
Fair and Diverse DPP-Based Data Summarization

L. Elisa Celis¹ Vijay Keswani¹ Damian Straszak¹ Amit Deshpande² Tarun Kathuria³ Nisheeth K. Vishnoi¹

Abstract

Sampling methods that choose a subset of the data proportional to its *diversity* in the feature space are popular for data summarization. However, recent studies have noted the occurrence of *bias* – e.g., under or over representation of a particular gender or ethnicity – in such data summarization methods. In this paper we initiate a study of the problem of outputting a diverse and *fair* summary of a given dataset. We work with a well-studied determinantal measure of diversity and corresponding distributions (DPPs) and present a framework that allows us to incorporate a general class of fairness constraints into such distributions. Designing efficient algorithms to sample from these constrained determinantal distributions, however, suffers from a complexity barrier; we present a fast sampler that is provably good when the input vectors satisfy a natural property. Our empirical results on both real-world and synthetic datasets show that the diversity of the samples produced by adding fairness constraints is not too far from the unconstrained case.

1. Introduction

A problem facing many services – from search engines and news feeds to machine learning – is data summarization: how can one select a small but representative, i.e., *diverse*, subset from a large dataset. For instance, Google Images outputs a small subset of images from its enormous dataset given a user query. Similarly, in training a learning algorithm one may be required to choose a subset of data points to train on as training on the entire dataset may be costly. However, data summarization algorithms prevalent in the online world have been recently shown to be biased with respect to sensitive attributes such as gender, race or ethnicity. For instance, a recent study found evidence of systematic

under-representation of women in search results (Kay et al., 2015). Concretely, the above work studied the output of Google Images for various search terms involving occupations and found, e.g., that for the search term “CEO”, the percentage of women in top 100 results was 11%, significantly lower than the ground truth of 27%. Through studies on human subjects, they also found that such misrepresentations have the power to influence people’s perception about reality. Beyond humans, since data summaries are used to train algorithms, there is a danger that these biases in the data might be passed on to the algorithms that use them; a phenomena that is being revealed more and more in automated data-driven processes in education, recruitment, banking, and judiciary systems, see (O’Neil, 2016).

A robust and widely deployed method for data summarization is to associate a *diversity score* to each subset and select a subset with probability proportional to this score; see (Hesabi et al., 2015). This paper focuses on a concrete geometric measure of diversity of a subset S of a dataset $\{v_x\}_{x \in X}$ of vectors – the *determinantal measure* denoted by $G(S)$ (Kulesza & Taskar, 2012); and the resulting probability distribution is called a determinantal point process (DPP). $G(S)$ generalizes the correlation measure for two vectors to multiple vectors and, intuitively, the larger $G(S)$, the more diverse is S in the feature space. Among benefits of $G(\cdot)$ are its overall simplicity, wide applicability – not depending on combinatorial properties of the data, and efficient computability. A potential downside might be the additional effort required in modeling, i.e., to represent the data in a suitable vector form so that the geometry of the dataset indeed corresponds to diversity.

Despite the well-acknowledged ability of DPPs to produce diverse subsets, unfortunately, there seems to be no obvious way to ensure that this also guarantees *fairness* in the DPP samples in the form of appropriate representation of sensitive attributes in the subset selected. Partially, this is due to the fact that fairness could mean different things in different contexts. For instance, consider a dataset in which each data point has a gender. One notion of fairness, useful in ensuring that the ground truth does not get distorted, is *proportional representation*: i.e., the distribution of sensitive characteristics in the output set should be identical to that of the input dataset (Kay et al., 2015). Another notion of fairness, argued to be necessary to reverse the effect of

¹École Polytechnique Fédérale de Lausanne (EPFL), Switzerland ²Microsoft Research, India ³UC Berkeley. Correspondence to: L. Elisa Celis <elisa.celis@epfl.ch>, Nisheeth K. Vishnoi <nisheeth.vishnoi@epfl.ch>.

historical biases (Koriyama et al., 2013), could be *equal representation* – the representation of sensitive characteristics should be equal *independent* of the ratio in the input dataset. While these measures of fairness have natural generalizations to the case when the number of sensitive types is more than two, and can be refined in several ways, one thing remains common: they all operate in the combinatorial space of sensitive attributes of the data points. Simple examples (see, e.g., Figure 1 in the Supplementary File) show that, in certain settings, geometric diversity does not imply fairness and vice-versa; however, there seems to be no intrinsic barrier in attaining both.

We initiate a rigorous study of the problem of incorporating fairness with respect to sensitive attributes of data in DPP-based sampling for data summarization. Our contributions are: A framework that can incorporate a wide class of notions of fairness with respect to disjoint sensitive attributes and, conditioned on being fair in the specified sense, outputs subsets where the probability of a set is still proportional to $G(\cdot)$. In particular, we model the problem as sampling from a *partition DPP* – the parts correspond to different sensitive attributes and the goal is to select a specified number of points from each. Unfortunately, the problem of sampling from partition DPPs has been recently shown to be intractable in a strong sense (Celis et al., 2017) and the question of designing fast algorithms for it, at the expense of being approximate, has been open. Our main technical result is a linear time algorithm (see Section 3.1) to sample from partition DPPs that is guaranteed to output samples from close to the DPP distribution under a natural condition on the data (see Definition 4). We prove that random data matrices satisfy this condition in Section 3.3. We run our algorithm on the Adult dataset (Blake & Merz, 1998) and a curated image dataset with various parameter settings and observe a marked improvement in fairness without compromising geometric diversity by much. A theoretical justification of this low *price of fairness* is provided in Section 4; while there have been few works on controlling fairness, ours is the first to give a rigorous, quantitative price of fairness guarantee in any setting. Overall, our work gives a general and rigorous algorithmic solution to the problem of controlling bias in DPP-based sampling algorithms for data summarization while maximizing diversity.

Related Work. There are several data pre-processing approaches to reduce bias in training data. For example, in (Kamiran & Calders, 2012) or (He & Garcia, 2009), bias is removed from training data by over- or under-sampling from the dataset with appropriately defined cardinality constraints on the parts of a partition. The sampling approach used is often either uniform or preferential (according to a problem-dependent ranking). We show that sampling using partition-DPPs has better results in ensuring diversity of the sampled subset than any such sampling method.

DPP-based sampling has been deployed for many data summarization tasks including text and images (Kulesza & Taskar, 2011), videos (Gong et al., 2014), documents (Lin & Bilmes, 2012), recommendation systems (Zhou et al., 2010), and sensors (Krause et al., 2008); and the study of DPPs with additional budget or resource constraints is of importance. While for unconstrained DPPs there are efficient algorithms to sample (Hough et al., 2006), the problem of sampling from constrained DPPs is intractable; see (Celis et al., 2017), where pseudopolynomial time algorithms for partition DPPs are presented. There is also work on approximate MCMC algorithms for sampling from various discrete point processes (see (Rebeschini & Karbasi, 2015; Anari et al., 2016) and the references therein), and algorithms that are efficient for constrained DPPs under certain restrictions on the data matrix and constraints (see (Li et al., 2016) and the references therein). To the best of our knowledge, ours is the first algorithm for constrained DPPs that is near-linear time. Our algorithm is a greedy, approximate algorithm, and can be considered an extension of a similar algorithm for unconstrained DPPs given by (Deshpande & Vempala, 2006). Finally, our work contributes towards an ongoing effort to measure, understand and incorporate fairness (e.g., see (Barocas & Selbst, 2015; Caliskan et al., 2017; Dwork et al., 2012; Zafar et al., 2017)) in fundamental algorithmic problems, including ranking (Celis et al., 2018b), voting (Celis et al., 2018a), and personalization (Celis & Vishnoi, 2017).

2. Our Model

In this section we present the formal notions, model and other theoretical constructs studied in this paper. X will denote the dataset and we let m denote its size. We assume that for each $x \in X$, we are given a (feature) vector $v_x \in \mathbb{R}^n$, where $n \leq m$ is the dimension of the data. Let V denote the $m \times n$ matrix whose rows correspond to the vectors v_x for $x \in X$. For a set $S \subseteq X$, we use V_S to denote the submatrix of V that is obtained by picking the rows of V corresponding to the elements of S . We can now describe geometric diversity formally.

Definition 1. (Geometric Diversity) *Given a dataset X and the corresponding feature vectors $V \in \mathbb{R}^{m \times n}$, the geometric diversity of a subset $S \subseteq X$ is defined as $G(S) := \det(V_S V_S^T)$, which is the squared volume of the parallelepiped spanned by the rows of V_S .*

This volume generalizes the correlation measure for two vectors to multiple vectors and, intuitively, the larger the volume, the more diverse is S in the feature space; see Figure 2 in the Supplementary File for an illustration. Geometric diversity gives rise to the following distribution on subsets known as a determinantal point process (DPP).

Definition 2. (DPPs and k -DPPs) *Given a dataset X and the corresponding feature vectors $V \in \mathbb{R}^{m \times n}$, the DPP is a distribution over subsets $S \subseteq X$ such that the probability*

$\mathbb{P}[S] \propto \det(V_S V_S^\top)$. The induced probability distribution over k -sized subsets is called k -DPP.

A characteristic of a DPP measure is that the inclusion of one item makes including other similar items less likely. Consequently, DPPs assign greater probability to subsets of points that are diverse; for example, a DPP prefers search results that cover multiple aspects of a user’s query, rather than the most popular one.

Our Algorithmic Framework: We are given a dataset X along with corresponding feature vectors $V \in \mathbb{R}^{m \times n}$ and a positive number $k \leq m$ that denotes the size of the subset or summary that needs to be generated. The dataset X is partitioned into p disjoint classes $X_1 \cup X_2 \cup \dots \cup X_p$, each corresponding to a sensitive class. A key feature of our model is that we do not fix one notion of fairness; rather, we allow for the specification of *fairness constraints* with respect to these sensitive classes. Formally, we do this by taking as input p natural numbers (k_1, k_2, \dots, k_p) such that $\sum_{j=1}^p k_j = k$ is the sample size. These numbers give rise to a *fair family of allowed subsets* defined to be $\mathcal{B} := \{S \subseteq X : |S \cap X_j| = k_j \text{ for all } j = 1, 2, \dots, p\}$. By setting (k_1, \dots, k_p) appropriately, the user can ensure their desired notion of fairness. For example, if the dataset has m_i items with the i -th sensitive attribute, then we can set $k_i := km_i/m$ to obtain proportional representation. Similarly, equal representation can be implemented by setting $k_i = k/p$ for all i .

The fair data summarization problem is to sample from a distribution that is supported on \mathcal{B} . However, there could be many such distributions; we pick one that is “closest” to the to the k -DPP described by V . We use the Kullback-Leibler (KL) divergence between distributions q and \tilde{q} defined as $D_{KL}(q||\tilde{q}) := \sum_S q_S \log \frac{q_S}{\tilde{q}_S}$.¹ The following lemma characterizes the distribution supported on \mathcal{B} that has the least KL-divergence to a given distribution (see Appendix B.1 in the Supplementary File for the proof).

Lemma 1. *Given a distribution \tilde{q} with support set \mathcal{C} , let $\mathcal{B} \subseteq \mathcal{C}$ and q be any distribution on \mathcal{B} . Then the optimal value of $\min_q D_{KL}(q||\tilde{q})$ is achieved by the distribution q^* , such that $q_S^* \propto \tilde{q}_S$, for $S \in \mathcal{B}$ and 0 otherwise.*

Thus, the distribution above can be thought of as the most diverse while being fair; we call it partition DPP, or P -DPP.

Definition 3. (P -DPP) *Given a dataset X , the corresponding feature vectors $V \in \mathbb{R}^{m \times n}$, a partition $X = X_1 \cup X_2 \cup \dots \cup X_p$ into p parts, and natural numbers k_1, \dots, k_p , P -DPP defines a distribution q^* over subsets $S \subseteq X$ of size $k = \sum_{i=1}^p k_i$ such that for all $S \in \mathcal{B}$ we*

¹Note that when there are only two parts, one can recover the percentages of elements from each part from the KL-distance. For multiple parts, the KL-distance is a natural (and general) single-dimensional function of the percentage vector with which to measure the deviation from the target distribution.

have $q_S^* := \frac{\det(V_S V_S^\top)}{\sum_{T \in \mathcal{B}} \det(V_T V_T^\top)}$, and $q_S^* = 0$ otherwise.

Given the results of (Celis et al., 2017), we know that sampling from P -DPPs is #P-hard and exact sampling algorithm for P -DPPs are unlikely. Correspondingly, the flexibility that our framework provides in specifying the fairness constraints comes at a computational cost. In this paper, we give a fast, approximate sampling algorithm for P -DPPs.

3. Our Algorithm

Notions of Volume and Projection. Let us recall the interpretation of determinants in terms of volumes. For $S \subseteq X$, V_S is the set of vectors $\{v_x\}_{x \in S}$. If the vectors in S are pairwise orthogonal, then the matrix $V_S V_S^\top$ is diagonal with entries $\{\|v_x\|^2\}_{x \in S}$ on the diagonal and, hence, $\det(V_S V_S^\top) = \prod_{x \in S} \|v_x\|^2$. In the general case, the determinant is not simply the (squared) product of the norms of vectors, however a similar formula still holds. Let $H \subseteq \mathbb{R}^n$ be any linear subspace and H^\perp be its orthogonal complement, i.e., $H^\perp := \{y \in \mathbb{R}^n \mid \langle x, y \rangle = 0 \text{ for all } x \in H\}$. Let $\Pi_H : \mathbb{R}^n \rightarrow \mathbb{R}^n$ be the orthogonal projection operator on the subspace H^\perp , i.e., whenever $w \in \mathbb{R}^n$ decomposes as $w_1 + w_2$ for $w_1 \in H$ and $w_2 \in H^\perp$, then $\Pi_H(w) = w_2$. By a slight abuse of notation, we also denote by Π_v the operator that projects a vector to another that is orthogonal to a given vector $v \in \mathbb{R}^n$, i.e., $\Pi_v(w) := w - \langle w, v \rangle / \|v\|^2$.

The following lemma is a simple generalization of the formula derived above for orthogonal families of vectors and inspires our algorithm for P -DPPs. The proof of this lemma is presented in Section B.3 in the Supplementary File.

Lemma 2 (Determinant Volume Lemma). *Let $w_1, \dots, w_k \in \mathbb{R}^n$ be the rows of a matrix $W \in \mathbb{R}^{k \times n}$, then $\det(WW^\top) = \prod_{i=1}^k \|\Pi_{H_i} w_i\|^2$, where H_i is the subspace spanned by $\{w_1, \dots, w_{i-1}\}$ for all $i = 1, 2, \dots, k$.*

3.1. Our Sample and Project Algorithm

Before we describe our algorithms for sampling from P -DPPs, it is instructive to consider the special case of k -DPPs itself and the simple “orthogonal” scenario – where all the vectors v_x , for $x \in X$, are pairwise orthogonal. In such a case, there is a simple iterative algorithm: sample $x \in X$ with probability $\propto \|v_x\|^2$, then add x to S and remove x from X ; repeat until $|S| = k$. It is intuitively clear, and not hard to prove, that the final probability of obtaining a given set S as a sample is proportional to $\prod_{x \in S} \|v_x\|^2 = \det(V_S V_S^\top)$ and, hence, recovers the k -DPP exactly.

In case of P -DPPs where all the vectors are pairwise orthogonal, and we need to sample k_i vectors from partition X_i , we can sample the required number of elements from each partition independently using the procedure in the previous paragraph. The orthogonality of the vectors and the disjointness of the parts implies that this sampling procedure gives the right probability distribution.

Algorithm 1 Sample-And-Project

```

1: Input:  $V, (X_1, \dots, X_p), (k_1, \dots, k_p)$ 
2:  $S \leftarrow \emptyset$ 
3:  $k \leftarrow k_1 + k_2 + \dots + k_p$ 
4: Let  $w_x := v_x$  for all  $x \in X$ 
5: while  $|S| < k$  do
6:   Pick any2  $i \in \{1, \dots, p\}$  such that  $|S \cap X_i| < k_i$ 
7:   Define  $q \in \mathbb{R}^{X_i}$  by  $q_x := \|w_x\|^2$  for  $x \in X_i$ 
8:   Sample  $\tilde{x} \in X_i$  from distribution  $\left\{ \frac{q_x}{\sum_{y \in X_i} q_y} \right\}_{x \in X_i}$ 
9:    $S \leftarrow S \cup \{\tilde{x}\}$ 
10:  Let  $v := w_{\tilde{x}}$ 
11:  For all  $x \in X$ , set  $w_x := \Pi_v(w_x)$ 
12: end while
    
```

However, when the vectors v_x are no longer pairwise orthogonal, the above heuristic can fail miserably. This is where we invoke Lemma 2. It suggests the following strategy: once we select a vector, then we should orthogonalize all the remaining vectors with respect to it before repeating the sampling procedure. For the case of k -DPPs, it can be shown that this heuristic outputs a set S with probability no more than $k!$ times its desired probability (Deshpande & Vempala, 2006). The $k!$ term is primarily because the k vectors can be chosen in any of the $k!$ orders. Taking this simple heuristic as a starting point and incorporating an additional idea to deal with partition constraints, we arrive at our *Sample and Project* algorithm – see Algorithm 1.

Given that we have made several simplifications and informal “jumps” when deriving the algorithm one cannot expect that the distribution over sets S produced by Algorithm 1 to be exactly the same as P -DPP. Later in this section we give evidence that in fact the distribution output by the “Sample and Project” heuristic can be formally related to the P -DPP distribution, and hence the constructed algorithm is provably an approximation to a P -DPP. However, we first note an attractive feature of this algorithm – it is fast and practical. For a $V \in \mathbb{R}^{m \times n}$ matrix and $k = \sum_{i=1}^p k_i$, Algorithm 1 can be implemented in $O(mnk)$ time.

Note that the size of the data for this problem is already $\Theta(mn)$, hence, the algorithm does only linear work per sampled point. For P -DPPs there is only one known exact algorithm which samples in time $m^{O(p)}$, which is polynomial only when $p = O(1)$ (Celis et al., 2017).

Another possible approach for sampling from DPPs is the Markov Chain Monte Carlo method. It was proved in (Anari et al., 2016) that Markov Chains can be used to sample from k -DPPs in time roughly $\tilde{O}(mk^4 + mn^2)$ given a “warm start”, i.e., a set S_0 of significant probability. This approach does not extend to P -DPPs – indeed in (Anari et al., 2016) the underlying probability distribution is required to be Strongly Rayleigh, a property which holds for k -DPPs, but fails for P -DPPs whenever the number of parts is at least

two. One can still formulate an analogous MCMC algorithm for the case of P -DPPs – it fails on specially crafted “bad instances” but seems to perform well on real world data. However, even ignoring the lack of provable guarantees for this algorithm, it does not seem possible to reduce its running time below $O(mk^4 + mn^2)$, which significantly limits its practical applicability.

3.2. Provable Guarantees for Our Algorithm

We now present a theorem which connects the output distribution of Algorithm 1 to the corresponding P -DPP. To establish such a guarantee we require the following assumption on the singular values of the matrices V_{X_i} .

Definition 4 (β -balance). *Let X be a set of m elements partitioned into p parts X_1, \dots, X_p and let $V \in \mathbb{R}^{m \times n}$ be a matrix. Denote by $\sigma_1 \geq \dots \geq \sigma_n$ the singular values of V and for each $i \in \{1, 2, \dots, p\}$, let $\sigma_{i,1} \geq \dots \geq \sigma_{i,n}$ denote the singular values of V_{X_i} . For $\beta \geq 1$, the partition X_1, \dots, X_p is called β -balanced with respect to V if for all $i \in \{1, \dots, p\}$ and for all $j \in \{1, \dots, n\}$, $\sigma_{i,j} \geq \frac{1}{\beta} \sigma_j$.*

The β -balance property informally requires that the diversity within each of the partitions V_{X_i} , relative to V , is significant. A more concrete geometric way to think about this condition is as follows: if one thinks of the positive semidefinite matrix $V^T V \in \mathbb{R}^{n \times n}$ as representing an ellipsoid in \mathbb{R}^n whose axes are the singular values, then the β -balance condition essentially says that the ellipsoids corresponding to each of the partitions are a β -approximation to that of V (see Figure 4 in the Supplementary File).

Importantly, Algorithm 1 never outputs a set $S \notin \mathcal{B}$, hence the only way its output distribution could significantly differ from the P -DPP would be if certain sets $S \in \mathcal{B}$ appeared in the output with larger probabilities than specified by the P -DPP. Our main theoretical result for *Sample and Project* is that for β -balanced instances we can control the scale at which such a violation can happen.

Theorem 1 (Approximation Guarantee). *Let X be a set of m elements partitioned into p parts X_1, \dots, X_p , a matrix $V \in \mathbb{R}^{m \times n}$ and integers k_1, \dots, k_p , such that X_1, \dots, X_p is a β -balanced partition with respect to V and $\sum_{j=1}^p k_j$. Let $\mathcal{B} \subseteq 2^X$ denote the following family of sets*

$$\mathcal{B} := \{S \subseteq X : |S \cap X_j| = k_j \text{ for all } j = 1, 2, \dots, p\}$$

Then Algorithm 1, with $V, (X_1, \dots, X_p)$ and (k_1, \dots, k_p) as input, returns a subset $S \in \mathcal{B}$ with probability $\tilde{q}(S) \leq \eta_k \cdot \beta^{2k} \cdot q_S^$ where $q_S^* = \frac{\det(V_S V_S^T)}{\sum_{T \in \mathcal{B}} \det(V_T V_T^T)}$, $k = \sum_{j=1}^p k_j$ and $\eta_k = k_1! \cdot k_2! \cdot \dots \cdot k_p!$.*

The proof of the approximation guarantee uses techniques inspired by (Deshpande & Vempala, 2006) who prove a similar bound for k -DPP sampling.

We use the following lemmas in the proof of the theorem. The proof of these lemmas appear in Appendix B.4 and Appendix B.5 in the Supplementary File.

Lemma 3. For any matrix $V \in \mathbb{R}^{m \times n}$ with $m \geq n \geq k$,

$$\sum_{i_1 < i_2 < \dots < i_k} \sigma_{i_1}^2 \sigma_{i_2}^2 \dots \sigma_{i_k}^2 = \sum_{S: |S|=k} \det(V_S V_S^\top)$$

where $\sigma_1, \sigma_2, \dots, \sigma_n$ are the singular values of V and V_S is the sub-matrix of V with rows corresponding to S .

Lemma 4. Given a β -balanced partition, Algorithm 1 returns a set S such that $\det(V_S V_S^\top)$ is non-zero with probability one.

Proof of Theorem 1. Let π be the random variable representing the ordered output of the algorithm. Suppose that the algorithm outputs the set $S = \{x_1, \dots, x_k\}$. Since the partition X_1, \dots, X_p is β -balanced with respect to V , by Lemma 4 the algorithm will always output a set which has non-zero determinant value, i.e., $\det(V_S V_S^\top) \neq 0$.

Consider any ordering of the set S , say, $\tau := (x_1, \dots, x_k)$. Let $H_j \subseteq \mathbb{R}^n$ denote the linear subspace spanned by the vectors corresponding to the first $j - 1$ elements, i.e., $\{v_{x_1}, \dots, v_{x_{j-1}}\}$. We also define a mapping $f : X \rightarrow \{1, \dots, p\}$ such that $f(x) = i$ if $x \in X_i$.

In the first iteration say we choose partition X_1 . Then the algorithm will sample an element from X_1 with probability proportional to the squared norm of the vector.

After $(j - 1)$ iterations w_x will be the orthogonal projection of v_x onto the subspace orthogonal to $\text{span}\{v_{x_1}, v_{x_2}, \dots, v_{x_{j-1}}\}$. This is a consequence of the fact that $(\prod_{v_{x_1}} \prod_{v_{x_2}} \dots \prod_{v_{x_{j-1}}}) = \Pi_{H_j}$.

Hence in the $(j - 1)$ -th iteration, $w_x = \Pi_{H_j}(v_x)$ for all $x \in X$. Therefore, the probability that the sequence τ is the output of the algorithm is

$$\mathbb{P}(\pi = \tau) = \prod_{j=1}^k \frac{\|\Pi_{H_j}(v_{x_j})\|^2}{\sum_{x \in X_{f(x_j)}} \|\Pi_{H_j}(v_x)\|^2}. \quad (1)$$

The numerator of (1) is $\det(V_S V_S^\top)$ by Lemma 2. Let D_{x_1, \dots, x_k} denote the denominator. For each term in the denominator $\sum_{x \in X_l} \|\Pi_{H_j}(v_x)\|^2 = \|V_{X_l} - V'_{X_l}\|_F^2$ where $\|\cdot\|_F$ denotes the Frobenius norm and V'_{X_l} is the rank $j - 1$ matrix with rows $\{v'_x\}_{x \in X_l}$ such that v'_x is the projection of vector v_x on H_j . By a result on low rank approximations (see Theorem 1), we can bound the above quantity as

$$\sum_{x \in X_l} \|\Pi_{H_j}(v_x)\|^2 \geq \sum_{t=j}^n \sigma_{l,t}^2 \geq \frac{1}{\beta^2} \sum_{t=j}^n \sigma_t^2$$

where $\sigma_{l,t}$ is the t -th singular value of V_{X_l} and second inequality is due to the β -balanced property of the partition.

Using above, the denominator of (1) becomes

$$D_{x_1, \dots, x_k} \geq \prod_{j=1}^k \frac{1}{\beta^2} \sum_{t=j}^n \sigma_t^2 \geq \frac{1}{\beta^{2k}} \sum_{t_1 < \dots < t_k} \sigma_{t_1}^2 \dots \sigma_{t_k}^2.$$

By applying Lemma 3, it then follows

$$D_{x_1, \dots, x_k} \geq \frac{1}{\beta^{2k}} \sum_{|S|=k} \det(V_S V_S^\top) \geq \frac{1}{\beta^{2k}} \sum_{S \in \mathcal{B}} \det(V_S V_S^\top).$$

Thus, $\mathbb{P}(\pi = \tau) \leq \beta^{2k} \frac{\det(V_S V_S^\top)}{\sum_{T \in \mathcal{B}} \det(V_T V_T^\top)}$. Since the order in which the partitions are considered by the algorithm is fixed, the vectors of each X_i in τ can be permuted amongst themselves and the output set will still be S . Correspondingly there are $\eta_k = k_1! \cdot k_2! \cdot \dots \cdot k_p!$ valid permutations of τ . Let T_S be the set of all valid permutations of elements of S , then $\bar{q}_S = \sum_{\tau \in T_S} \mathbb{P}(\pi = \tau) \leq \eta_k \cdot \beta^{2k} \cdot q_S^*$. ■

3.3. β -balanced property for random data

For a given matrix $V \in \mathbb{R}^{m \times n}$, suppose we choose the partitions randomly. For each element $x \in X$, we put x in X_i with probability $1/p$. Using the Matrix Chernoff bounds (Tropp, 2012), we prove the following theorem.

Theorem 2. Assume that all the rows v_j (for $j \in X = \{1, 2, \dots, m\}$) of $V \in \mathbb{R}^{m \times n}$ satisfy $v_j^\top (V^\top V)^{-1} v_j \leq \frac{\delta^2}{8p \log(np)}$, where $\delta \in (0, 1)$ is a constant. If X is randomly partitioned into $X = X_1 \cup X_2 \cup \dots \cup X_p$ then with probability at least $\frac{1}{e}$, the partition X_1, \dots, X_p is β -balanced with respect to V , for $\beta = \sqrt{(1 + \delta)p}$.

The proof of this theorem is given in Appendix B.6 in the Supplementary File. The quantity $v_j^\top (V^\top V)^{-1} v_j$ is also called the *statistical leverage score* of v_j with respect to $V^\top V$. For two partitions, the theorem states that if the leverage score of all rows is $O(\frac{1}{\log n})$, then the partitions are β -balanced for $\beta \approx \sqrt{2}$.

4. Price of Fairness

In this section we present conditions under which the k -DPP and P -DPP distributions are close to each other. Note that the support of a P -DPP is a subset of the support of the corresponding k -DPP. Thus, a natural definition of the *price of fairness* is the KL-divergence between them.

Definition 5 (Price of Fairness). Given a matrix $V \in \mathbb{R}^{m \times n}$, partitions X_1, \dots, X_p and integers k_1, \dots, k_p , let $k = k_1 + \dots + k_p$. Suppose q is the distribution defined by k -DPP over subsets of size k and q^* is the distribution defined by P -DPP over subsets with k_i elements from each X_i . Then, the price of fairness is $D_{KL}(q^* || q)$.

We define the following property for the input data and analyze its price of fairness.

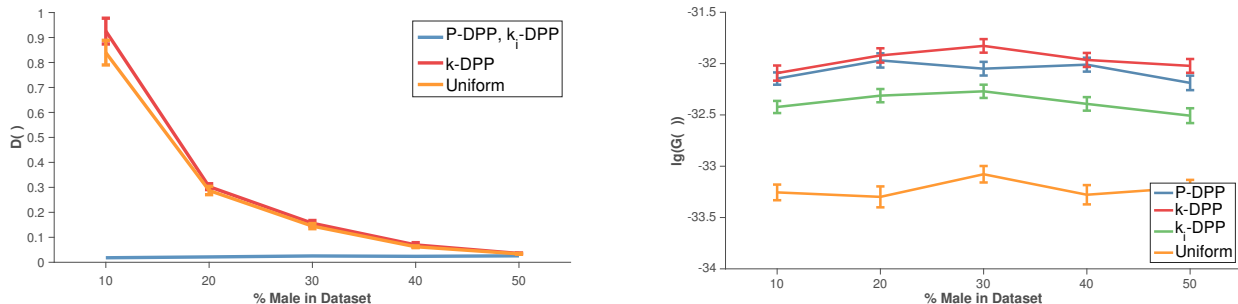


Figure 1. The mean relative unfairness measure $D(\cdot) = D^{\text{un}}(\cdot)$ with respect to the uniform distribution over 4 classes, and the logarithm of the geometric diversity $\lg(G(\cdot))$ are reported in the left and right figures respectively for $n = 200$ repetitions. Error bars represent the standard error of the mean.

Definition 6 (δ -drop). For $0 \leq \delta \leq 1$, the partition X_1, \dots, X_p is called a δ -drop partition with respect to V and k_1, \dots, k_p if for all $i \in \{1, \dots, p\}$, $\sigma_{i, k_i+1} \leq \delta \sigma_{i, k_i}$. Here $\sigma_{i, j}$ is the j -th largest singular value of V_{X_i} .

Roughly, this says that, if δ is small, then each of the matrices V_{X_i} is effectively a rank- k_i matrix. Such a notion of low effective rank appears frequently in the machine learning literature (Roy & Vetterli, 2007; Drineas et al., 1999). We prove the following theorem that asserts that if the δ -drop condition is satisfied, then we can be sure that most of the probability mass is concentrated on subsets which satisfy partition constraints. In such a case, sampling a k sized subset using any k -DPP algorithm will output a subset which satisfies partition constraints with high probability. The proof of the theorem is provided in the Appendix B.7 in the Supplementary File.

Theorem 3. Let $\varepsilon \in (0, 1)$ and suppose that the partition X_1, \dots, X_p is δ -drop w.r.t. V and k_1, \dots, k_p , with $\delta \leq \frac{\varepsilon}{nN_0}$ and $N_0 := \binom{k+p-1}{p-1}$. If $n \geq \sqrt{2}k \cdot \left(\frac{\gamma}{\sigma_n}\right)^2$ (with $\gamma := \max\{\sigma_{i,1}\}_i$, where $\sigma_{i,1}$ is the largest singular value of V_{X_i} and σ_n is the smallest non-zero singular value of V) then the price of ensuring fairness is $D_{KL}(q^*||q) \leq \log \frac{1}{(1-\varepsilon)}$.

5. Empirical Results

5.1. Algorithms and Baselines

In each simulation, we compare several different probability distributions from which to select k samples from a dataset: As benchmarks we consider the (unconstrained) distributions, k -DPP (see Def 2), and UNIF, which selects a uniformly random subset of size k from the dataset X . We compare this against different methods which select from a fair family of allowed subsets, P -DPP (see Def 3), and k_i -DPP (see Def 7 below).

Definition 7. (k_i -DPP) Given a dataset X , the corresponding feature vectors $V \in \mathbb{R}^{m \times n}$, a partition $X = X_1 \cup \dots \cup X_p$ into p parts, and numbers k_1, \dots, k_p , k_i -DPP defines a distribution over $k_1 + \dots + k_p$ -sized subsets $S \subseteq X$ that is a product distribution: for each i , we obtain

a sample $S_i \subseteq X_i$ of size k_i independently with probability proportional to $\mathbb{P}[S_i] \propto \det(V_{S_i} V_{S_i}^T)$, and combine these samples to output $S = S_1 \cup \dots \cup S_p$.

Algorithms for k_i -DPPs are simply obtained by independently using a k -DPP sampler with $k = k_i$ on each part X_i . For sampling from all the above listed distribution we use the Sample and Project algorithm as described in Section 3.1.

Metrics. In each simulation, we report the geometric diversity $G(\cdot)$ (see Def 1) and the fairness as measured by the KL-divergence from the desired frequency over parts. Formally, given a probability distribution q over the p parts of the dataset, we define the relative unfairness measure of a set $S \subseteq X$ as $D^q(S) := D_{KL}(q||s)$, where $s = (s_1, \dots, s_p)$ denotes the vector of frequencies, i.e., $s_i = \frac{|X_i \cap S|}{|S|}$ for $i = 1, 2, \dots, p$. In particular, typically we want to have $D^q(\cdot)$ as small as possible – ideally equal to 0. When $q_i = 1/p$ for all i , we refer to D^q as D^{un} . When $q_i = |X_i|/m$, we refer to D^q as D^{prop} .

5.2. Empirical Results on the Image Dataset

Curated Dataset. We gathered a collection of images curated using Google image search as follows: Four search terms were used: (a) “Scientist Male”, (b) “Scientist Female”, (c) “Painter Male”, and (d) “Painter Female” (Image-dataset).

Following (Kulesza & Taskar, 2011), each image was processed with the `vlf` toolbox to obtain sets of 128-dimensional SIFT descriptors (Lowe, 1999; Vedaldi & Fulkerson, 2008). All such descriptors are collected in a single set and subsampled to roughly 10% of its total size. The resulting set of $\approx 10^4$ descriptors was clustered using the k -means algorithm where $k = 128$ is the number of means. The feature vector for an image is the normalized histogram of the nearest clusters to the descriptors in the image.

Empirical Results on the Biased Datasets. Our goal is to understand how the bias in the underlying dataset can affect the performance of the different sampling distributions with

Sampling Method		Gender						Ethnicity					
		$D^{\text{un}}(\cdot)$		$D^{\text{prop}}(\cdot)$		$\log G(\cdot)$		$D^{\text{un}}(\cdot)$		$D^{\text{prop}}(\cdot)$		$\log G(\cdot)$	
		mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
Unconstr.	UNIF	0.075	0.019	0.001	0.002	-67	41	0.357	0.050	0.001	0.001	-67	41
	k -DPP	0.027	0.009	0.011	0.005	489	11	0.268	0.038	0.005	0.004	487	12
Equal	k_i -UNIF	0	0	0.069	0	-31	35	0	0	0.282	0	16	32
	k_i -DPP	0	0	0.069	0	410	16	0	0	0.282	0	366	16
	P -DPP	0	0	0.069	0	490	11	0	0	0.282	0	476	12
Prop.	k_i -UNIF	0.074	0	0	0	-64	29	0.358	0	0	0	-65	35
	k_i -DPP	0.074	0	0	0	409	17	0.358	0	0	0	426	15
	P -DPP	0.074	0	0	0	482	13	0.358	0	0	0	488	12

Table 1. We report the unfairness ($D^{\text{un}}(\cdot)$) with respect to the uniform distribution over parts, and $D^{\text{prop}}(\cdot)$ with respect to the “proportional” distribution, i.e. as in the whole dataset) and diversity ($\log G(\cdot)$) for the different sampling methods on the Adult dataset when (a) the sensitive attribute is Gender or (b) the sensitive attribute is Ethnicity. Sets of size 400 were selected, and 100 samples were taken for each. For the samplers that match fairness constraints, we consider both selecting subsets with equal representation and selecting subsets with proportional representation. We note that P -DPP has the highest diversity out of all constrained sampling methods regardless of the method of representation. Moreover, the diversity of P -DPP matches that of the unconstrained k -DPP for Gender under proportional representation and for Ethnicity under equal representation.

respect to fairness and geometric diversity. We include all female (b and d) images, but vary how many of the male images (a and c) appear in the dataset in order to create biased sets that have between 10% to 50% male images. The male images are selected uniformly at random from the set of all male scientists and male artists for each repetition of the simulation. We sample 40 images from each biased dataset; roughly the number that fits on the first page of an image search result. We conduct 200 repetitions. We place fairness constraints so that P -DPP and k_i -DPP select exactly 50% of their samples from the male (a and c) images and female (b and d) images, regardless of the bias in the underlying dataset. Note that we *do not* enforce constraints across scientist (a and b) images and artist (c and d) images, but measure the unfairness $D^{\text{un}}(\cdot)$ with respect to all four attributes.

Results. With respect to $D^{\text{un}}(\cdot)$, P -DPP significantly outperforms k -DPP, and UNIF (paired one-sided t -tests, $p < 0.05$), see Figure 1. As expected, the bias in the underlying dataset can dramatically affect the fairness of UNIF and k -DPP as neither approach is designed to correct for such biases. However, P -DPP and k_i -DPP both enforce fairness constraints; note that this is despite the fact that the sampling was only equal with respect to gender and not profession. The latter does not appear to affect the outcome here.

With respect to the diversity $G(\cdot)$, P -DPP has significantly higher $G(\cdot)$ than UNIF and k_i -DPP (paired one-sided t -tests, $p < 0.05$). Moreover, P -DPP performs comparatively to k -DPP; the mean diversity of k -DPP is higher, but not significantly so. Thus, we observe that, when the underlying data is biased, there is a tradeoff between $D^{\text{un}}(\cdot)$ (for which P -DPP performs best) and $G(\cdot)$ (for which k -DPP performs best); however the differences in geometric diversity are negligible while differences in unfairness can be very large.

5.3. Empirical Results on Real-World Dataset

The Adult Dataset. The Adult income dataset (Blake & Merz, 1998) consists of roughly 45000 records of subjects each with 14 features such as age, ethnicity, education and a binary label indicating whether a subject’s incomes is above or below 50K USD.³ This dataset has been widely studied in the context of fairness (see, (Yang & Stoyanovich, 2017; Zafar et al., 2017; Zemel et al., 2013; Zadrozny, 2004)).

In preprocessing the data we filter out incomplete entries, and from the remaining ones we pick a random subset of 5000 records for our simulations. We vectorize the data as follows: Categorical fields (with a small number of possible values) we turn into sets of binary fields. As the dimension n of such feature vectors is quite small – 50 – the DPP framework allows sampling sets of cardinality at most $k \leq 50$. For this reason we enrich the feature vectors in a standard way – by adding pairwise products of all existing features as separate ones – this, after removing redundant columns, yields feature vectors of dimension 992.

Empirical Results on Equal and Proportional Representation. We conduct our simulations across either gender or ethnicity as the sensitive attribute. For the former, we use the gender categories provided in the dataset; all entries were labeled either male (68.3%) or female (31.7%). For the latter, we use the ethnicity categories provided in the dataset; we consider the partition Caucasian (85.7%) and non-Caucasian (14.3%).

In addition to the algorithms mentioned above, we report the performance of an additional benchmark k_i -UNIF, which selects a uniformly random subset of size k_i from X_i . In our subsampling, we consider both equal representation, where

³Data downloaded from <https://archive.ics.uci.edu/ml/datasets/adult>.

Sampling Method		Before Scaling				After scaling			
		$D^{\text{un}}(\cdot)$		$\log G(\cdot)$		$D^{\text{un}}(\cdot)$		$\log G(\cdot)$	
		mean	std	mean	std	mean	std	mean	std
Unconstrained	UNIF	0.066	0	455.7	1.4	0.064	0	228.6	215.8
	k -DPP	0.063	0	457.3	1.3	5.2×10^{-6}	0	397.4	11.6
	Scale-And-Sample	5.2×10^{-6}	0	457.5	1.1	-	-	-	-
Constrained	k_i -UNIF	0	0	455.7	1.3	0	0	226.5	20.8
	P -DPP	0	0	457.2	1.1	0	0	397.5	9.2

Table 2. We report the unfairness ($D^{\text{un}}(\cdot)$ with respect to the uniform distribution over parts) and diversity for the different sampling methods on a random dataset before and after scaling the singular values by a factor of $1/n$. In this simulation we have $m = 200$ vectors of dimension $n = 150$ divided into two partitions (partition 1 has $\frac{m}{3}$ elements and partition 2 has $\frac{2m}{3}$ elements), and we want to sample 50 elements from each partition ($k = 100$).

each attribute makes up of 50% of the selected points, and proportional representation, where each attribute is represented with the same ratio as in the original population.

Results. We observe that P -DPP has the highest diversity out of all constrained sampling methods regardless of the proportion of representation or sensitive attribute; see Table 1. Surprisingly, the diversity of P -DPP matches that of the unconstrained k -DPP for Gender under proportional representation and for Ethnicity under equal representation. In the other two settings – Gender under equal representation and Ethnicity under proportional representation – the P -DPP score is lower than that of k -DPP, but minimally so, and outperforms k_i -DPP by several standard deviations.

We note that k_i -UNIF, although it has very poor geometric diversity as a whole, performs better under equal representation than it does under proportional representation. This fact suggests that there could be value in selecting sensitive attributes equally beyond the consideration of fairness.

The fact that P -DPP performs so well, especially when significantly changing the distribution of sensitive attributes (e.g., for ethnicity, from 14.3% non-Caucasian to 50% non-Caucasian), is quite surprising. Overall, it appears that one can support very dramatic changes to the underlying distributions of attributes with minimal or even zero loss to geometric diversity by using our P -DPP algorithm.

5.4. Empirical Results on the Price of Fairness

We look at the effect of the scaling of singular values, suggested by Theorem 3, on the sampled subsets of our Algorithm. In this simulation we take an instance of random vectors and use different sampling methods to sample a subset from the dataset, and report the $D^{\text{un}}(\cdot)$ and $\log G(\cdot)$ value of the sampled subset. Following this, we scale the tail singular values of the partition matrices by $\delta = O(1/n)$ and again report the $D^{\text{un}}(\cdot)$ and $\log G(\cdot)$ values.

We also present a heuristic approach, Scale-And-Sample, for constrained sampling which will use any k -DPP algorithm as a sub-routine. The algorithm is simple. For each V_{X_i} , scale the smallest $(n - k_i)$ singular values by $1/n$. Then sample a $\sum_{i=1}^p k_i$ sized subset using any k -DPP algorithm.

Results. The results are presented in Table 2. It can be seen that after scaling the tail singular values of the partition matrices, the mean $D^{\text{un}}(\cdot)$ value for k -DPP is very low, and resembles closely the constrained sampling case. We also note that the Scale-And-Sample approach to constrained sampling suggested earlier performs very well. The mean relative unfairness measure $D^{\text{un}}(\cdot)$ is almost zero. Furthermore, the value of the geometric diversity parameter $\log G(\cdot)$ is also similar to unscaled P -DPP.

6. Conclusion and Future Work

In this paper we initiated the study of fair and diverse DPP-based sampling for data summarization. We provide a novel and fast algorithm that can sample from a DPP that satisfy fairness constraints based on the desired proportion of samples with a given attribute. Our algorithm gives provably good guarantees when the data matrix satisfies a natural β -balance property. We prove that a large class of datasets satisfy the β -balance condition. We define a notion of *price of fairness*, the KL-divergence between the fairness constrained distribution and the unconstrained distribution and theoretically show that, when the data satisfies reasonable properties, this price would be low. We further show in silico that adding fairness constraints results in minimal loss to diversity, even when the underlying dataset is very biased, or when the proportion of attributes is changed significantly.

Several challenging problems remain from a technical standpoint; naturally, a first question would be whether the theorems can be improved either by attaining better approximation guarantees, or by weakening the necessary conditions. Extending these results to arbitrary group structures (as opposed to partitions) would be very relevant, but appears to be significantly more challenging.

From a practical point of view, it remains to be seen what effect de-biasing a sampler has on the end result of an ML algorithm (e.g., classification), both on its accuracy and on the output bias. Indeed, this P -DPP model can be used to pre-process the training data by taking a fair subsample; evaluating the performance of ML algorithms in this regard would be an interesting direction for future research.

References

- Anari, N., Oveis Gharan, S., and Rezaei, A. Monte carlo markov chain algorithms for sampling strongly rayleigh distributions and determinantal point processes. In *Proceedings of the 29th Conference on Learning Theory, COLT 2016*, pp. 103–115, 2016.
- Barocas, S. and Selbst, A. D. *Big Data’s Disparate Impact*. SSRN eLibrary, 2015. URL <https://books.google.co.in/books?id=oMNsrgEACAAJ>.
- Blake, C. and Merz, C. UCI machine learning repository, 1998. URL <http://archive.ics.uci.edu/ml>.
- Caliskan, A., Bryson, J. J., and Narayanan, A. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017. ISSN 0036-8075.
- Celis, L. E. and Vishnoi, N. K. Fair personalization. *Fairness, Accountability, and Transparency (FAT)*, 2017.
- Celis, L. E., Deshpande, A., Kathuria, T., Straszak, D., and Vishnoi, N. K. On the complexity of constrained determinantal point processes. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, APPROX/RANDOM 2017*, pp. 36:1–36:22, 2017.
- Celis, L. E., Huang, L., and Vishnoi, N. K. Multiwinner voting with fairness constraints. *International Joint Conference on Artificial Intelligence (IJCAI)*, 2018a.
- Celis, L. E., Straszak, D., and Vishnoi, N. K. Ranking with fairness constraints. In *International Colloquium on Automata, Languages and Programming (ICALP)*, 2018b.
- Deshpande, A. and Vempala, S. Adaptive sampling and fast low-rank matrix approximation. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, 9th International Workshop on Approximation Algorithms for Combinatorial Optimization Problems, APPROX 2006 and 10th International Workshop on Randomization and Computation, RANDOM 2006*, pp. 292–303, 2006.
- Drineas, P., Frieze, A. M., Kannan, R., Vempala, S., and Vinay, V. Clustering in large graphs and matrices. In *ACM-SIAM Symposium on Discrete Algorithms*, volume 99, pp. 291–299. Citeseer, 1999.
- Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, ITCS ’12*, pp. 214–226. ACM, 2012. ISBN 978-1-4503-1115-1.
- Gong, B., Chao, W., Grauman, K., and Sha, F. Diverse sequential subset selection for supervised video summarization. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014*, pp. 2069–2077, 2014.
- He, H. and Garcia, E. A. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009.
- Hesabi, Z. R., Tari, Z., Goscinski, A., Fahad, A., Khalil, I., and Queiroz, C. *Data Summarization Techniques for Big Data—A Survey*, pp. 1109–1152. Springer New York, New York, NY, 2015. ISBN 978-1-4939-2092-1.
- Hough, J. B., Krishnapur, M., Peres, Y., Virág, B., et al. Determinantal processes and independence. *Probability surveys*, 3:206–229, 2006.
- Image-dataset. URL goo.gl/hNukfP.
- Kamiran, F. and Calders, T. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, 2012.
- Kay, M., Matuszek, C., and Munson, S. A. Unequal representation and gender stereotypes in image search results for occupations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, CHI ’15*, pp. 3819–3828. ACM, 2015. ISBN 978-1-4503-3145-6.
- Koriyama, Y., Macé, A., Treibich, R., and Laslier, J.-F. Optimal apportionment. *Journal of Political Economy*, 121(3):584–608, 2013.
- Krause, A., Singh, A. P., and Guestrin, C. Near-optimal sensor placements in gaussian processes: Theory, efficient algorithms and empirical studies. *Journal of Machine Learning Research*, 9:235–284, 2008.
- Kulesza, A. and Taskar, B. k -dpps: Fixed-size determinantal point processes. In *Proceedings of the 28th International Conference on Machine Learning, ICML 2011*, pp. 1193–1200, 2011.
- Kulesza, A. and Taskar, B. Determinantal point processes for machine learning. *Foundations and Trends in Machine Learning*, 5(2-3):123–286, 2012.
- Li, C., Jegelka, S., and Sra, S. Efficient sampling for k -determinantal point processes. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, AISTATS 2016*, pp. 1328–1337, 2016.
- Lin, H. and Bilmes, J. A. Learning mixtures of submodular shells with application to document summarization. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 479–490, 2012.

- Lowe, D. G. Object recognition from local scale-invariant features. In *ICCV*, pp. 1150–1157, 1999.
- O’Neil, C. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown/Archetype, 2016. ISBN 9780553418828.
- Rebeschini, P. and Karbasi, A. Fast mixing for discrete point processes. In *Proceedings of The 28th Conference on Learning Theory, COLT 2015*, pp. 1480–1500, 2015.
- Roy, O. and Vetterli, M. The effective rank: A measure of effective dimensionality. In *15th European Signal Processing Conference, 2007*, pp. 606–610. IEEE, 2007.
- Tropp, J. A. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012.
- Vedaldi, A. and Fulkerson, B. Vlfeat: An open and portable library of computer vision algorithms., 2008. URL <http://www.vlfeat.org/>.
- Yang, K. and Stoyanovich, J. Measuring fairness in ranked outputs. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management, 2017*, pp. 22:1–22:6, 2017.
- Zadrozny, B. Learning and evaluating classifiers under sample selection bias. In *Machine Learning, Proceedings of the Twenty-first International Conference (ICML 2004)*, 2004.
- Zafar, M. B., Valera, I., Gomez-Rodriguez, M., and Gummadi, K. P. Fairness constraints: Mechanisms for fair classification. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017*, pp. 962–970, 2017.
- Zemel, R. S., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013*, pp. 325–333, 2013.
- Zhou, T., Kuscsik, Z., Liu, J.-G., Medo, M., Wakeling, J. R., and Zhang, Y.-C. Solving the apparent diversity-accuracy dilemma of recommender systems. *Proceedings of the National Academy of Sciences*, 107(10):4511–4515, 2010.