Guarantees for Spectral Clustering with Fairness Constraints

Matthäus Kleindessner [1] **Samira Samadi** [2] **Pranjal Awasthi** [1] **Jamie Morgenstern** [2]

Abstract

Given the widespread popularity of spectral clustering (SC) for partitioning graph data, we study a version of constrained SC in which we try to incorporate the fairness notion proposed by Chierichetti et al. (2017). According to this notion, a clustering is fair if every demographic group is approximately proportionally represented in each cluster. To this end, we develop variants of both normalized and unnormalized constrained SC and show that they help find fairer clusterings on both synthetic and real data. We also provide a rigorous theoretical analysis of our algorithms on a natural variant of the stochastic block model, where h groups have strong inter-group connectivity, but also exhibit a "natural" clustering structure which is fair. We prove that our algorithms can recover this fair clustering with high probability.

1. Introduction

Machine learning (ML) has recently seen an explosion of applications in settings to guide or make choices directly affecting people. Examples include applications in lending, marketing, education, and many more. Close on the heels of the adoption of ML methods in these everyday domains have been any number of examples of ML methods displaying unsavory behavior towards certain demographic groups. These have spurred the study of *fairness* of ML algorithms. Numerous mathematical formulations of fairness have been proposed for supervised learning settings, each with their strengths and shortcomings in terms of what they disallow and how difficult they may be to satisfy (e.g., Dwork et al., 2012; Hardt et al., 2016; Kleinberg et al., 2017; Zafar et al., 2017). Somewhat more recently, the community has begun to study appropriate notions of fairness for unsupervised learning settings (e.g., Chierichetti et al., 2017; Celis et al., 2018a;b; Samadi et al., 2018; Kleindessner et al., 2019).

In particular, the work of Chierichetti et al. (2017) proposes a notion of fairness for clustering: namely, that each cluster has proportional representation from different demographic groups. Their paper provides approximation algorithms that incorporate this fairness notion for k-center and k-median clustering. The follow-up work of Schmidt et al. (2018) extends this to k-means clustering. These papers open up an important line of work that aims at studying the following questions for clustering: a) *How to incorporate fairness constraints into popular clustering objectives and algorithms?* and b) *What is the price of fairness?* For example, the experimental results of Chierichetti et al. (2017) indicate that fair clusterings come with a significant increase in the k-center or k-median objective value. While the above works focus on clustering data sets in Euclidean / metric spaces, many clustering problems involve graph data. On such data, spectral clustering (SC; von Luxburg, 2007) is the method of choice in practice. In this paper, we extend the above line of work by studying the implications of incorporating the fairness notion of Chierichetti et al. (2017) into SC.

The contributions of this paper are as follows:

- We show how to incorporate the constraints that in each cluster, every group should be represented with the same proportion as in the original data set into the SC framework. For continuity with prior work (as discussed above; also see Section 5), we refer to these constraints as *fairness* constraints and speak of *fair* clusterings. However, the terms *proportionality* and *proportional* would be a more formal description of our goal. Our approach to incorporate the fairness constraints is analogous to existing versions of constrained SC that try to incorporate must-link constraints (see Section 5). In contrast to the work of Chierichetti et al. (2017), which always yields a fair clustering no matter how much the objective value increases compared to an unfair clustering, our approach does not guarantee that we end up with a fair clustering. Rather, our approach guides SC to find a good *and* fair clustering if such a one exists.
- Indeed, we prove that our algorithms find a good and fair clustering in a natural variant of the famous stochastic block model that we propose. In our variant, h demographic

[1]Department of Computer Science, Rutgers University, NJ [2]College of Computing, Georgia Tech, GA. Correspondence to: Matthäus Kleindessner <matthaeus.kleindessner@rutgers.edu>, Samira Samadi <ssamadi6@gatech.edu>, Pranjal Awasthi <pranjal.awasthi@rutgers.edu>, Jamie Morgenstern <jamiemmt@cs.gatech.edu>.

Proceedings of the 36th International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

groups have strong inter-group connectivity, but also exhibit a "natural" clustering structure that is fair. We provide a rigorous analysis of our algorithms showing that they can recover this fair clustering with high probability. To the best of our knowledge, such an analysis has not been done before for constrained versions of SC.

- We conclude by giving experimental results on real-world data sets where proportional clustering can be a desirable goal, comparing the proportionality and objective value of standard SC to our methods. Our experiments confirm that our algorithms tend to find fairer clusterings compared to standard SC. A surprising finding is that in many real data sets *achieving higher proportionality often comes at minimal cost*, namely, that our methods produce clusterings that are fairer, but have objective values very close to those of clusterings produced by standard SC. This complements the results of Chierichetti et al. (2017), where achieving fairness constraints *exactly* comes at a significant cost in the objective value, and indicates that in some scenarios fairness and objective value need not be at odds with one other.

Notation For $n \in \mathbb{N}$, we use $[n] = \{1, \dots, n\}$. I_n denotes the $n \times n$-identity matrix and $0_{n\times m}$ is the $n \times m$-zero matrix. $\mathbf{1}_n$ denotes a vector of length n with all entries equaling 1. For a matrix $A \in \mathbb{R}^{n\times m}$, we denote the transpose of A by $A^T \in \mathbb{R}^{m\times n}$. For $A \in \mathbb{R}^{n\times n}$, $\mathrm{Tr}(A)$ denotes the trace of A, that is $\mathrm{Tr}(A) = \sum_{i=1}^n A_{ii}$. If we say that a matrix is positive (semi-)definite, this implies that the matrix is symmetric.

2. Spectral Clustering

To set the ground and introduce terminology, we review spectral clustering (SC). There are several versions of SC (von Luxburg, 2007). For ease of presentation, here we focus on unnormalized SC (Hagen & Kahng, 1992). In Appendix A, we adapt all findings of this section and the following Section 3 to normalized SC (Shi & Malik, 2000).

Let $G = (V, E)$ be an undirected graph on $V = [n]$. We assume that each edge between two vertices i and j carries a positive weight $W_{ij} > 0$ encoding the strength of similarity between the verices. If there is no edge between i and j, we set $W_{ij} = 0$. We assume that $W_{ii} = 0$ for all $i \in [n]$. Given $k \in \mathbb{N}$, unnormalized SC aims to partition V into k clusters with minimum value of the RatioCut objective function as follows (see von Luxburg, 2007, for details): for a clustering $V = C_1 \dot\cup \dots \dot\cup C_k$ we have

$$\mathrm{RatioCut}(C_1, \dots, C_k) = \sum_{l=1}^{k} \frac{\mathrm{Cut}(C_l, V \setminus C_l)}{|C_l|}, \quad (1)$$

where

$$\mathrm{Cut}(C_l, V \setminus C_l) = \sum_{i\in C_l, j\in V\setminus C_l} W_{ij}.$$

Algorithm 1 Unnormalized SC

Input: weighted adjacency matrix $W \in \mathbb{R}^{n\times n}$; $k \in \mathbb{N}$

Output: a clustering of $[n]$ into k clusters

- compute the Laplacian matrix $L = D - W$
- compute the k smallest (respecting multiplicities) eigenvalues of L and the corresponding orthonormal eigenvectors (written as columns of $H \in \mathbb{R}^{n\times k}$)
- apply k-means clustering to the rows of H

Let $W = (W_{ij})_{i,j\in[n]}$ be the weighted adjacency matrix of G and D be the degree matrix, that is a diagonal matrix with the vertex degrees $d_i = \sum_{j\in[n]} W_{ij}$, $i \in [n]$, on the diagonal. Let $L = D - W$ denote the unnormalized graph Laplacian matrix. Note that L is positive semi-definite. A key insight is that if we encode a clustering $V = C_1 \dot\cup \dots \dot\cup C_k$ by a matrix $H \in \mathbb{R}^{n\times k}$ with

$$H_{il} = \begin{cases} 1/\sqrt{|C_l|}, & i \in C_l, \\ 0, & i \notin C_l \end{cases}, \quad (2)$$

then $\mathrm{RatioCut}(C_1, \dots, C_k) = \mathrm{Tr}(H^T L H)$. Hence, in order to minimize the RatioCut function over all possible clusterings, we could instead solve

$$\min_{H\in\mathbb{R}^{n\times k}} \mathrm{Tr}(H^T L H) \text{ subject to } H \text{ is of form (2).} \quad (3)$$

Spectral clustering relaxes this minimization problem by replacing the requirement that H has to be of form (2) with the weaker requirement that $H^T H = I_k$, that is it solves

$$\min_{H\in\mathbb{R}^{n\times k}} \mathrm{Tr}(H^T L H) \text{ subject to } H^T H = I_k. \quad (4)$$

Since L is symmetric, it is well known that a solution to (4) is given by a matrix H that contains some orthonormal eigenvectors corresponding to the k smallest eigenvalues (respecting multiplicities) of L as columns (Lütkepohl, 1996, Section 5.2.2). Consequently, the first step of SC is to compute such an optimal H by computing the k smallest eigenvalues and corresponding eigenvectors. The second step is to infer a clustering from H. While there is a one-to-one correspondence between a clustering and a matrix of the form (2), this is not the case for a solution H to the relaxed problem (4). Usually, a clustering of V is inferred from H by applying k-means clustering to the rows of H. We summarize unnormalized SC as Algorithm 1. Note that, in general, there is no guarantee on how close the RatioCut value of the clustering obtained by Algorithm 1 to the RatioCut value of an optimal clustering (solving (3)) is.

3. Adding Fairness Constraints

We now extend the above setting to incorporate fairness constraints. Suppose that the data set V contains h groups V_s

such that $V = \dot{\cup}_{s\in[h]} V_s$. Chierichetti et al. (2017) proposed a notion of fairness for clustering asking that every cluster contains approximately the same number of elements from each group V_s. For a clustering $V = C_1 \dot{\cup} \ldots \dot{\cup} C_k$, define the balance of cluster C_l as

$$\text{balance}(C_l) = \min_{s\neq s'\in[h]} \frac{|V_s \cap C_l|}{|V_{s'} \cap C_l|} \in [0,1]. \tag{5}$$

The higher the balance of each cluster, the fairer is the clustering according to the notion of Chierichetti et al. (2017). For any clustering, we have $\min_{l\in[k]} \text{balance}(C_l) \leq \min_{s\neq s'\in[h]} |V_s|/|V_{s'}|$, so that this fairness notion is actually asking for a clustering in which in every cluster, each group is (approximately) represented with the same fraction as in the whole data set V. The following lemma shows how to incorporate this goal into the RatioCut minimization problem (3) using a linear constraint on H.

Lemma 1 (Fairness constraints as linear constraint on H)**.** *For $s \in [h]$, let $f^{(s)} \in \{0,1\}^n$ be the group-membership vector of V_s, that is $f_i^{(s)} = 1$ if $i \in V_s$ and $f_i^{(s)} = 0$ otherwise. Let $V = C_1 \dot{\cup} \ldots \dot{\cup} C_k$ be a clustering that is encoded as in (2). We have, for every $l \in [k]$,*

$$\forall s \in [h-1] : \sum_{i=1}^{n} \left(f_i^{(s)} - \frac{|V_s|}{n} \right) H_{il} = 0 \quad \Leftrightarrow$$
$$\forall s \in [h] : \frac{|V_s \cap C_l|}{|C_l|} = \frac{|V_s|}{n}.$$

Proof. This simply follows from

$$\sum_{i=1}^{n} \left(f_i^{(s)} - \frac{|V_s|}{n} \right) H_{il} = \frac{|V_s \cap C_l|}{\sqrt{|C_l|}} - \frac{|V_s| \cdot |C_l|}{n\sqrt{|C_l|}}.$$

and $|C_l| = \sum_{s=1}^{h} |V_s \cap C_l|$. □

Hence, if we want to find a clustering that minimizes the RatioCut objective function and is as fair as possible, we have to solve

$$\min_{H\in\mathbb{R}^{n\times k}} \text{Tr}(H^T L H) \text{ subject to } H \text{ is of form (2)} \text{ and } F^T H = 0_{(h-1)\times k}, \tag{6}$$

where $F \in \mathbb{R}^{n\times(h-1)}$ is the matrix that has the vectors $f^{(s)} - (|V_s|/n) \cdot \mathbf{1}_n$, $s \in [h-1]$, as columns. In the same way as we have relaxed (3) to (4), we may relax the minimization problem (6) to

$$\min_{H\in\mathbb{R}^{n\times k}} \text{Tr}(H^T L H) \text{ subject to } H^T H = I_k \text{ and } F^T H = 0_{(h-1)\times k}. \tag{7}$$

Our proposed approach to incorporate the fairness notion by Chierichetti et al. (2017) into the SC framework consists of solving (7) instead of (4) (and, as before, applying k-means clustering to the rows of an optimal H in order to infer a clustering). Our approach is analogous to the numerous versions of constrained SC that try to incorporate must-link constraints ("vertices A and B should end up in the same cluster") by putting a linear constraint on H (e.g., Yu & Shi, 2004; Kawale & Boley, 2013; see Section 5).

Algorithm 2 Unnormalized SC with fairness constraints

Input: weighted adjacency matrix $W \in \mathbb{R}^{n\times n}$; $k \in \mathbb{N}$; group-membership vectors $f^{(s)} \in \{0,1\}^n$, $s \in [h]$

Output: a clustering of $[n]$ into k clusters

- compute the Laplacian matrix $L = D - W$
- Let F be a matrix with columns $f^{(s)} - \frac{|V_s|}{n} \cdot \mathbf{1}_n$, $s \in [h-1]$
- compute a matrix Z whose columns form an orthonormal basis of the nullspace of F^T
- compute the k smallest (respecting multiplicities) eigenvalues of $Z^T L Z$ and the corresponding orthonormal eigenvectors (written as columns of Y)
- apply k-means clustering to the rows of $H = ZY$

Next, we describe a straightforward way to solve (7), which is also discussed by Yu & Shi (2004). It is easy to see that $\text{rank}(F) = \text{rank}(F^T) = h-1$. We need to assume that $k \leq n-h+1$ since otherwise (7) does not have any solution. Let $Z \in \mathbb{R}^{n\times(n-h+1)}$ be a matrix whose columns form an orthonormal basis of the nullspace of F^T. We can substitute $H = ZY$ for $Y \in \mathbb{R}^{(n-h+1)\times k}$, and then, using that $Z^T Z = I_{(n-h+1)}$, problem (7) becomes

$$\min_{Y\in\mathbb{R}^{(n-h+1)\times k}} \text{Tr}(Y^T Z^T L Z Y) \text{ subj. to } Y^T Y = I_k. \tag{8}$$

Similarly to problem (4), a solution to (8) is given by a matrix Y that contains some orthonormal eigenvectors corresponding to the k smallest eigenvalues (respecting multiplicities) of $Z^T L Z$ as columns. We then set $H = ZY$.

This way of solving (7) gives rise to our "fair" version of unnormalized SC as stated in Algorithm 2. Note that just as there is no guarantee on the RatioCut value of the output of Algorithm 1 or Algorithm 2 compared to the RatioCut value of an optimal clustering, in general, there is also no guarantee on how fair the output of Algorithm 2 is. We still refer to Algorithm 2 as our *fair* version of unnormalized SC. Similarly to how we proceeded here, in Appendix A, we incorporate the fairness constraints into normalized SC and state our fair version of normalized SC as Algorithm 3.

One might wonder why we do not simply run standard SC on each group V_s separately in order to derive a fair version. In Appendix D we show why such an idea does not work.

Computational complexity We provide a complete dis-

cussion of the complexity of our algorithms in Appendix B. With the implementations as stated, the complexity of both Algorithm 2 and Algorithm 3 is $\mathcal{O}(n^3)$ regarding time and $\mathcal{O}(n^2)$ regarding space, which is the same as the worst-case complexity of standard SC when the number of clusters can be arbitrary. One could apply one of the techniques suggested in the existing literature on constrained spectral clustering to speed up computation (e.g., Yu & Shi, 2004, or Xu et al., 2009; see Section 5), but most of these techniques only work for $k = 2$ clusters.

4. Analysis on Variant of the Stochastic Block Model

In this section, our goal is to model data sets that have two or more meaningful ground-truth clusterings, of which only one is fair, and show that our algorithms recover the fair ground-truth clustering. If there was only one meaningful ground-truth clustering and this clustering was fair, then any clustering algorithm that is able to recover the ground-truth clustering (e.g., standard SC) would be a fair algorithm. To this end, we define a variant of the famous stochastic block model (SBM; Holland et al., 1983). The SBM is a random graph model that has been widely used to study the performance of clustering algorithms, including standard SC (see Section 5 for related work). In the traditional SBM there is a ground-truth clustering of the vertex set $V = [n]$ into k clusters, and in a random graph generated from the model, two vertices i and j are connected with a probability that only depends on which clusters i and j belong to.

In our variant of the SBM we assume that $V = [n]$ comprises h groups $V = V_1 \dot\cup \ldots \dot\cup V_h$ and is partitioned into k ground-truth clusters $V = C_1 \dot\cup \ldots \dot\cup C_k$ such that $|V_s \cap C_l|/|C_l| = \eta_s$, $s \in [h], l \in [k]$, for some $\eta_1, \ldots, \eta_h \in (0,1)$ with $\sum_{s=1}^{h} \eta_s = 1$. Hence, in every cluster each group is represented with the same fraction as in the whole data set V and this ground-truth clustering is fair. Now we define a random graph on V by connecting two vertices i and j with a certain probability $\Pr(i,j)$ that only depends on whether i and j are in the same cluster (or not) and on whether i and j are in the same group (or not). More specifically, we have

$$\Pr(i,j) = \begin{cases} a, & i \text{ and } j \text{ in same cluster and in same group,} \\ b, & i \text{ and } j \text{ not in same cluster, but in same group,} \\ c, & i \text{ and } j \text{ in same cluster, but not in same group,} \\ d, & i \text{ and } j \text{ not in same cluster and not in same group,} \end{cases} \tag{9}$$

and assume that $a > b > c > d$. As in the ordinary SBM, connecting i and j is independent of connecting i' and j' for $\{i,j\} \neq \{i',j'\}$. Every edge is assigned a weight of $+1$, that is no two connected vertices are considered more similar to each other than any two other connected vertices.

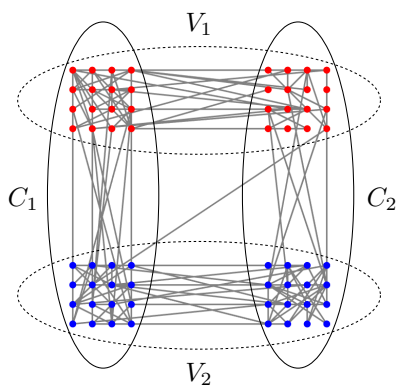

Figure 1. Example of a graph generated from our variant of the SBM. There are two meaningful ground-truth clusterings into two clusters: $V = C_1 \dot\cup C_2$ and $V = V_1 \dot\cup V_2$. Only the first one is fair.

An example of a graph generated from our model (with $h = k = 2$ and $\eta_1 = \eta_2 = 1/2$) can be seen in Figure 1. We can see that there are two meaningful ground-truth clusterings into two clusters: $V = C_1 \dot\cup C_2$ and $V = V_1 \dot\cup V_2$. Among these two clusterings, only $V = C_1 \dot\cup C_2$ is fair since $\mathrm{balance}(C_i) = 1$ while $\mathrm{balance}(V_i) = 0$ for $i \in \{1,2\}$. Note that the clustering $V = V_1 \dot\cup V_2$ has a smaller RatioCut value than $V = C_1 \dot\cup C_2$ because there are more edges between $V_s \cap C_1$ and $V_s \cap C_2$ ($s = 1$ or $s = 2$) than between $V_1 \cap C_l$ and $V_2 \cap C_l$ ($l = 1$ or $l = 2$). As we will see in the experiments in Section 6 (and can also be seen from the proof of the following Theorem 1), for such a graph, standard SC is very likely to return the unfair clustering $V = V_1 \dot\cup V_2$ as output. In contrast, our fair versions of SC return the fair clustering $V = C_1 \dot\cup C_2$ with high probability:

Theorem 1 (SC with fairness constraints succeeds on variant of stochastic block model)**.** *Let $V = [n]$ comprise $h = h(n)$ groups $V = V_1 \dot\cup \ldots \dot\cup V_h$ and be partitioned into $k = k(n)$ ground-truth clusters $V = C_1 \dot\cup \ldots \dot\cup C_k$ such that for all $s \in [h]$ and $l \in [k]$*

$$|V_s| = \frac{n}{h}, \quad |C_l| = \frac{n}{k}, \quad \frac{|V_s \cap C_l|}{|C_l|} = \frac{1}{h}. \tag{10}$$

Let G be a random graph constructed according to our variant of the stochastic block model (9) with probabilities $a = a(n)$, $b = b(n)$, $c = c(n)$, $d = d(n)$ satisfying $a > b > c > d$ and $a \geq C \ln n/n$ for some $C > 0$.

Assume that we run Algorithm 2 or Algorithm 3 (stated in Appendix A) on G, where we apply a $(1+M)$-approximation algorithm to the k-means problem encountered in the last step of Algorithm 2 or Algorithm 3, for some $M > 0$. Then, for every $r > 0$, there exist constants $\widehat{C}_i = \widehat{C}_i(C, r)$ and $\widetilde{C}_i = \widetilde{C}_i(C, r)$, $i \in \{1,2\}$, such that the following is true:

- ***Unnormalized SC with fairness constraints***

  If

$$\frac{a \cdot k^3 \cdot \ln n}{(c-d)^2 \cdot n} < \frac{\widehat{C}_1}{1+M}, \tag{11}$$

  then with probability at least $1 - n^{-r}$, the clustering returned by Algorithm 2 misclassifies at most

$$\widetilde{C}_1 \cdot (1+M) \cdot \frac{a \cdot k^2 \cdot \ln n}{(c-d)^2} \tag{12}$$

  many vertices.

- ***Normalized SC with fairness constraints***

  Let $\lambda_1 = \frac{n}{kh}(a + (h-1)c) + \frac{n(k-1)}{kh}(b + (h-1)d)$. If

$$\frac{\sqrt{k \cdot a \cdot n \ln n}}{\lambda_1 - a} < \frac{\widehat{C}_2}{1+M} \quad \textit{and} \quad \frac{a \cdot k^4 \cdot \ln n}{(c-d)^2 \cdot n} < \frac{\widehat{C}_2}{1+M}, \tag{13}$$

  then with probability at least $1 - n^{-r}$, the clustering returned by Algorithm 3 misclassifies at most

$$\widetilde{C}_2 \cdot (1+M) \cdot \left[\frac{a \cdot k^3 \cdot \ln n}{(c-d)^2} + \frac{a \cdot n^2 \cdot \ln n}{(\lambda_1 - a)^2}\right] \tag{14}$$

  many vertices.

We make several remarks on Theorem 1:

1. By "misclassifies at most x many vertices" we mean that, considering the index l of the cluster C_l that a vertex belongs to as the vertex's class label, there exists a permutation of cluster indices $1, \dots, k$ such that up to this permutation the clustering returned by our algorithm predicts the correct class label for all but x many vertices.

2. The condition (11) is satisfied, for n sufficiently large and assuming that $M \in \mathcal{O}(\ln k)$ (see the next remark), in various regimes: assuming that $k \in \mathcal{O}(n^s)$ for some $s \in [0, 1/3)$, it is satisfied in the dense regime $a, b, c, d \sim \text{const}$, but also in the sparse regime $a, b, c, d \sim \text{const} \cdot (\ln n/n)^q$ for some $q \in [0, 1-3s)$.

The same is true for condition (13), but here we require $s \in [0, 1/4)$ and $q \in [0, 1-4s)$. We suspect that condition (13), with respect to k, is stronger than necessary. We also suspect that the error bound in (14) is not tight with respect to k. Note that in (14), both in the dense and in the sparse regime, the term $a \cdot k^3 \cdot \ln n/(c-d)^2$ is dominating over the term $a \cdot n^2 \cdot \ln n/(\lambda_1 - a)^2$ by the factor k^3.

Both in the dense and in the sparse regime, under these assumptions on s, q and M, the error bounds (12) and (14) divided by n, that is the fraction of misclassified vertices, tends to zero as n goes to infinity. Using the terminology prevalent in the literature on community detection in SBMs (see Section 5), we may say that our algorithms are *weakly consistent* or solve the *almost exact recovery* problem.

3. There are efficient approximation algorithms for the k-means problem in $\mathbb{R}^l$. An algorithm by Ahmadian et al. (2017) achieves a constant approximation factor and has running time polynomial in n, k and l, where n is the number of data points. There is also the famous $(1+\varepsilon)$-approximation algorithm by Kumar et al. (2004) with running time linear in n and l, but exponential in k and $1/\varepsilon$. The algorithm most widely used in practice (e.g., as default method in MATLAB) is k-means++, which is a randomized $\mathcal{O}(\ln k)$-approximation algorithm (Arthur & Vassilvitskii, 2007).

4. We show empirically in Section 6 that our algorithms are also able to find the fair ground-truth clustering in a graph constructed according to our variant of the SBM when (10) is not satisfied, that is when the clusters are of different size or the balance of the fair ground-truth clustering is smaller than 1 (i.e., $\eta_s \neq 1/h$ for some $s \in [h]$). For Algorithm 3, the violation of (10) can be more severe than for Algorithm 2. In general, we observe Algorithm 3 to outperform Algorithm 2. This is in accordance with standard SC, for which normalized SC has been observed to outperform unnormalized SC (von Luxburg, 2007; Sarkar & Bickel, 2015).

The proof of Theorem 1 can be found in Appendix C. It consists of two technical challenges (described here only for the unnormalized case). The first one is to compute the eigenvalues and eigenvectors of the matrix $Z^T \mathcal{L} Z$, where $\mathcal{L}$ is the expected Laplacian matrix of the random graph G and Z is the matrix computed in Algorithm 2. Let $\mathcal{Y}$ be a matrix containing some orthonormal eigenvectors corresponding to the k smallest eigenvalues of $Z^T \mathcal{L} Z$ as columns and Y be a matrix containing orthonormal eigenvectors corresponding to the k smallest eigenvalues of $Z^T L Z$, where L is the observed Laplacian matrix of G. The second challenge is to prove that with high probability, ZY is close to $Z\mathcal{Y}$. For doing so we make use of the famous Davis-Kahan sinΘ Theorem (Davis & Kahan, 1970). After that, we can use existing results about k-means clustering of perturbed eigenvectors (Lei & Rinaldo, 2015) to derive the theorem.

5. Related Work

Spectral clustering and stochastic block model SC is one of the most prominent clustering techniques, with a long history and an abundance of related papers. See von Luxburg (2007) or Nascimento & de Carvalho (2011) for general introductions and an overview of the literature. There are numerous papers on constrained SC, where the goal is to incorporate prior knowledge about the target clustering (usually in the form of must-link and / or cannot-link constraints) into the SC framework (e.g., Yu & Shi, 2001; 2004; Joachims, 2003; Lu & Carreira-Perpinan, 2008; Xu et al., 2009; Wang & Davidson, 2010; Eriksson et al., 2011; Maji et al., 2011; Kawale & Boley, 2013; Khoreva et al., 2014; Wang et al., 2014; Cucuringu et al., 2016). Most of

these papers are motivated by the use of SC in image or video segmentation. Closely related to our work are the papers by Yu & Shi (2004); Xu et al. (2009); Eriksson et al. (2011); Kawale & Boley (2013), which incorporate the prior knowledge by imposing a linear constraint in the RatioCut or NCut optimization problem analogously to how we derived our fair versions of SC. These papers provide efficient algorithms to solve the resulting optimization problems. However, the iterative algorithms by Xu et al. (2009); Eriksson et al. (2011); Kawale & Boley (2013) only work for $k = 2$ clusters. The method by Yu & Shi (2004) works for arbitrary k and could be used to speed up the computation of a solution of (7) or (18) compared to our straightforward way as implemented by Algorithm 2 and Algorithm 3, respectively, but requires to modify the eigensolver in use.

The stochastic block model (SBM; Holland et al., 1983) is the canonical model to study the performance of clustering algorithms. There exist several variants of the original model such as the degree-corrected SBM or the labeled SBM. For a recent survey see Abbe (2018). In the labeled SBM, vertices can carry a label that is correlated with the ground-truth clustering. This is quite the opposite of our model, in which the group-membership information is "orthogonal" to the ground-truth clustering. Several papers show the consistency (i.e., the capability to recover the ground-truth clustering) of different versions of SC on the SBM or the degree-corrected SBM under different assumptions (Rohe et al., 2011; Fishkind et al., 2013; Qin & Rohe, 2013; Lei & Rinaldo, 2015; Joseph & Yu, 2016; Su et al., 2017). For example, Rohe et al. (2011) show consistency of normalized SC assuming that the minimum expected vertex degree is in $\Omega(n/\sqrt{\log n})$, while Lei & Rinaldo (2015) show that SC based on the adjacency matrix is consistent requiring only that the maximum expected degree is in $\Omega(\sqrt{\log n})$. Note that these papers also make assumptions on the eigenvalues of the expected Laplacian or adjacency matrix while all assumptions and guarantees stated in our Theorem 1 directly depend on the connection probabilities a, b, c, d of our model. We are not aware of any work providing consistency results for constrained SC methods as we do in this paper.

Fairness By now, there is a huge body of work on fairness in machine learning. For a recent paper providing an overview of the literature on fair classification see Donini et al. (2018). Our paper adds to the literature on fair methods for unsupervised learning tasks (Chierichetti et al., 2017; Celis et al., 2018a;b; Samadi et al., 2018; Schmidt et al., 2018). Note that all these papers assume to know which demographic group a data point belongs to just as we do. We discuss the pieces of work most closely related to our paper.

Chierichetti et al. (2017) proposed the notion of fairness for clustering underlying our paper. It is based on the fairness notion of disparate impact (Feldman et al., 2015) and the $p\%$-rule (Zafar et al., 2017), respectively, which essentially say that the output of a machine learning algorithm should be independent of a sensitive attribute. In their paper, Chierichetti et al. focus on k-median and k-center clustering. For the case of a binary sensitive attribute, that is there are only two demographic groups, they provide approximation algorithms for the problems of finding a clustering with minimum k-median / k-center cost under the constraint that all clusters have some prespecified level of balance. Subsequently, Rösner & Schmidt (2018) provide an approximation algorithm for such a fair k-center problem with multiple groups. Schmidt et al. (2018) build upon the fairness notion and techniques of Chierichetti et al. and devise an approximation algorithm for the fair k-means problem, assuming that there are only two groups of the same size.

6. Experiments

In this section, we present a number of experiments. We first study our fair versions of spectral clustering, Algorithm 2 and Algorithm 3, on synthetic data generated according to our variant of the SBM and compare our algorithms to standard SC. We also study how robust our algorithms are with respect to a certain perturbation of our model. We then compare our algorithms to standard SC on real network data. We implemented all algorithms in MATLAB[1]. We used the built-in function for k-means clustering with all parameters set to their default values except for the number of replicates, which we set to 10. In the following, all plots show average results obtained from running an experiment for 100 times.

6.1. Synthetic Data

We run experiments on our variant of the SBM introduced in Section 4. To asses the quality of a clustering we measure the fraction of misclassified vertices w.r.t. the fair ground-truth clustering (see Section 4), which we refer to as error.

In the experiments of Figure 2, we study the performance of standard unnormalized and normalized SC and of our fair versions, Algorithm 2 and Algorithm 3, as a function of n. Due to the high running time of Algorithm 3 (see Section 3), we only run it up to $n = 4000$. All plots show the error of the methods, except for the fourth plot in the first row, which shows their running time. We study several parameter settings. For the plots in the first row, Assumption (10) in Theorem 1 is satisfied, that is $|V_s \cap C_l| = \frac{n}{kh}$ for all $s \in [h]$ and $l \in [k]$. In this case, in accordance with Theorem 1, both Algorithm 2 and Algorithm 3 are able to recover the fair ground-truth clustering if n is just large enough while standard SC always fails to do so. Algorithm 3 yields significantly better results than Algorithm 2 and requires much

[1]The code is available on https://github.com/matthklein/fair_spectral_clustering.

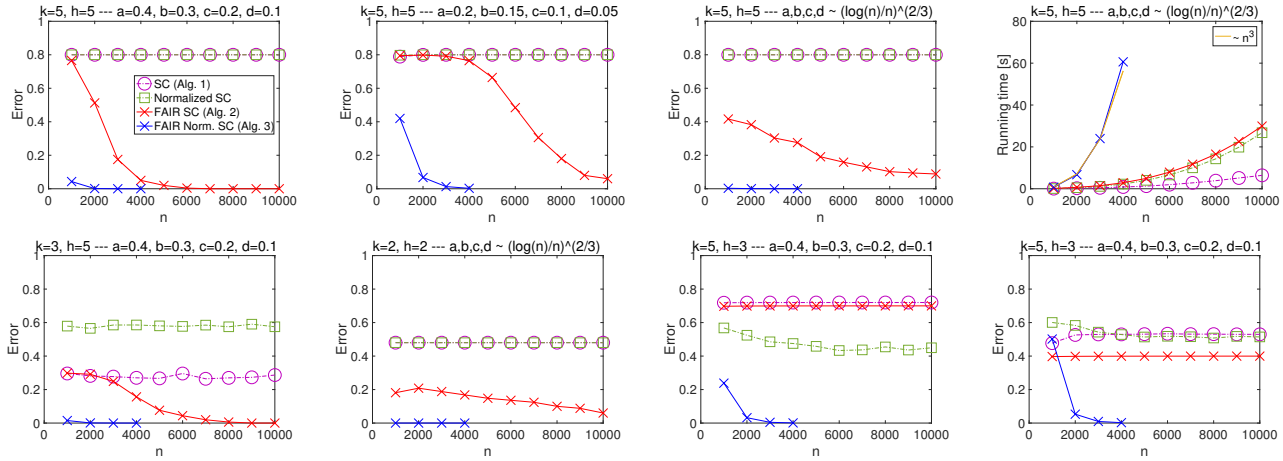

Figure 2. Performance of standard spectral clustering and our fair versions on our variant of the stochstic block model as a function of n for various parameter settings. Error is the fraction of misclassified vertices w.r.t. the fair ground-truth clustering (see Section 4). **First row:** Assumption (10) in Theorem 1 is satisfied, that is $|V_s \cap C_l| = \frac{n}{kh}$, $s \in [h]$, $l \in [k]$. **Second row:** Assumption (10) is not satisfied. **2nd row, 1st plot:** $\frac{|C_1|}{n} = \frac{7}{10}$, $\frac{|C_2|}{n} = \frac{|C_3|}{n} = \frac{15}{100}$ and $\frac{|V_s \cap C_l|}{|C_l|} = \frac{1}{5}$, $s \in [5]$, $l \in [5]$; **2nd plot:** $\frac{|C_1|}{n} = \frac{6}{10}$, $\frac{|C_2|}{n} = \frac{4}{10}$ and $\frac{|V_1 \cap C_l|}{|C_l|} = \frac{6}{10}$, $\frac{|V_2 \cap C_l|}{|C_l|} = \frac{4}{10}$, $l \in [2]$; **3rd plot:** $\frac{|C_1|}{n} = \frac{|C_2|}{n} = \frac{3}{10}$, $\frac{|C_3|}{n} = \frac{2}{10}$, $\frac{|C_4|}{n} = \frac{|C_5|}{n} = \frac{1}{10}$ and $\frac{|V_1 \cap C_l|}{|C_l|} = \frac{5}{10}$, $\frac{|V_2 \cap C_l|}{|C_l|} = \frac{3}{10}$, $\frac{|V_3 \cap C_l|}{|C_l|} = \frac{2}{10}$, $l \in [5]$; **4th plot:** $\frac{|C_1|}{n} = \frac{6}{10}$, $\frac{|C_2|}{n} = \ldots = \frac{|C_5|}{n} = \frac{1}{10}$ and $\frac{|V_1 \cap C_l|}{|C_l|} = \frac{5}{10}$, $\frac{|V_2 \cap C_l|}{|C_l|} = \frac{3}{10}$, $\frac{|V_3 \cap C_l|}{|C_l|} = \frac{2}{10}$, $l \in [5]$.

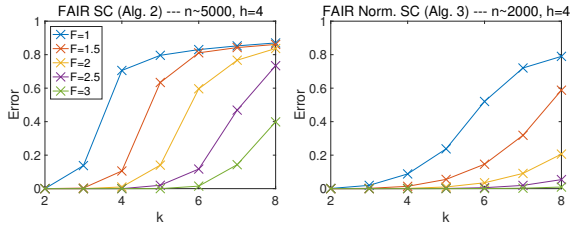

Figure 3. Error of our algorithms as a function of k. We consider $a = F \cdot \frac{25}{100}$, $b = F \cdot \frac{2}{10}$, $c = F \cdot \frac{15}{100}$, $d = F \cdot \frac{1}{10}$ for various values of F. **Left:** Alg. 2, $n \approx 5000$. **Right:** Alg. 3, $n \approx 2000$.

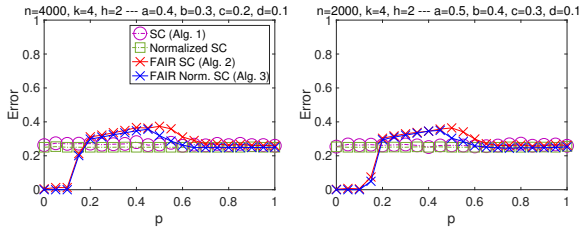

Figure 4. Error of standard spectral clustering and our fair versions as a function of the perturbation parameter p.

smaller values of n for achieving zero error. This comes at the cost of a higher running time of Algorithm 3 (still it is in $\mathcal{O}(n^3)$ as claimed in Section 3). The run-time of Algorithm 2 is the same as the run-time of standard normalized SC. For the plots in the second row, Assumption (10) in Theorem 1 is not satisfied. We consider various scenarios of cluster sizes $|C_l|$ and group sizes $|V_s|$ (however, we always have $|V_s \cap C_l|/|C_l| = |V_s|/n$, $s \in [h]$, $l \in [k]$, so that $C_1 \dot\cup \ldots \dot\cup C_k$ is as fair as possible). When the cluster sizes are different, but the group sizes are all equal to each other (1st plot in the 2nd row) or Assumption (10) is only slightly violated (2nd plot), both Algorithm 2 and Algorithm 3 are still able to recover the fair ground-truth clustering. Compared to the plots in the first row, Algorithm 2 requires a larger value of n though, even though k is smaller. Algorithm 3 achieves (almost) zero error already for $n = 1000$ in these scenarios. When Assumption (10) is strongly violated (3rd and 4th plot), Algorithm 2 fails to recover the fair ground-truth clustering, but Algorithm 3 still succeeds.

In the experiments shown in Figure 3, we study the error of Algorithm 2 (left plot) and Algorithm 3 (right plot) as a function of k when n is roughly fixed. More precisely, for $k \in \{2, \ldots, 8\}$ and $h = 4$, we have $n = kh\lceil \frac{5000}{kh} \rceil$ (Alg. 2; left) or $n = kh\lceil \frac{2000}{kh} \rceil$ (Alg. 3; right), which allows for fair ground-truth clusterings satisfying (10). We consider connection probabilities $a = F \cdot \frac{25}{100}$, $b = F \cdot \frac{2}{10}$, $c = F \cdot \frac{15}{100}$, $d = F \cdot \frac{1}{10}$ for $F \in \{1, 1.5, 2, 2.5, 3\}$. Unsurprisingly, for both Algorithm 2 and Algorithm 3 the error is monotonically increasing with k. The rate of increase critically depends on F (or the probabilities a, b, c, d). For Algorithm 2, this is even more severe. There is only a small range in which the various curves exhibit polynomial growth, which makes it impossible to empirically evaluate whether our error guarantees (12) and (14) are tight with respect to k.

In the experiments of Figure 4, we consider a perturbation of our model as follows: first, for $n = 4000$ (left plot) or

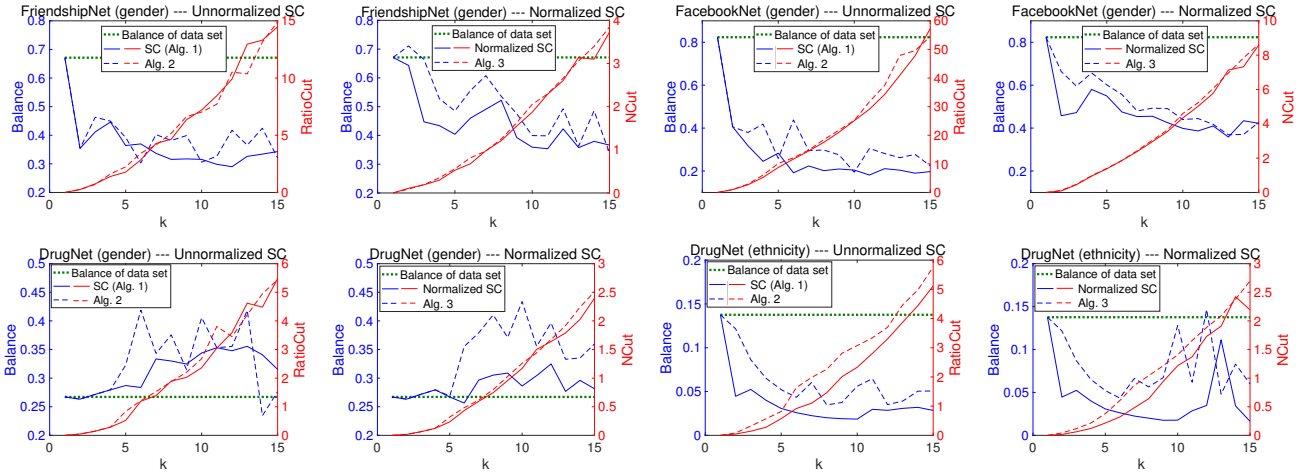

Figure 5. Balance (left axis) and RatioCut / NCut value (right axis) of standard SC and our fair versions as a function of k on real networks.

$n = 2000$ (right plot), $k = 4$ and $h = 2$ we generate a graph from our model just as before (Assumption (10) is satisfied; in particular, the two groups have the same size), but then we assign some of the vertices in the first group to the other group. Concretely, for a perturbation parameter $p \in [0, 1]$, each vertex in the first group is assigned to the second one with probability p independently of each other. The case $p = 0$ is our model without any perturbation. If $p = 1$, there is only one group and our algorithms technically coincide with standard unnormalized or normalized SC. The two plots show the error of our algorithms and standard SC as a function of p. Both our algorithms show the same behavior. They are robust against the perturbation up to $p = 0.15$. They yield the same error as standard SC for $p \geq 0.7$.

6.2. Real Data

In the experiments of Figure 5, we evaluate the performance of standard unnormalized and normalized SC versus our fair versions on real network data. The quality of a clustering is measured through its "Balance" (defined as the average of the balance (5) over all clusters; shown on left axis of the plots) and its RatioCut (1) or NCut (15) value (right axis). All networks that we are working with are the largest connected component of an originally unconnected network.

The first row of Figure 5 shows the results as a function of the number of clusters k for two high school friendship networks (Mastrandrea et al., 2015). Vertices correspond to students and are split into two groups of males and females. FRIENDSHIPNET has 127 vertices and an edge between two students indicates that one of them reported friendship with the other one. FACEBOOKNET consists of 155 vertices and an edge between two students indicates friendship on Facebook. As we can see from the plots, compared to standard SC, our fair versions improve the output clustering's balance (by 10% / 15% / 34% / 10% on average over k) while almost not changing its RatioCut or NCut value.

The second row shows the results for DRUGNET, a network encoding acquaintanceship between drug users in Hartford, CT (Weeks et al., 2002). In the left two plots, the network consists of 185 vertices split into two groups of males and females (we had to remove some vertices for which the gender was not known). In the right two plots, the network has 193 vertices split into three ethnic groups of African Americans, Latinos and others. Again, our fair versions of SC quite significantly improve the balance of the output clustering over standard SC (by 5% / 18% / 86% / 167% on average over k). However, in the right two plots we also observe a moderate increase of the RatioCut or NCut value.

7. Discussion

In this work, we presented an algorithmic approach towards incorporating fairness constraints into the SC framework. We provided a rigorous analysis of our algorithms and proved that they can recover fair ground-truth clusterings in a natural variant of the stochastic block model. Furthermore, we provided strong empirical evidence that often in real data sets, it is possible to achieve higher demographic proportionality at minimal additional cost in the clustering objective.

An important direction for future work is to understand the price of fairness in the SC framework if one needs to satisfy the fairness constraints exactly. One way to achieve this would be to run the fair k-means algorithm of Schmidt et al. (2018) in the last step of our Algorithms 2 or 3. We want to point out that the algorithm of Schmidt et al. currently does not extend beyond two groups of the same size. Second, our experimental results on the stochastic block model provide evidence that our algorithms are robust to moderate levels of perturbations in the group assignments. Characterizing this robustness rigorously is an intriguing open problem.

Acknowledgements

This research is supported by a Rutgers Research Council Grant and a Center for Discrete Mathematics and Theoretical Computer Science (DIMACS) postdoctoral fellowship.

References

Abbe, E. Community detection and stochastic block models: Recent developments. *Journal of Machine Learning Research (JMLR)*, 18:1–86, 2018.

Ahmadian, S., Norouzi-Fard, A., Svensson, O., and Ward, J. Better guarantees for k-means and Euclidean k-median by primal-dual algorithms. In *Symposium on Foundations of Computer Science (FOCS)*, 2017.

Arthur, D. and Vassilvitskii, S. k-means++: The advantages of careful seeding. In *Symposium on Discrete Algorithms (SODA)*, 2007.

Bai, Z., Demmel, J., Dongarra, J., Ruhe, A., and van der Vorst, H. (eds.). *Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide*. Society for Industrial and Applied Mathematics, 2000.

Bhatia, R. *Matrix Analysis*. Springer, 1997.

Boucheron, S., Lugosi, G., and Bousquet, O. Concentration inequalities. In *Advanced Lectures on Machine Learning*. Springer, 2004.

Celis, L. E., Keswani, V., Straszak, D., Deshpande, A., Kathuria, T., and Vishnoi, N. K. Fair and diverse DPP-based data summarization. In *International Conference on Machine Learning (ICML)*, 2018a.

Celis, L. E., Straszak, D., and Vishnoi, N. K. Ranking with fairness constraints. In *International Colloquium on Automata, Languages and Programming (ICALP)*, 2018b.

Chierichetti, F., Kumar, R., Lattanzi, S., and Vassilvitskii, S. Fair clustering through fairlets. In *Neural Information Processing Systems (NIPS)*, 2017.

Cucuringu, M., Koutis, I., Chawla, S., Miller, G., and Peng, R. Simple and scalable constrained clustering: A generalized spectral method. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2016.

Davis, C. and Kahan, W. M. The rotation of eigenvectors by a perturbation. III. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

Donini, M., Oneto, L., Ben-David, S., Shawe-Taylor, J., and Pontil, M. Empirical risk minimization under fairness constraints. In *Neural Information Processing Systems (NeurIPS)*, 2018.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. Fairness through awareness. In *Innovations in Theoretical Computer Science Conference (ITCS)*, 2012.

Eriksson, A., Olsson, C., and Kahl, F. Normalized cuts revisited: A reformulation for segmentation with linear grouping constraints. *Journal of Mathematical Imaging and Vision*, 39(1):45–61, 2011.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. Certifying and removing disparate impact. In *ACM International Conference on Knowledge Discovery and Data Mining (KDD)*, 2015.

Fishkind, D. E., Sussman, D. L., Tang, M., Vogelstein, J. T., and Priebe, C. E. Consistent adjacency-spectral partitioning for the stochastic block model when the model parameters are unknown. *SIAM Journal on Matrix Analysis and Applications*, 34(1):23–39, 2013.

Golub, G. H. and Van Loan, C. F. *Matrix Computations*. John Hopkins University Press, 2013.

Hagen, L. and Kahng, A. B. New spectral methods for ratio cut partitioning and clustering. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 11(9):1074–1085, 1992.

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. In *Neural Information Processing Systems (NIPS)*, 2016.

Holland, P. W., Laskey, K. B., and Leinhardt, S. Stochastic blockmodels: First steps. *Social Networks*, 5:109–137, 1983.

Joachims, T. Transductive learning via spectral graph partitioning. In *International Conference on Machine Learning (ICML)*, 2003.

Joseph, A. and Yu, B. Impact of regularization on spectral clustering. *The Annals of Statistics*, 44(4):1765–1791, 2016.

Kawale, J. and Boley, D. Constrained spectral clustering using L1 regularization. In *SIAM International Conference on Data Mining (SDM)*, 2013.

Khoreva, A., Galasso, F., Hein, M., and Schiele, B. Learning must-link constraints for video segmentation based on spectral clustering. In *German Conference on Pattern Recognition (GCPR)*, 2014.

Kleinberg, J., Mullainathan, S., and Raghavan, M. Inherent trade-offs in the fair determination of risk scores. In *Innovations in Theoretical Computer Science Conference (ITCS)*, 2017.

Kleindessner, M., Awasthi, P., and Morgenstern, J. Fair k-center clustering for data summarization. In *International Conference on Machine Learning (ICML)*, 2019.

Kumar, A., Sabharwal, Y., and Sen, S. A simple linear time $(1+\varepsilon)$-approximation algorithm for k-means clustering in any dimensions. In *Symposium on Foundations of Computer Science (FOCS)*, 2004.

Lei, J. and Rinaldo, A. Consistency of spectral clustering in stochastic block models. *The Annals of Statistics*, 43(1): 215–237, 2015.

Li, M., Lian, X.-C., Kwok, J. T.-Y., and Lu, B.-L. Time and space efficient spectral clustering via column sampling. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

Lu, Z. and Carreira-Perpinan, M. A. Constrained spectral clustering through affinity propagation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

Lütkepohl, H. *Handbook of Matrices*. Wiley & Sons, 1996.

Maji, S., Vishnoi, N. K., and Malik, J. Biased normalized cuts. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

Mastrandrea, R., Fournet, J., and Barrat, A. Contact patterns in a high school: a comparison between data collected using wearable sensors, contact diaries and friendship surveys. *PloS ONE*, 10(9):1–26, 2015. Data available on http://www.sociopatterns.org/datasets/high-school-contact-and-friendship-networks/.

Nascimento, M. C. V. and de Carvalho, A. C. P. L. F. Spectral methods for graph clustering - A survey. *European Journal of Operational Research*, 211(2):221–231, 2011.

Qin, T. and Rohe, K. Regularized spectral clustering under the degree-corrected stochastic blockmodel. In *Neural Information Processing Systems (NIPS)*, 2013.

Rohe, K., Chatterjee, S., and Yu, B. Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, 39(4):1878–1915, 2011.

Rösner, C. and Schmidt, M. Privacy preserving clustering with constraints. In *International Colloquium on Automata, Languages, and Programming (ICALP)*, 2018.

Samadi, S., Tantipongpipat, U., Morgenstern, J., Singh, M., and Vempala, S. The price of fair PCA: One extra dimension. In *Neural Information Processing Systems (NeurIPS)*, 2018.

Sarkar, P. and Bickel, P. J. Role of normalization in spectral clustering for stochastic blockmodels. *The Annals of Statistics*, 43(3):962–990, 2015.

Schmidt, M., Schwiegelshohn, C., and Sohler, C. Fair coresets and streaming algorithms for fair k-means clustering. arXiv:1812.10854 [cs.DS], 2018.

Shi, J. and Malik, J. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

Su, L., Wang, W., and Zhang, Y. Strong consistency of spectral clustering for stochastic block models. arXiv:1710.06191 [stat.ME], 2017.

von Luxburg, U. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

Vu, V. Q. and Lei, J. Minimax sparse principal subspace estimation in high dimensions. *The Annals of Statistics*, 41(6):2905–2947, 2013.

Wang, X. and Davidson, I. Flexible constrained spectral clustering. In *ACM International Conference on Knowledge Discovery and Data Mining (KDD)*, 2010.

Wang, X., Qian, B., and Davidson, I. On constrained spectral clustering and its applications. *Data Mining and Knowledge Discovery*, 28(1):1–30, 2014.

Weeks, M. R., Clair, S., Borgatti, S. P., Radda, K., and Schensul, J. J. Social networks of drug users in high-risk sites: Finding the connections. *AIDS and Behavior*, 6(2):193–206, 2002. Data available on https://sites.google.com/site/ucinetsoftware/datasets/covert-networks/drugnet.

Xu, L., Li, W., and Schuurmans, D. Fast normalized cut with linear constraints. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

Yan, D., Huang, L., and Jordan, M. I. Fast approximate spectral clustering. In *ACM International Conference on Knowledge Discovery and Data Mining (KDD)*, 2009.

Yu, S. X. and Shi, J. Grouping with bias. In *Neural Information Processing Systems (NIPS)*, 2001.

Yu, S. X. and Shi, J. Segmentation given partial grouping constraints. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(2):173–183, 2004.

Zafar, M. B., Valera, I., Rodriguez, M. G., and Gummadi, K. P. Fairness constraints: Mechanisms for fair classification. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017.