A Deferred Proofs of Section 3

Lemma 3.1. *Given a correlation clustering instance G, a fairlet decomposition $\mathcal{P}$ for G, and a clustering $\mathcal{C}$ of G, there exists a clustering $\mathcal{C}'$ of $G^{\mathcal{P}}$ such that*

$$\textsc{cost}(G^{\mathcal{P}}, \mathcal{C}') \leq \textsc{cost}(G, \mathcal{C}) + \textsc{fcost}^{out}(\mathcal{P}).$$

Proof. To show existence of the claimed clustering $\mathcal{C}'$, we devise a randomized algorithm and bound the expected cost of the clustering output of this algorithm. We abuse notation and let clustering $\mathcal{C}'$ be the output of this randomized algorithm on $G^{\mathcal{P}}$. Given fairlet decomposition $\mathcal{P}$, the algorithm first picks a representative r_i for each partition P_i in $\mathcal{P}$ uniformly at random. Next the algorithm defines clustering $\mathcal{C}'$ based on where r_i is assigned in $\mathcal{C}$: if r_i is placed in cluster C_j, it places the vertex p_i of $G^{\mathcal{P}}$ in the cluster C'_j.

Next, we bound the expected cost of $\mathcal{C}'$ in $G^{\mathcal{P}}$. Let us fix two vertices p_i and p_j in $G^{\mathcal{P}}$. We first consider the case $\sigma(p_i, p_j) > 0$; the other case follows by a similar argument. The clustering $\mathcal{C}'$ incurs a cost of $|\sigma(p_i, p_j)|$ if p_i and p_j are assigned to different clusters, which happens when the two representatives r_i and r_j are in different clusters in $\mathcal{C}$. Now, if $\sigma(r_i, r_j) = +1$, then $\mathcal{C}$ also pays a cost of 1 for separating r_i and r_j and if $\sigma(r_i, r_j) = -1$, then the edge (r_i, r_j) is an edge with the minority sign and it contributes to $\textsc{fcost}^{out}(P_i, P_j)$. Hence, if we denote the cost of the edge between p_i and p_j in the clustering $\mathcal{C}'$ by $\textsc{cost}_{\mathcal{C}'}(p_i p_j)$, we can bound the expected value of this cost as follows:

$$\begin{aligned}\mathbf{E}[\textsc{cost}_{\mathcal{C}'}(p_i, p_j)] \quad \leq \quad & |\sigma(p_i, p_j)| \cdot E[\textsc{cost}_{\mathcal{C}}(r_i, r_j) \\ & + \mathbb{1}\left(\sigma(p_i, p_j) * \sigma(r_i, r_j) < 0\right)],\end{aligned}$$

where $\mathbb{1}(A)$ is an indicator function having value 1 if A is true and zero otherwise. Now since r_i and r_j are picked uniformly at random,

$$\begin{aligned}&\mathbf{E}[\textsc{cost}_{\mathcal{C}'}(p_i, p_j)] \\ \leq\ & \frac{|\sigma(p_i, p_j)|}{|P_i| \cdot |P_j|} \cdot (\textsc{cost}_{\mathcal{C}}(P_i, P_j) + \textsc{fcost}^{out}(P_i, P_j)).\end{aligned}$$

This follows from the fact that there are $|P_i| \cdot |P_j|$ many possible pairs (r_i, r_j) to be selected. Summing the above over all p_i, p_j and using $|\sigma(p_i, p_j)| \leq |P_i| \cdot |P_j|$, we get

$$\mathbf{E}\big[\textsc{cost}(G^{\mathcal{P}}, \mathcal{C}')\big] \leq \textsc{cost}(\mathcal{C}) + \textsc{fcost}^{out}(\mathcal{P}). \quad \square$$

Lemma 3.2. *Assume $\mathcal{C}$ is a clustering of $G^{\mathcal{P}}$, and let $\mathcal{C}'$ be the clustering computed in line 4 of Algorithm 1 for G. Then we have*

$$\textsc{cost}(G, \mathcal{C}') \leq \textsc{cost}(G^{\mathcal{P}}, \mathcal{C}) + \textsc{fcost}(\mathcal{P}).$$

Proof. Any edge (u, v) contributing to the cost of the clustering $\mathcal{C}'$ is either a negative edge inside a fairlet or an edge between two fairlets that are clustered in disagreement with $\sigma(u, v)$ in $\mathcal{C}$. Negative edges inside fairlets are counted in $\textsc{fcost}^{in}(\mathcal{P})$. An edge (u, v) between fairlets P_i and P_j that is clustered in disagreement with $\sigma(u, v)$ either has the same sign as the majority sign of $E(P_i, P_j)$, or as the minority sign of $E(P_i, P_j)$. The edges in the former case are counted in $\textsc{cost}(G^{\mathcal{P}}, \mathcal{C})$, and the edges in the latter case are counted in $\textsc{fcost}^{out}(\mathcal{P})$. Therefore, the total cost of $\mathcal{C}'$ is at most $\textsc{cost}(G^{\mathcal{P}}, \mathcal{C}) + \textsc{fcost}^{in}(\mathcal{P}) + \textsc{fcost}^{out}(\mathcal{P})$. $\square$

Lemma 3.3. *For any constrained correlation clustering instance G, and any constrained clustering $\mathcal{C}$ of G, there is a fairlet decomposition $\mathcal{P}$ of G satisfying $\textsc{fcost}(\mathcal{P}) \leq \textsc{cost}(G, \mathcal{C})$.*

Proof. We can simply take $\mathcal{P}$ to be the same as the clustering $\mathcal{C}$. It is easy to observe that this is a valid fairlet decomposition. To bound $\textsc{fcost}(\mathcal{P})$, it is enough to note that each edge counted in $\textsc{fcost}^{in}(\mathcal{P})$ also imposes a cost of 1 in $\mathcal{C}$ (as it is a negative edge inside a cluster), and for any two clusters C_i and C_j, the number of positive edges between C_i and C_j is at least $\textsc{fcost}^{out}(C_i, C_j)$. Summing over all these inequalities, we obtain that $\textsc{fcost}(\mathcal{P})$ is at most $\textsc{cost}(G, \mathcal{C})$. $\square$

B Deferred Proofs of Section 4

Lemma 4.1. *For any fairlet decomposition $\mathcal{P}$, we have*

$$\textsc{mcost}(\mathcal{P}) \leq 2 \cdot \textsc{fcost}(\mathcal{P}).$$

Proof. Consider a fairlet P_i in $\mathcal{P}$. We define a vector $\mu \in [0, 1]^n$ indexed by the vertices of G as follows: $\mu_u = \text{majority}(\{\phi(v)_u : v \in P_i\})$. By the definition of $\textsc{mcost}$, we have

$$\begin{aligned}\textsc{mcost}(P_i) \quad &= \quad \min_{u \in [0,1]^n} \sum_{v \in P_i} d(u, v) \\ &\leq \quad \sum_{v \in P_i} |\mu - \phi(v)|. \qquad (1)\end{aligned}$$

On the other hand, for every vertex u, if we denote $N^-(u) = \{v : (u, v) \in E^-\}$ and $N^+(u) = \{u\} \cup \{v : (u, v) \in E^+\}$, we have

$$\begin{aligned}\sum_{v \in P_i} |\mu_u - \phi(v)_u| &= |\{v \in P_i : \phi(v)_u \neq \mu_u|\}| \\ &= \min(|N^-(u) \cap P_i|, |N^+(u) \cap P_i|).\end{aligned}$$

For $u \notin P_i$, the above quantity is precisely $\textsc{fcost}^{out}(P_i, \{u\})$. For $u \in P_i$, the above quantity

is at most $|N^-(u) \cap P_i|$, which is the number of negative edges in P_i incident on u. Therefore, the sum of this quantity over all u can be bounded by

$$\sum_{u \in V} \sum_{v \in P_i} |\mu_u - \phi(v)_u| \leq 2 \cdot \textsc{fcost}^{in}(P_i) + \textstyle\sum_{u \in V \setminus P_i} \textsc{fcost}^{out}(P_i, \{u\}). \tag{2}$$

Finally, by the definition of $\textsc{fcost}^{out}$, we have $\textsc{fcost}^{out}(P_i, S) + \textsc{fcost}^{out}(P_i, T) \leq \textsc{fcost}^{out}(P_i, S \cup T)$ for any two disjoint sets S and T. Therefore,

$$\sum_{u \in V \setminus P_i} \textsc{fcost}^{out}(P_i, \{u\}) \leq \sum_{j \neq i} \textsc{fcost}^{out}(P_i, P_j).$$

Combining this with Equations (1) and (2), we get $\textsc{mcost}(\mathcal{P}) \leq 2 \cdot \textsc{fcost}^{in}(\mathcal{P}) + \textsc{fcost}^{out}(\mathcal{P}) \leq 2 \cdot \textsc{fcost}(\mathcal{P})$. □

Lemma 4.2. *Let $\mathcal{P}$ be any fairlet decomposition and let $f = \max_{P \in \mathcal{P}} |P|$. Then,*

$$\textsc{fcost}(\mathcal{P}) \leq 2f \cdot \textsc{mcost}(\mathcal{P}).$$

Proof. Consider a fairlet P_i and define the vector μ as in the proof of Lemma 4.1. It is easy to see that

$$\begin{aligned} \textsc{mcost}(P_i) &= \min_{x \in [0,1]^n} \sum_{v \in P_i} |x - \phi(v)| \\ &= |\mu - \phi(v)| \\ &= \sum_{u \in V} \min(|N^-(u) \cap P_i|, |N^+(u) \cap P_i|). \end{aligned}$$

As in the proof of Lemma 4.1, for every $u \notin P_i$, the summand in the above expression is precisely $\textsc{fcost}^{out}(P_i, \{u\})$. For $u \in P_i$, this quantity is zero if u has no negative edge to any other vertex in P_i, and is at least 1 otherwise. Therefore, since $|P_i| \leq f$, this quantity is always at least the number of negative edges from u to other vertices in P_i divided by f. Therefore,

$$\textsc{mcost}(P_i) \geq \frac{2}{f} \cdot \textsc{fcost}^{in}(P_i) + \sum_{j : j \neq i} \sum_{u \in P_j} \textsc{fcost}^{out}(P_i, \{u\}).$$

Summing over all i and rearranging the terms, we obtain:

$$\textsc{mcost}(\mathcal{P}) \geq \frac{2}{f} \cdot \textsc{fcost}^{in}(\mathcal{P}) + \sum_{i<j} \left(\sum_{u \in P_j} \textsc{fcost}^{out}(P_i, \{u\}) + \sum_{u \in P_i} \textsc{fcost}^{out}(P_j, \{u\}) \right). \tag{3}$$

Now, we fix $i < j$ and bound the summand in the above expression. Without loss of generality, we assume $|P_i| \geq |P_j|$. We consider the following cases:

Case 1: There are at most $|P_i|/2$ vertices u in P_i with $\textsc{fcost}^{out}(P_j, \{u\}) = 0$. In this case, we have $\sum_{u \in P_i} \textsc{fcost}^{out}(P_j, \{u\}) \geq \frac{|P_i|}{2} \geq \frac{|P_i| \cdot |P_j|}{2f} \geq \frac{\textsc{fcost}^{out}(P_i, P_j)}{2f}$.

Case 2: There are at least $|P_i|/2$ vertices u in P_i with $\textsc{fcost}^{out}(P_j, \{u\}) = 0$. Let S be the set of such vertices. By the definition of $\textsc{fcost}^{out}(P_j, \{u\})$, any $u \in S$ must have either positive edges to all vertices in P_j, or negative edges to all of them. Assume x vertices in S have positive edges to all vertices in P_j and y of them have negative edges, for some x, y with $x + y = |S| \geq |P_i|/2$. We further consider the following cases:

- **Case 2a:** If $x = 0$, then every vertex u in P_j has at least $P_i/2$ positive edges to vertices in P_i (namely, at least to those in S). Therefore, $\textsc{fcost}^{out}(P_i, \{u\}) = |E^- \cap E(P_i, \{u\})|$. Thus, $\sum_{u \in P_j} \textsc{fcost}^{out}(P_i, \{u\}) = |E^- \cap E(P_i, P_j)| \geq \textsc{fcost}^{out}(P_i, P_j)$.
- **Case 2b:** If $y = 0$, an argument similar to case 2a shows that $\sum_{u \in P_j} \textsc{fcost}^{out}(P_i, \{u\}) = |E^+ \cap E(P_i, P_j)| \geq \textsc{fcost}^{out}(P_i, P_j)$.
- **Case 2c:** If $x \geq 1$ and $y \geq 1$, then each vertex in P_j has at least one positive edge and at least one negative edge to P_i. Therefore, for every $u \in P_j$, $\textsc{fcost}^{out}(P_i, \{u\}) \geq 1$. Thus, $\sum_{u \in P_j} \textsc{fcost}^{out}(P_i, \{u\}) \geq |P_j| \geq \frac{|P_i| \cdot |P_j|}{f} \geq \frac{\textsc{fcost}^{out}(P_i, P_j)}{f}$.

In all of the above cases, we have:

$$\sum_{u \in P_j} \textsc{fcost}^{out}(P_i, \{u\}) + \sum_{u \in P_i} \textsc{fcost}^{out}(P_j, \{u\}) \geq \frac{\textsc{fcost}^{out}(P_i, P_j)}{2f}.$$

This, together with (3), implies:

$$\begin{aligned} \textsc{mcost}(\mathcal{P}) &\geq \frac{2}{f} \cdot \textsc{fcost}^{in}(\mathcal{P}) + \frac{1}{2f} \cdot \textsc{fcost}^{out}(\mathcal{P}) \\ &\geq \frac{1}{2f} \cdot \textsc{fcost}(\mathcal{P}). \quad \square \end{aligned}$$

Theorem 4.6. *For $\alpha = 1/2$, there is a 256-approximation algorithm for fair correlation clustering.*

Proof. From Theorem 4.3 and Lemma 4.5, we get a 24-approximation algorithm for solving fairlet decomposition with minimum FCOST ($f = 3, \gamma = 2$). From Lemma 3.5, there is a $2 \cdot 2.06 \cdot (1.5)^2 = 9.27$-approximation algorithm for unconstrained correlation clustering. Combining, we get a 255.75-approximation algorithm for fair correlation clustering. □

Theorem 4.8. *For $\alpha = 1/C$, there is a $(16.48C^2)$-approximation algorithm for fair correlation clustering.*

Proof. From Theorem 4.3 and Lemma 4.5, we get a $4C^2$-approximation algorithm for solving fairlet decomposition with minimum FCOST ($f = C, \gamma = C$). From Lemma 3.5, there is a $2 \cdot 2.06 \cdot 1 = 4.12$-approximation algorithm for unconstrained correlation clustering. Combining, we get a $(16.48C^2)$-approximation algorithm for fair correlation clustering. □

Theorem 4.10. *For $\alpha = 1/t$, given an γ-approximation for fair decomposition with median cost, there exists an $O(t\gamma)$-approximation algorithm for fairlet correlation clustering.*

Proof. Let $\mathcal{P}$ be the output of the output of the γ-approximation algorithm on the metric space (M, d) obtained from correlation instance G. Let fairlet decomposition $\mathcal{P}'$ be obtained from $\mathcal{P}$ by applying Lemma 4.9 and assigning each fairlet to a center minimizing the median cost of the fairlet. Since dedicating a center to a subset of points assigned to the same center in $\mathcal{P}$ can only decrease the median cost, $\text{MCOST}(\mathcal{P}') \leq \text{MCOST}(\mathcal{P})$. From Theorem 4.3 and Lemma 4.5, there is a $((8t-4)\gamma)$-approximation algorithm for solving fairlet decomposition with minimum FCOST ($f = 2t - 1, \gamma = \gamma_{\mathcal{A}}$). Since the size of each fairlet is at least t, applying Lemma 3.5, there is a $2 \cdot 2.06 \cdot (\frac{2t-1}{t})^2 < 16.48$-approximation algorithm for solving unconstrained correlation clustering. Now applying Theorem 3.4, we get an $O(t\gamma)$-approximation algorithm for fair correlation clustering. □

C Supplemental Experimental Results

Here, we report additional experimental results.

C.1 Description of the datasets

We describe more in detail the datasets used.

amazon: Vertices represents products on the Amazon website (Leskovec et al., 2007) and positive edges connect products co-reviewed by the same user (all missing edges are treated as negative). We set the color of each item to its category. Further, we use 1000 vertices equally distributed among 2 popular book categories *Nonfiction* and *Literature & Fiction* for a total of $\sim$ 106,000 positive edges.

reuters: This graph is extracted from a dataset, which was used in previous fair clustering work (Ahmadian et al., 2019) and includes 50 English language articles from each of up to 16 authors (for a total of up to 800 texts).This dataset is available at `archive.ics.uci.edu/ml/datasets/Reuter_50_50`. We transform each text into a 10-dimensional vector using Gensim's Doc2Vec with standard parameters, as in previous work (Ahmadian et al., 2019), and we create one vertex for each text. Then we use a threshold on the dot product of the embedding vectors. Through this operation, we set the top $\theta \in \{0.25, 0.50, 0.75\}$ fraction of edges via dot products as $+1$'s ,and the remaining edges are assigned -1's. Note that the colors represent the text authors.

victorian: Similarly, for the victorian dataset, available at `archive.ics.uci.edu/ml/datasets/Victorian+Era+Authorship+Attribution`. We use texts from up to 16 English-language authors from the Victorian era. Each text consists of 1,000-word sequences obtained from a book written by one of these authors (we use the training dataset). The data was extracted and processed as in Gungor (2018). From each document, we extract a 10-dimensional vector using Gensim's Doc2Vec with the standard parameter settings again, and we assign the author id as color, as in prior work (Ahmadian et al., 2019). We use 100 texts from each author, create one vertex for each text, and set the top $\theta \in \{0.25, 0.50, 0.75\}$ fraction of pairwise dot product edges as positive, and the remaining edges as negative. All graphs are unweighted and complete.

C.2 Other experimental results

In Table 3 we report an overview of the results of the various algorithms for a dataset extracted from Amazon involving 250 vertex for each of 4 colors corresponding to the book categories *Literature & Fiction, Nonfiction, Business & Investing, Computers & Internet.*

Similarly in Table 4 we report the results for reuters, $\theta = 0.50$, $c = 8$ colors.

Finally, in Table 5 we report an evaluation of our algorithms in a variety of datasets and for different number of colors.

Notice how in all cases the results matches qualitatively the results reported in the main paper.

Algorithm	Error	Imbalance for 1/2	Imbalance for equality
Local	0.005	0.375	0.541
Pivot	0.009	0.365	0.529
Match + Local	0.006	0	0.518
Rep. Match + Local	0.070	0	0
Single	0.828	0	0
Rand	0.173	0	0

Table 3: Experimental results for amazon, $C = 4$ colors.

Algorithm	Error	Imbalance for 1/2	Imbalance for equality
Local	0.239	0.036	0.344
Pivot	0.35	0.024	0.298
Match + Local	0.251	0	0.310
Rep. Match + Local	0.416	0	0
Single	0.501	0	0
Rand	0.500	0	0

Table 4: Experimental results for reuters, $\theta = 0.50$, $C = 8$ colors.

dataset	C	Error Local	Error Rep. Match + Local
reuters, $\theta = 0.25$	2	0.096	0.230
	4	0.120	0.244
	8	0.133	0.252
	16	0.146	0.255
reuters, $\theta = 0.50$	2	0.181	0.350
	4	0.191	0.336
	8	0.239	0.416
	16	0.258	0.391
reuters, $\theta = 0.75$	2	0.188	0.199
	4	0.211	0.227
	8	0.237	0.250
	16	0.220	0.250
victorian, $\theta = 0.25$	2	0.109	0.212
	4	0.141	0.210
	8	0.161	0.212
	16	0.150	0.222
victorian, $\theta = 0.50$	2	0.183	0.348
	4	0.228	0.311
	8	0.249	0.319
	16	0.232	0.343
victorian, $\theta = 0.75$	2	0.203	0.237
	4	0.225	0.245
	8	0.218	0.246
	16	0.215	0.250

Table 5: Experimental results for various datasets and number of colors.

Additional baselines. We further experimented with two other greedy baselines. First, we tried the following (unfair) greedy baseline: in an arbitrary order, iterate over the vertices, and for each vertex, add it to either the current cluster with most positive neighbors (if it exists) or to a singleton cluster. More precisely, we assign the vertex to the best current cluster, if it is connected with more positive edges than negative edges to it, otherwise we leave the vertex as a singleton. Unsurprisingly, this unfair baseline is worse than all other unfair baselines we considered in terms of error and it has also a large imbalance, so we omit the results.

We also tested a fair greedy baseline for $\alpha = \frac{1}{2}$, for $C = 2$: sort all pairs of different color vertices by distance in the Hamming space in an increasing order, and assign vertices to clusters of size 2 with a greedy matching algorithm over this order. This creates fair clusters but again, we observe that this baseline to be close to that of RAND and as such we omit the results.

Running time. All experiments have been conducted on commodity hardware. Each run of an algorithm completed in less than an hour. In our experiments, our fair algorithms have a running time in the same order of magnitude of that of the local search heuristic. For instance, for reuters, $\theta = 0.50$, the ratio of mean running time of MATCH +LOCAL and REP. MATCH + LOCAL w.r.t. LOCAL was 90% and 29%, respectively. For victorian, $\theta = 0.50$ it was 123% and 41%, respectively.