theory*, 59(7), 2013.

Condon, A. and Karp, R.M. Algorithms for graph partitioning on the planted partition model. *Random Structures and Algorithms*, 18(2):116–140, 2001.

Demaine, Erik D., Emanuel, Dotan, Fiat, Amos, and Immorlica, Nicole. Correlation clustering in general weighted graphs. *Theoretical Computer Science*, 2006.

Giesen, J. and Mitsche, D. Reconstructing many partitions using spectral techniques. In *FOCS*, 2005.

Giotis, Ioannis and Guruswami, Venkatesan. Correlation clustering with a fixed number of clusters. *Theory of Computing*, 2(1):249–266, 2006.

Gomes, Ryan G, Welinder, Peter, Krause, Andreas, and Perona, Pietro. Crowdclustering. In *NIPS*, 2011.

Jalali, Ali, Chen, Yudong, Sanghavi, Sujay, and Xu, Huan. Clustering partially observed graphs via convex optimization. In *ICML*, 2011.

Khajehnejad, M. A., Xu, Weiyu, Avestimehr, A. S., and Hassibi, Babak. Analyzing weighted ℓ_1 minimization for sparse recovery with nonuniform sparse models. *IEEE Transactions on Signal Processing*, 59(5), 2011.

Krishnamurthy, A., Balakrishnan, S., Xu, M., and Singh, A. Efficient active algorithms for hierarchical clustering. *arXiv preprint arXiv:1206.4672*, 2012.

Krishnamurthy, B. An improved min-cut algorithm for partitioning VLSI networks. *IEEE Trans. Computers*, 1984.

Krishnaswamy, Anilesh K, Oymak, Samet, and Hassibi, Babak. A simpler approach to weighted ℓ_1 minimization. In *ICASSP*, pp. 3621–3624, 2012.

Li, Xiaodong. Compressed sensing and matrix completion with constant proportion of corruptions. *Constructive Approximation*, 37(1):73–99, 2013.

Lin, Z., Chen, M., Wu, L., and Ma, Y. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. Technical Report UILU-ENG-09-2215, UIUC, 2009.

Mathieu, C. and Schudy, W. Correlation clustering with noisy input. In *Proceedings of the 21st Annual ACM-SIAM Symposium on Discrete Algorithms*, 2010.

McSherry, F. Spectral partitioning of random graphs. In *FOCS*, pp. 529–537, 2001.

Mishra, N., R. Schreiber, I. Stanton, and Tarjan, R. E. Clustering social networks. *Algorithms and Models for Web-Graph, Springer*, 2007.

Oymak, S. and Hassibi, B. Finding dense clusters via low rank + sparse decomposition. arXiv:1104.5186v1, 2011.

Oymak, Samet, Khajehnejad, M. Amin, and Hassibi, Babak. Weighted compressed sensing and rank minimization. In *ICASSP*, pp. 3736–3739, 2011.

Rohe, K., Chatterjee, S., and Yu, B. Spectral clustering and the high-dimensional stochastic block model. *Annals of Statistics*, 39:1878–1915, 2011.

Shamir, O. and Tishby, N. Spectral Clustering on a Budget. In *AISTATS*, 2011.

Shamir, R. and Tsur, D. Improved algorithms for the random cluster graph model. *Random Structures and Algorithms*, 31(4):418–449, 2007.

Tropp, J.A. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

Voevodski, K., Balcan, M.F., Roglin, H., Teng, S.H., and Xia, Y. Efficient clustering with limited distance information. *arXiv preprint arXiv:1009.5168*, 2010.

Yahoo!-Inc. Graph partitioning. Available at http://research.yahoo.com/project/2368, 2009.

Yi, Jinfeng, Jin, Rong, Jain, Anil K, and Jain, Shaili. Crowdclustering with sparse pairwise labels: A matrix completion approach. In *AAAI Workshop on Human Computation*, 2012.