Aggregating Inconsistent Information: Ranking and Clustering

Nir Ailon[*]
Dept. of Computer Science
Princeton University
Princeton, NJ

Moses Charikar[†]
Dept. of Computer Science
Princeton University
Princeton, NJ

Alantha Newman[‡]
Dept. of Computer Science
RWTH Aachen
Aachen, Germany

ABSTRACT

We address optimization problems in which we are given contradictory pieces of input information and the goal is to find a globally consistent solution that minimizes the number of disagreements with the respective inputs. Specifically, the problems we address are rank aggregation, the feedback arc set problem on tournaments, and correlation and consensus clustering. We show that for all these problems (and various weighted versions of them), we can obtain improved approximation factors using essentially the same remarkably simple algorithm. Additionally, we almost settle a long-standing conjecture of Bang-Jensen and Thomassen and show that unless NP⊆BPP, there is no polynomial time algorithm for the problem of minimum feedback arc set in tournaments.

Categories and Subject Descriptors

F.2.0 [**Analysis of Algorithms and Problem Complexity**]: General

General Terms

Algorithms, Theory

[*]Email: nailon@cs.princeton.edu.

[†]Email: moses@cs.princeton.edu.
Supported by NSF ITR grant CCR-0205594, DOE Early Career Principal Investigator award DE-FG02-02ER25540, NSF CAREER award CCR-0237113, an Alfred P. Sloan Fellowship and a Howard B. Wentz Jr. Junior Faculty Award.
[‡]Email: alantha@cs.rwth-aachen.de.
Supported in part by the EU within the 6th Framework Programme under contract 001907 (DELIS). Work done while visiting Princeton University, supported by Moses Charikar's Alfred P. Sloan fellowship.

STOC'05, May 22-24, 2005, Baltimore, Maryland, USA.

Keywords

Rank Aggregation, Consensus Clustering, Correlation Clustering, Minimum Feedback Arc Set, Tournament

1. INTRODUCTION

The problem of aggregating inconsistent information from many different sources arises in numerous contexts and disciplines. For example, the problem of ranking a set of contestants or a set of alternatives based on possibly conflicting preferences is a central problem in the areas of voting and social choice theory. Specifically, combining k different complete ranked lists on the same set of n elements into a single ranking, which best describes the preferences expressed in the given k lists, is known as the problem of *rank aggregation*. This problem dates back to as early as the late 18th century when Condorcet and Borda each proposed voting systems for elections with more than two candidates [11, 8]. There are numerous applications in sports, databases, and statistics [14, 17] in which it is necessary to effectively combine rankings from different sources. Another example of aggregating information is the problem of integrating possibly contradictory clusterings from existing data sets into a single representative cluster. This problem is known as *consensus clustering* or *ensemble clustering* and can be applied to remove noise and inconguencies from data sets [18] or combine information from multiple classifiers [30].

In the last half century, rank aggregation has been studied and defined from a mathematical perspective. In particular, Kemeny proposed a precise criterion for determining the "best" aggregate ranking[1] [24, 23]. Given n candidates and k permutations of the candidates, $\{\pi_1, \pi_2, \ldots, \pi_k\}$, a *Kemeny optimal* ranking of the candidates is the ranking π that minimizes $\sum_i^k d(\pi, \pi_i)$, where $d(\pi_j, \pi_k)$ denotes[2] the number of pairs of candidates that are ranked in different orders by π_j and π_k. For example, if $\pi_j = (1, 2, 3, 4)$ and $\pi_k = (2, 3, 1, 4)$, then $d(\pi_j, \pi_k) = 2$ since elements 1 and 2 appear in different orders in the two rankings as do elements 1 and 3. In other words, a Kemeny optimal ranking minimizes the number of pairwise *disagreements* with the given k rankings. Throughout this paper we will slightly abuse terminology and refer to the problem of finding a Kemeny optimal ranking as the RANK-AGGREGATION problem.

[1]Historically known as *Kemeny aggregation*.
[2]The distance function $d(\cdot, \cdot)$ is known as the *Kendall tau* distance.

More recently, the RANK-AGGREGATION problem has been studied from a computational perspective. Finding a Kemeny optimal ranking is NP-hard [7] and remains NP-hard even when there are only four input lists to aggregate [14]. This motivates the problem of finding a ranking that *approximately* minimizes the number of disagreements with the given input rankings. Several 2-approximation algorithms [12, 14] are known. In fact, if we take the best of the input rankings, then the number of disagreements between this ranking and the k input rankings is no more than twice optimal.

The problem of finding a globally consistent ranking based on possibly conflicting pairwise information arises in the well studied problem of feedback arc set in a digraph—specifically as the feedback arc set problem on *tournaments*. Throughout the paper, we refer to this problem as FAS-TOURNAMENT. A tournament is a directed graph $G = (V, A)$ such that for each pair of vertices $i, j \in V$, either $(i, j) \in A$ of $(j, i) \in A$. The minimum feedback arc set is the smallest set $A' \subseteq A$ such that $(V, A - A')$ is acyclic. The size of this set is exactly the minimal number of backward edges induced by a linear ordering of V. This problem turns out to be useful in studying RANK-AGGREGATION, but is also interesting in its own right. For example, imagine a sports tournament where each player plays against every other player once: How should we rank the players based on these possibly non-transitive (inconsistent) outcomes? The complementary problem to finding a minimum feedback arc set is the *maximum acyclic subgraph* problem, also known as the *linear ordering* problem.

The RANK-AGGREGATION problem can be cast as a special case of *weighted* FAS-TOURNAMENT, where the objective is to minimize the total weight of backward edges in a linear order of the vertices. When the weight of edge (i, j) is the fraction of input rankings that order i before j, solving RANK-AGGREGATION is equivalent to solving this weighted FAS-TOURNAMENT instance.

The last problem we consider is that of clustering objects based on complete but possibly conflicting pairwise information. An instance of this problem can be represented by a graph with a vertex for each object and a '+' or a '-' for each pair of vertices, indicating that two elements should be in the same or different clusters, respectively. The goal is to cluster the elements so as to minimize the number of '-' edges within clusters and '+' edges crossing clusters. We refer to this problem as CORRELATION-CLUSTERING (on complete graphs). An analog to RANK-AGGREGATION is known as the CONSENSUS-CLUSTERING problem. In this problem, we are given k clusterings of the same set of n elements. The goal is to find a clustering that minimizes the number of pairwise disagreements with the given k clusterings.

1.1 Previous Work

The minimum feedback arc set problem can be approximated to within $O(\log n \log\log n)$ in general graphs [16, 28] and has (at least) the same approximation hardness as the vertex cover problem [22], which is 1.36 [13]. More than a decade ago, Bang-Jensen and Thomassen conjectured that the FAS-TOURNAMENT problem is NP-hard [5]. However, for the past decade, no progress has been made on settling this conjecture. In contrast, the minimum feedback *vertex* set problem on tournaments is NP-hard [29] and is approximable to within 2.5 [9]

We are not aware of any approximation for FAS-TOURNAMENT that improves on the bound for the general feedback arc set problem. The complementary maximization problem on tournaments seems to be easier from an approximation standpoint. Arora, Frieze and Kaplan [4] and Frieze and Kannan [19] gave PTASs for the maximum acyclic subgraph problem in dense graphs, which implies a PTAS for the problem on tournaments.

There are two well-known factor 2-approximation algorithms for RANK-AGGREGATION. One such approximation is to pick one of the k given permutations at random. We will call this algorithm PICK-A-PERM. The *Spearman's footrule* distance between two permutations π_i and π_j on n elements is defined to be: $F(\pi_i, \pi_j) = \sum_{k=1}^{n} |\pi_i(k) - \pi_j(k)|$. The footrule distance is no more than twice the Kemeny distance [12] and can be computed in polynomial time via a minimum cost matching [14, 15]. These observations yield another 2-approximation.

CORRELATION-CLUSTERING has been studied both on general and complete graphs. Both minimization and maximizing versions have been investigated. Bansal, Blum and Chawla gave the first constant factor approximation for the problem of minimizing disagreements on the complete graph [6]. This factor was improved to 4 by rounding a linear program [10]. The weighted version of this problem in which edges have fractional $\pm$ assignments has also been studied. Each edge is assigned fractional values w_{ij}^{+} and w_{ij}^{-} rather than a discrete $'+'$ or $'-'$ label. When the edge weights satisfy the *probability constraints* (i.e. $w_{ij}^{+} + w_{ij}^{-} = 1$ for all edges), the best previous approximation factor was 7 [20, 6]. When the edge weights satisfy the probability and the triangle inequality constraints (see Section 1.2), the best previous approximation factor was 3 [20].

CORRELATION-CLUSTERING on complete graphs is MAX-SNP-hard [10] and CONSENSUS-CLUSTERING is NP-hard [31]. However, CONSENSUS-CLUSTERING is not known to be NP-hard if the number of input clusters is constant [18]. Analogously to RANK-AGGREGATION, choosing the best cluster out of the given k input clusters (algorithm PICK-A-CLUSTER) is an expected 2-approximation algorithm and 2 was the best previously known approximation factor for this problem.

1.2 Our Results

We give improved approximation algorithms for the following optimization problems: (i) FAS-TOURNAMENT, (ii) RANK-AGGREGATION, (iii) CORRELATION-CLUSTERING and (iv) CONSENSUS-CLUSTERING. We show that they can all be approximated using essentially the same remarkably simple algorithm. For example, the algorithm for FAS-TOURNAMENT, called FAS-PIVOT, is as follows: First, we pick a random vertex i to be the "pivot" vertex. Second, we place all vertices connected to i with an in-edge on the left side of i and all vertices connected to i with an out-edge on the right side of i. We then recurse on the two tournaments induced by the vertices on each side.

The analysis of FAS-PIVOT yields a 3-approximation algorithm for FAS-TOURNAMENT, improving on the best-known previous factor of $O(\log n \log\log n)$. Our analysis relies on a new technique for arguing a lower bound for FAS-TOURNAMENT by demonstrating a fractional packing of edge disjoint directed triangles. We apply this algorithm to RANK-AGGREGATION as follows. We convert the RANK-AGGRE-

GATION instance into a weighted FAS-TOURNAMENT instance, which we then convert to an unweighted FAS-TOURNAMENT instance using the *majority* tournament (see Definition 1). Finally we run FAS-PIVOT on this majority tournament. Although this algorithm by itself is yet another 2-approximation, the following is an 11/7-approximation: run both FAS-PIVOT and PICK-A-PERM and output the best solution. This improved approximation ratio is due to the fact that each algorithm does well on instances in which the other algorithm does poorly.

For CORRELATION-CLUSTERING and CONSENSUS-CLUSTERING we present similar combinatorial algorithms and analyses which, interestingly, give results that are analogous to the results for FAS-TOURNAMENT and RANK-AGGREGATION and improve upon previously known approximation factors.

A simple lower bound on the value of an optimal solution for weighted FAS-TOURNAMENT is to take the sum over all vertices $i < j$ of $\min\{w_{ij}, w_{ji}\}$. In contrast, our analysis uses a stronger lower bound based on the weight of directed triangles ("*bad triangles*") in the majority graph. Interestingly, the analysis of our simple combinatorial algorithm bounds the integrality gap of a natural LP relaxation for FAS-TOURNAMENT. In fact, it demonstrates an LP dual solution based on probabilities of random events occuring during the execution [3]. A similar analysis is done for CONSENSUS-CLUSTERING, with a different notion of "*bad triplets*".

Our analysis is applied to various cases of weighted FAS-TOURNAMENT (resp. weighted CORRELATION-CLUSTERING). More precisely, we analyze the following cases:

(i) **Probability Constraints:** $w_{ij}+w_{ji} = 1$ (resp. $w^+_{ij}+ w^-_{ij} = 1$) for all $i, j \in V$.

(ii) **Triangle Inequality:** $w_{ij} \leq w_{ik} + w_{kj}$ (resp. $w^-_{ij} + w^-_{jk} \leq w^-_{jk}$) for all $i, j, k \in V$.

(iii) **Aggregation:** Edge weights are a convex combination of actual permutations (resp. clusters). Constraints (i) and (ii) are implied in this case.

Table 1 summarizes the approximation factors we achieve for the different scenarios with the combinatorial algorithms. Additionally, we consider LP relaxations for FAS-TOURNAMENT and CORRELATION-CLUSTERING. After choosing a pivot vertex, instead of deterministically placing vertices on the right or left side (in FAS-PIVOT), or in a cluster (in CC-PIVOT), we decide randomly based on LP values. This results in vastly improved approximation factors. We state improvements on the approximation guarantees obtainable via our LP rounding techniques based on inequalities that are proven in [1].

Finally, we show that FAS-TOURNAMENT has no polynomial time algorithm assuming NP$\not\subseteq$BPP. The question of NP-hardness of FAS-TOURNAMENT has been a long-standing conjecture of Bang-Jensen and Thomassen [5]. We show a randomized reduction from the problem of finding a minimum feedback arc set in general digraphs (which is known to be NP-hard) to the special case of tournaments. This proof has been recently derandomized by Noga Alon [2], and the conjecture is therefore proven completely. We present the weaker randomized version here.

[3]We will not pursue the discussion on the dual LP in this extended abstract.

	Ordering	Clustering
Unweighted Tournaments	3 (*)	3 (4)[10]
Probability Constraints (i)	5 (*)	5 (9) [10, 6]
Triangle Inequality (ii)	3 (*)	N/A (**)
Probability Constraints + Triangle Inequality (i,ii)	2 (*)	2 (3) [20]
Aggregation (iii)	11/7 (2)	11/7 (2)

Table 1: The previous best-known factors are shown in parentheses. (*) The best-known factor was the $O(\log n \log\log n)$ algorithm [16, 28] for digraphs. (**) Our techniques cannot directly be applied to weighted CORRELATION-CLUSTERING with triangle inequality but no probability constraints.

1.3 Organization

In Section 2, we give precise problem statements and definitions. In Section 3, we present FAS-PIVOT and analyze its approximation guarantee, introducing the basic ideas we use throughout the paper. In Section 4, we extend these ideas to approximate weighted FAS-TOURNAMENT . In Section 5, we further extend our techniques to approximate RANK-AGGREGATION. In Section 6, we discuss CORRELATION-CLUSTERING and CONSENSUS-CLUSTERING. In Section 7, we extend our ideas to round LP's for FAS-TOURNAMENT and CORRELATION-CLUSTERING. In Section 8, we prove hardness results for FAS-TOURNAMENT. In Section 9 we discuss open problems and future work.

2. PRELIMINARIES AND DEFINITIONS

We study the following problems in this paper. In what follows, we fix a ground set $V = \{1, \ldots, n\}$.

FAS-TOURNAMENT: (Minimum Feedback Arc Set in Tournaments) We are given a tournament $G = (V, A)$ (a digraph with either $(i, j) \in A$ or $(j, i) \in A$ for all distinct $i, j \in V$). We want to find a permutation π on V minimizing the number of pairs i, j such that $i <_\pi j$ and $(j, i) \in A$ (backward edges w.r.t. π)[4]. In a *weighted* FAS-TOURNAMENT instance, we are given weights $w_{ij} \geq 0$ for all ordered $i, j \in V$. We want to find a permutation π on V minimizing $\sum_{i,j:i<_\pi j} w_{ji}$. Clearly, the unweighted case can be encoded as a 0/1 weighted case.

RANK-AGGREGATION: We are given a list of k permutations (rankings) $\pi_1, .., \pi_k$ on V. We want to find a permutation π minimizing the sum of distances $\sum_{i=1}^{k} d(\pi, \pi_i)$, where $d(\pi, \rho)$ is the number of pairs i, j such that $i <_\pi j$ but $j <_\rho i$ (the Kemeny distance).

CORRELATION-CLUSTERING: Between any two unordered $i, j \in V$ we either have a $(+)$ or a $(-)$ relation. We let E^+ (resp. E^-) denote the set of pairs $i \neq j$ which are $(+)$-related (resp. $(-)$-related). We want to find disjoint clusters $C_1, \ldots, C_m$ covering V and minimizing the num-

[4]By $i <_\pi j$ we mean that π ranks i before j.

ber of disagreement pairs ((+) pairs in different clusters or (−) pairs in the same cluster). In a *weighted* CORRELATION-CLUSTERING instance, we assign for each pair i, j two weights $w_{ij}^+ \geq 0$ and $w_{ij}^- \geq 0$. The cost of a clustering will now be the sum of w_{ij}^+ over all i, j in different clusters, plus the sum of w_{ij}^- over all i, j in the same cluster. Clearly, the unweighted case can be encoded as a 0/1 weighted case.

CONSENSUS-CLUSTERING: We are given a list of k different clusterings $\mathcal{C}_1, \ldots, \mathcal{C}_k$ of V, and we wish to find one clustering $\mathcal{C}$ that minimizes $\sum_{i=1}^k d(\mathcal{C}, \mathcal{C}_i)$, where the distance $d(\mathcal{C}, \mathcal{D})$ between two clusterings is the number of unordered paris $i, j \in V$ that are clustered together by one and separated by the other.

DEFINITION 1. *Given an instance (V, w) of weighted* FAS-TOURNAMENT*, we define the unweighted* majority *tournament $G_w = (V, A_w)$ as follows: $(i, j) \in A_w$ if $w_{ij} > w_{ji}$. If $w_{ij} = w_{ji}$, then we decide $(i, j) \in A_w$ or $(j, i) \in A_w$ arbitrarily. Given an instance (V, w^+, w^-) of weighted* CORRELATION-CLUSTERING*, we define the unweighted* majority *instance (V, E_w^+, E_w^-) as follows: $(i, j) \in E_w^+$ if $w_{ij}^+ > w_{ij}^-$, and $(i, j) \in E_w^-$ if $w_{ij}^- > w_{ij}^+$. If $w_{ij}^+ = w_{ij}^-$, then we decide arbitrarily.*

Note that although the majority instances depend on the weights of the weighted instances, they are *unweighted* instances.

3. MINIMUM FEEDBACK ARC SET IN TOURNAMENTS

Let $G = (V, A)$ be a FAS-TOURNAMENT instance. We present the following algorithm FAS-PIVOT for approximating it.

```
FAS-PIVOT(G = (V, A))

  Set V_L → ∅, V_R → ∅.
  Pick random pivot i ∈ V.

  For all vertices j ∈ V \ {i}:
      If (j, i) ∈ A then
          Add j to V_L (place j on left side).
      Else (If (i, j) ∈ A)
          Add j to V_R (place j on right side).

  Let G_L = (V_L, A_L) be tournament induced by V_L.
  Let G_R = (V_R, A_R) be tournament induced by V_R.

  Return order FAS-PIVOT(G_L), i, FAS-PIVOT(G_R).
    (Concatenation of left recursion, i, and right recursion.)
```

THEOREM 2. *Algorithm* FAS-PIVOT *is a randomized expected 3-approximation algorithm for* FAS-TOURNAMENT.

PROOF. Let C^{OPT} denote the cost of an optimal solution. Let C^{PIV} denote the cost of FAS-PIVOT on $G = (V, A)$. We want to show that $E[C^{PIV}] \leq 3C^{OPT}$.

An edge $(i, j) \in A$ becomes a backward edge if and only if there exists a third vertex k such that (i, j, k) form a directed triangle[5] in G and k was chosen as a pivot when all three were input to the same recursive call. Pivoting on k would then place i to its right and j to its left, rendering edge (i, j) backward. In this case, we will charge a unit cost of the backward edge (i, j) to the directed triangle (i, j, k). Let T denote the set of directed triangles. For a directed triangle $t \in T$, denote by A_t the event that one of its vertices is chosen as pivot when all three are part of the same recursive call. Let p_t denote the probability of event A_t. Now we observe, that a triangle t is charged a unit cost exactly when A_t occurs, and it can be charged at most once. Therefore, the expected cost of FAS-PIVOT is exactly $E[C^{PIV}] = \sum_{t \in T} p_t$.

Clearly, if we had a set of *edge disjoint* triangles, then a lower bound for C^{OPT} would be its cardinality. This is also true *fractionally*: If $\{\beta_t\}_{t \in T}$ is a system of nonnegative weights on triangles in T such that for all $e \in A$, $\sum_{t: e \in t} \beta_t \leq 1$, then $C^{OPT} \geq \sum_{t \in T} \beta_t$. Indeed, consider the following LP relaxation for the problem: minimize $\sum_{e \in A} x_e$, subject to $x_{e_1} + x_{e_2} + x_{e_3} \geq 1$ for edge sets $\{e_1, e_2, e_3\} \in T$, and $x_e \geq 0$ for all $e \in A$. The solution to this LP clearly lower bounds C^{OPT}. It is easy to show that a packing $\{\beta_t\}$ is a feasible solution to the dual LP, hence a lower bound on the optimal.

We will demonstrate such a *packing* using the probabilities p_t. Let $t = (i, j, k)$ be some triangle. Conditioned on the event A_t, each one of the 3 vertices of t was the breaking vertex with probability 1/3, because all vertices input to a recursive call are chosen as pivot with equal probability. Therefore, any edge $e = (i, j)$ of t becomes a backward edge with probability 1/3 (still, conditioned on A_t). Let B_e denote the event that e becomes a backward edge. Clearly, if A_t occured then B_e occured if and only if k was the first chosen pivot among i, j, k. Then we get that for all $t \in T$ and $e \in t$,

$$\Pr[B_e \wedge A_t] = \Pr[B_e | A_t] \Pr[A_t] = \frac{1}{3} p_t.$$

The main observation of this proof is as follows: For two different triangles $t, t' \in T$ sharing an edge e, the events $B_e \wedge A_t$ and $B_e \wedge A_{t'}$ are disjoint. Indeed, if e is charged to triangle t, then the endpoints of e are split between two different recursive calls, and event $A_{t'}$ cannot occur. Therefore, for all $e \in E$,

$$\sum_{t: e \in t} \frac{1}{3} p_t \leq 1 \ . \tag{1}$$

So $\{p_t/3\}_{t \in T}$ is a fractional packing of T. Thus, $C^{OPT} \geq \sum_{t \in T} p_t/3 = E[C^{PIV}]/3$, as required. □

4. MINIMUM FEEDBACK ARC SET IN WEIGHTED TOURNAMENTS

Let (V, w) be a weighted FAS-TOURNAMENT instance. We suggest the following approximation algorithm: construct the unweighted majority tournament $G_w = (V, A_w)$ and return the ordering generated by FAS-PIVOT(G_w). We analyze this algorithm.

For an edge $e = (i, j) \in A_w$, we let $w(e) = w_{ij}$, and $\bar{w}(e) = w_{ji} = 1 - w(e) \leq w(e)$. Fix an optimal solution π^*, and let $c^*(e)$ denote the cost incurred to it by $e = (i, j) \in A_w$, that is, $c^*(e) = w(e)$ if $j <_{\pi^*} i$, else $c^*(e) = \bar{w}(e)$. So $C^{OPT} = \sum_{e \in A_w} c^*(e)$. Let T denote the set of directed triangles in G_w. For any $t = (e_1, e_2, e_3) \in T$, we define $c^*(t) = c^*(e_1) + c^*(e_2) + c^*(e_3)$, $w(t) = w(e_1) + w(e_2) + w(e_3)$. Finally, let C^{PIV} denote the cost the solution returned by FAS-PIVOT(V, G_w).

[5] In what follows we will use (i, j, k) to denote the directed triangle $i \to j, j \to k, k \to i$. It will be clear from the context whether a triangle is the set of its vertices or its edges.

THEOREM 3. *For an instance (V, w) of weighted* FAS-TOURNAMENT*, if there exists a constant $\alpha > 0$ such that $w(t) \leq \alpha c^*(t)$ for all $t \in T$, then $E[C^{PIV}] \leq \alpha C^{OPT}$, i.e.* FAS-PIVOT$(G_w)$ *is an expected α-approximation solution.*

PROOF. We generalize techniques presented in Section 3. When FAS-PIVOT is run on G_w, an edge $e \in A_w$ is *heavily charged* if it becomes a backward edge, and thus incurs the heavy cost $w(e)$. It is *lightly* charged if it incurs the light cost $\bar{w}(e)$. Clearly, $e = (i, j) \in A_w$ is *heavily charged* if and only if a third vertex k is chosen as pivot when all three i, j, k are in the same recursive call, and (i, j, k) form a directed triangle in G_w. We charge this cost to triangle $t = (i, j, k)$. Again we consider the set T of directed triangles in G_w, and their corresponding events A_t with probability p_t (see Section 3). Fix a triangle $t \in T$ with edges e_1, e_2, e_3. Conditioned on A_t, each of e_1, e_2 and e_3 are equally likely to be heavily charged, so the expected charge of t is $\frac{1}{3}p_t w(t)$. The probability that an edge $e \in A_w$ does not incur a heavy cost (not charged to a triangle $t \in T$) is exactly $1 - \sum_{t:e\in t} \frac{1}{3}p_t$. Therefore, $E[C^{PIV}] = B^{PIV} + F^{PIV}$, where

$$B^{PIV} = \sum_{t\in T} \frac{1}{3} p_t w(t)$$

$$F^{PIV} = \sum_{e\in A_w} \left(1 - \sum_{t:e\in t} \frac{1}{3} p_t\right) \bar{w}(e).$$

We rearrange the sum $C^{OPT} = \sum_{e\in T} c^*(e)$ as $C^{OPT} = B^{OPT} + F^{OPT}$, where

$$B^{OPT} = \sum_{t\in T} \frac{1}{3} p_t c^*(t)$$

$$F^{OPT} = \sum_{e\in A_w} \left(1 - \sum_{t:e\in t} \frac{1}{3} p_t\right) c^*(e) \ .$$

Notice that for all $e \in A_w$, the term $(1 - \sum_{t:e\in t} \frac{1}{3}p_t)$ is nonnegative (see Section 3). Obviously, $F^{PIV} \leq F^{OPT}$, because $\bar{w}(e) \leq c^*(e)$ for any $e \in A_w$. Therefore, if for some $\alpha > 0$, $w(t) \leq \alpha c^*(t)$ for all t, then $E[C^{PIV}] \leq \alpha C^{OPT}$ as required. □

LEMMA 4. *If the weights satisfy the probability constraints ($w_{ij} + w_{ji} = 1$), then $w(t) \leq 5c^*(t)$ for all $t \in T$. If the weights satisfy the triangle inequality constraints ($w_{ij} \leq w_{ik} + w_{kj}$), then $w(t) \leq 3c^*(t)$ for all $t \in T$. If the weights satisfy the combined constraints, then $w(t) \leq 2c^*(t)$ for all $t \in T$.*

PROOF. First assume probability constraints on the weights. In this case, we claim that $w(t) \leq 5c^*(t)$. Indeed, in this case $w(e) \geq 1/2$ for all $e \in A_w$, and $\bar{w}(e) = 1 - w(e)$. Fix a triangle t containing edges e_1, e_2, e_3, and assume

$$1/2 \leq w(e_1) \leq w(e_2) \leq w(e_3) \leq 1 \ . \qquad (2)$$

Clearly, $w(t) = w(e_1) + w(e_2) + w(e_3) \leq 2 + w(e_1)$. Any solution has to direct at least one of the edges in t backwards, therefore $c^*(t) \geq w(e_1)$. Since $w(e_1) \in [1/2, 1]$, we therefore have $w(t) \leq 5c^*(t)$. Consequently, FAS-PIVOT has an expected approximation ratio of at most 5 on weighted tournament instances with probability constraints on the weights. Assume that in addition to the probability constraints, the weights satisfy the triangle inequality. So $w(t) \leq 2$. But the optimal solution has to pay the price of at least one backward edge, so

$$c^*(t) \geq w(e_1) + \bar{w}(e_2) + \bar{w}(e_3) \geq 2w(e_1)$$

(the right inequality follows from the triangle inequality $\bar{w}(e_2) + \bar{w}(e_3) \geq w(e_1)$). Finally, $w(e_1) \geq 1/2$ and therefore $c^*(t) \geq 1$ and $w(t) \leq 2c^*(t)$. Consequently FAS-PIVOT has an expected approximation ratio of at most 2 on weighted tournament instances with both triangle inequality and probability constraints on the weights.

Now we assume that the edge weights satisfy the triangle inequality, but not necessarily the probability constraints. Fix $t \in T$ with edge weights $w(e_1), w(e_2), w(e_3)$. Assume (2) holds. Showing that $c^*(t) \geq w(e_3)$ will prove the lemma for this case. There are 6 possible ways in which the optimal solution can order the vertices of t. For 3 possiblities, e_3 becomes a backward edge, and therefore $c^*(t) \geq w(e_3)$. The other 3 are analyzed case by case. If e_3 and e_2 are forward edges but e_1 is a backward edge, then $c^*(t) = w(e_1)+\bar{w}(e_2)+w(e_3) \geq \bar{w}(e_1)+\bar{w}(e_2)+\bar{w}(e_3)$. But by the triangle inequality, $\bar{w}(e_1)+\bar{w}(e_2) \geq w(e_3)$, so $c^*(t) \geq w(e_3)$, as required. If e_3 and e_1 are forward edges but e_2 is a backward edge, we argue similarly. If e_3 is a forward edge but e_1, e_2 are backward edges, then $c^*(t) = w(e_1) + w(e_2) + \bar{w}(e_3) \geq \bar{w}(e_1) + \bar{w}(e_2) + \bar{w}(e_3)$, which is, again, at least $w(e_3)$ by the triangle inequality. This completes all the cases. □

Combining Theorem 3 and Lemma 4, we get

THEOREM 5. *Running algorithm* FAS-PIVOT *on G_w gives an expected 5, 3 and 2 approximation for the probability constraints case, the triangle inequality constraints case, and the combined constraints case, respectively.*

5. AN IMPROVED APPROXIMATION RATIO FOR RANK AGGREGATION

Let $\{\pi_1, \ldots, \pi_k\}$ be a RANK-AGGREGATION instance over some V. Consider the corresponding equivalent weighted FAS-TOURNAMENT instance (V, w) (where w_{ij} is the fraction of inputs ranking i before j). Clearly, this weight system $\{w_{ij}\}$ is a convex combination of acyclic tournaments. Therefore, by linearity, the edge weights obey the probability constraints and the triangle inequality constraints. Theorem 5 shows that we get a 2 approximation for this case, but the additional structure in these instances allows us to improve upon this factor. As stated in the introduction, there already exists a well known 2-approximation algorithm for RANK-AGGREGATION:

```
PICK-A-PERM({π_1, π_2, ... π_k})
  Output a permutation π_i chosen uniformly at random
  from the input permutations.
```

(In practice, we can pick the permutation π_i that minimizes the cost, but we use the randomized version for the analysis). Let C^{PAP} denote the cost of PICK-A-PERM on the RANK-AGGREGATION instance. Let $G_w = (V, A_w)$ be the corresponding unweighted *majority* tournament. Using notation from Section 4, Let $z(e) = 2w(e)\bar{w}(e)$. We claim that

$$E[C^{PAP}] = \sum_{e\in A_w} z(e) \ . \qquad (3)$$

Indeed, edge $e \in A_w$ becomes a backward (resp. forward) edge with probability $\bar{w}(e)$ (resp. $w(e)$), in which case it incurs the cost of $w(e)$ (resp. $\bar{w}(e)$). For a directed triangle $t = (e_1, e_2, e_3) \in T$, we let $z(t) = z(e_1) + z(e_2) + z(e_3)$. The following theorem shows how to analyze a "convex combination" of FAS-PIVOT and PICK-A-PERM:

THEOREM 6. *If there exist constants $\beta \in [0,1]$ and $\gamma > 0$ such that*

$$\beta w(t) + (1-\beta)z(t) \le \gamma c^*(t) \text{ for all } t \in T, \text{and}$$
$$\beta + 2(1-\beta) \le \gamma,$$

then the best of FAS-PIVOT *and* PICK-A-PERM *is a γ-approximation for* RANK-AGGREGATION.

PROOF. We use the notation $C^{OPT}, F^{OPT}, B^{OPT}, c^*(e)$, $c^*(t)$ defined in Section 4. We rearrange (3) as $E[C^{PAP}] = B^{PAP} + F^{PAP}$, where

$$B^{PAP} = \sum_{t \in T} \frac{1}{3} p_t z(t)$$
$$F^{PAP} = \sum_{e \in A_w} (1 - \sum_{t: e \in t} \frac{1}{3} p_t) z(e) \ .$$

Clearly, $F^{PAP} \le 2F^{OPT}$, because $z(e) \le 2c^*(e)$ for any $e \in A_w$ and $(1 - \sum_{t:e\in t} \frac{1}{3} p_t) \ge 0$.

If we now have β, γ as in the statement of the theorem, then

$$\begin{aligned}
&\beta E[C^{PIV}] + (1-\beta)E[C^{PAP}] \\
&\quad = \beta B^{PIV} + (1-\beta)B^{PAP} + \beta F^{PIV} + (1-\beta)F^{PAP} \\
&\quad = \sum_{t \in T} \frac{1}{3} p_t (\beta w(t) + (1-\beta) z(t)) \\
&\qquad\quad + \sum_{e \in A_w} \left(1 - \sum_{t:e\in t} \frac{1}{3} p_t\right) (\beta \bar{w}(e) + (1-\beta) z(e)) \\
&\quad \le \sum_{t \in T} \frac{1}{3} p_t (\beta w(t) + (1-\beta) z(t)) \\
&\qquad\quad + \sum_{e \in A_w} \left(1 - \sum_{t:e\in t} \frac{1}{3} p_t\right) (\beta c^*(e) + (1-\beta) 2c^*(e)) \\
&\quad \le \sum_{t \in T} \frac{1}{3} p_t \gamma c^*(t) + \sum_{e \in A_w} \left(1 - \sum_{t:e\in t} \frac{1}{3} p_t\right) \gamma c^*(e) \\
&\quad = \gamma C^{OPT} \ ,
\end{aligned}$$

as required. □

LEMMA 7. *For all $t \in T$,*

$$\frac{3}{7} w(t) + \frac{4}{7} z(t) \le \frac{11}{7} c^*(t) \ .$$

PROOF. We want to show that

$$f(t) = \frac{3}{7} w(t) + \frac{4}{7} z(t) - \frac{11}{7} c^*(t) \le 0,$$

where (slightly changing notation) $t = (w_1, w_2, w_3)$ and

$$\begin{aligned}
w(t) &= w_1 + w_2 + w_3 \\
z(t) &= 2w_1(1-w_1) + 2w_2(1-w_2) + 2w_3(1-w_3) \\
c^*(t) &= 1 - w_2 + 1 - w_3 + w_1 \\
1/2 &\le w_1 \le w_j \le 1 \text{ for } j = 2,3 \\
w_1 + w_2 + w_3 &\le 2
\end{aligned}$$

The proof can be completed by finding the global maximum of $f(t)$ on the defined polytope using standard techniques of multivariate calculus. We omit the details from this version. □

Theorem 8 follows from Theorem 6 and Lemma 7, using $\beta = 3/7$ and $\gamma = 11/7$:

THEOREM 8. *The best of* FAS-PIVOT *on G_w and* PICK-A-PERM *is an expected* 11/7 *approximation for* RANK-AGGREGATION.

6. CORRELATION CLUSTERING AND CONSENSUS CLUSTERING

In this section, we show how to apply the techniques presented in Section 3 to CORRELATION-CLUSTERING and CONSENSUS-CLUSTERING. Disagreements in the output solution can also be charged to *bad triplets*, which will be defined shortly. The bad triplets replace the role taken by the directed triangles in tournaments. Let (V, E^+, E^-) be a CORRELATION-CLUSTERING instance. Our algorithm CC-PIVOT, which is an analog of FAS-PIVOT, is defined as follows:

```
CC-Pivot(G = (V, E+, E-))

  Pick random pivot i ∈ V.
  Set C = {i}, V' = ∅.

  For all j ∈ V, j ≠ i:
      If (i,j) ∈ E+ then
          Add j to C
      Else (If (i,j) ∈ E-)
          Add j to V'

  Let G' be the subgraph induced by V'.

  Return clustering C, CC-Pivot(G') .
```

As in the analysis of FAS-PIVOT, a pair i, j incurs a unit cost if a third vertex k is chosen as pivot when the triplet (i, j, k) is in the same recursive call, and there are two $(+)$ and one $(-)$ relations among i, j, k (doesn't matter in which order). A triplet (i, j, k) is therefore a *bad triplet* if it has two $(+)$ and one $(-)$ relations[6]. Let T denote the set of (not necessarily disjoint) bad triplets. For each $t = (i, j, k) \in T$ we define A_t as the event that *all three i, j, k are in the same recursive call when the first one among them was chosen as pivot*. Let p_t denote the probability of A_t. The analysis continues identically to that of FAS-PIVOT.

THEOREM 9. *Algorithm* CC-PIVOT *is a randomized expected* 3*-approximation algorithm for* CORRELATION-CLUSTERING.

Now let (V, w^+, w^-) be a weighted CORRELATION-CLUSTERING instance. Unlike weighted FAS-TOURNAMENT, we will *only* consider weight systems that satisfy the probability constraints $w^+_{ij} + w^-_{ij} = 1$. We create the unweighted majority CORRELATION-CLUSTERING instance $G_w = (V, E^+_w, E^-_w)$

[6]A CORRELATION-CLUSTERING instance with no bad triplets induces a consistent clustering, just as a tournament with no 3-cycles is acyclic. Our algorithms have an optimal cost of 0 on these instances.

and return the clustering generated by CC-Pivot(G_w). Using the same analysis as in Section 4, we can show that this algorithm gives an expected 5 approximation.

Triangle inequality constraints in weighted Correlation-Clustering have the following form: for all i, j, k, $w^+_{ij} + w^+_{jk} + w^-_{ik} \leq 2$. (Equivalently, $w^-_{ik} \leq w^-_{ij} + w^-_{jk}$.) Theorem 10 is analogous to Theorem 5:

Theorem 10. *Algorithm* CC-Pivot *on G_w is a 5 (resp. 2) approximation for weighted* Correlation-Clustering *with probability constraints (resp. with probability and triangle inequality constraints).*

The proof is almost identical to that of Theorem 5, with "$++-$" (bad) triplets in G_w replacing the role of directed (bad) triangles in tournaments.

Solving Consensus-Clustering is equivalent to solving weighted Correlation-Clustering with w^+_{ij} (resp. w^-_{ij}) as the fractional number of input clusters with a $(+)$ (resp. $(-)$) relation between i and j. This weighted Correlation-Clustering instance obeys both the probability constraints and the triangle inequality constraints, but we can do better than the 2 approximation guaranteed by Theorem 10. Analysis almost identical to the one in Section 5 gives an expected 11/7 approximation for this case. The CC-Pivot is coupled with Pick-A-Cluster, which is defined analogously to Pick-A-Perm: Simply return a cluster chosen uniformly at random from the list.

Theorem 11. *The best of* CC-Pivot *on G_w and* Pick-A-Cluster *has an expected approximation ratio of at most $\frac{11}{7}$ for* Consensus-Clustering.

7. USING THE PIVOT SCHEME FOR ROUNDING THE LP

We show how the techniques introduced above can be used for rounding the LP's for Fas-Tournament and Correlation-Clustering. We consider the LP's given in Figure 1 [27, 10]. Given a solution to the LP, we consider algorithms FasLP-Pivot and CCLP-Pivot (Figure 1) for rounding the solutions for Fas-Tournament and Correlation-Clustering, respectively. The main idea of these algorithms is that, after we choose some pivot, we use the LP solution variables to randomly decide where to put all other vertices, instead of deciding greedily.

Theorem 12. FasLP-Pivot *returns a ranking with an expected cost of at most* 2.5 *(resp.* 2*) times the LP solution for* Fas-Tournament*, when the weights satisfy the probability constraints (resp. the probability constraints and the triangle inequality constraints). The best of* FasLP-Pivot *and* Pick-A-Perm *returns a ranking with an expected cost of at most* 4/3 *times the LP solution for* Rank-Aggregation. CCLP-Pivot *returns a clustering with an expected cost of at most* 2.5 *(resp.* 2*) times the LP solution for* Correlation-Clustering*, when the weights satisfy the probability constraints (resp. the probability constraints and the triangle inequality constraints). The best of* CCLP-Pivot *and* Pick-A-Cluster *returns a ranking with an expected cost of at most* 4/3 *times the LP solution for* Consensus-Clustering.

Note that these bounds imply bounds on the integrality gaps of the LP relaxation for the different cases.

Proof. We prove these bounds by reducing the problem to proving global bounds of certain multinomials in high dimensional polytopes. We start with the analysis of FasLP-Pivot. A similar analysis is done for CCLP-Pivot.

Let C^{PIV}_{LP} denote the cost of the ordering returned by the rounding algorithm FasLP-Pivot. We have the notion of pairs i, j that are charged *dangerously* and *safely*. The safe edges are charged when one of their endpoints is chosen as pivot, and the other endpoint is in the same recursive call. The expected cost of pairs that are charged safely in FasLP-Pivot is

$$x_{ij}w_{ji} + x_{ji}w_{ij} \ , \tag{4}$$

which is exactly the contribution to the LP solution. We let c^*_{ij} denote expression (4). So the value of the LP solution is $C_{LP} = \sum_{i<j} c^*_{ij}$.

A pair i, j is charged dangerously when a third vertex k is chosen as pivot, all three i, j, k are in the same recursive call, and i, j are placed on opposite sides of k. The charge is w_{ij} (resp. w_{ji}) if j (resp. i) is placed on the left side of k and i (resp. j) on its right. In either case, we charge this cost to the triplet i, j, k. We let T denote the set of all triplets of distinct vertices, and for any $t = \{i, j, k\} \in T$ we denote by A_t the event that all of i, j, k are in the same recursive call when the first one among them is chosen as pivot. Let p_t denote the probability of A_t. Let B^t_{ij} denote the event that (i, j) is dangerously charged to triangle t, in that order (i to the left, j to the right). Then we have for any $t = \{i, j, k\}$,

$$\Pr[A_t \wedge B^t_{ij}] = \Pr[A_t] \Pr[B^t_{ij} | A_t] = \frac{1}{3} p_t x_{ik} x_{kj} \ .$$

(The 1/3 comes from the fact that conditioned on A_t, each one of i, j, k was equally likely to be the pivot vertex.) Denote $p^t_{ij} = \frac{1}{3} x_{ik} x_{kj}$. So the total expected charge to a triplet $t = \{i, j, k\}$ is $p_t y(t)$, where

$$y(t) = p^t_{ij} w_{ji} + p^t_{ji} w_{ij} + p^t_{jk} w_{kj} + p^t_{kj} w_{jk} + p^t_{ki} w_{ik} + p^t_{ik} w_{ki}.$$

Now we notice that for any $t = \{i, j, k\}$ and $t' = \{i, j, k'\}$ (two triplets sharing a pair i, j), the events $A_t \wedge (B^t_{ij} \vee B^t_{ji})$ and $A_{t'} \wedge (B^{t'}_{ij} \vee B^{t'}_{ji})$ are disjoint, because a pair i, j can be split into two different recursion branches only once. Thus,

$$\sum_{t : i,j \in t} p_t (p^t_{ij} + p^t_{ji}) \leq 1 \ .$$

The above expression is exactly the probability that the pair i, j is dangerously charged. Therefore, the total expected cost of FasLP-Pivot is $E[C^{PIV}_{LP}] = B^{PIV}_{LP} + F^{PIV}_{LP}$, where

$$B^{PIV}_{LP} = \sum_t p_t y(t)$$

$$F^{PIV}_{LP} = \sum_{i<j} \left(1 - \sum_{t : i,j \in t} p_t (p^t_{ij} + p^t_{ji}) \right) c^*_{ij} \ .$$

The following expression is a rearrangement of the sum $C_{LP} = \sum_{i<j} c^*_{ij}$: $C_{LP} = B_{LP} + F_{LP}$, where

$$B_{LP} = \sum_t p_t \sum_{\{i,j\} \subseteq t} (p^t_{ij} + p^t_{ji}) c^*_{ij}$$

$$F_{LP} = \sum_{i<j} \left(1 - \sum_{t : i,j \in t} p_t (p^t_{ij} + p^t_{ji}) \right) c^*_{ij} \ .$$

LP for weighted Fas-Tournament

$$\text{minimize } \sum_{i<j}(x_{ij}w_{ji} + x_{ji}w_{ij}) \text{ s.t.}$$
$$x_{ik} \le x_{ij} + x_{jk} \text{ for all distinct } i,j,k$$
$$x_{ij} + x_{ji} = 1 \text{ for all } i \ne j$$
$$x_{ij} \ge 0 \text{ for all } i \ne j$$

LP for weighted Correlation-Clustering

$$\text{minimize } \sum_{i<j}(x^+_{ij}w^-_{ji} + x^-_{ji}w^+_{ij}) \text{ s.t.}$$
$$x^-_{ik} \le x^-_{ij} + x^-_{jk} \text{ for all distinct } i,j,k$$
$$x^+_{ij} + x^-_{ij} = 1 \text{ for all } i \ne j$$
$$x^-_{ij}, x^+_{ij} \ge 0 \text{ for all } i \ne j$$

FasLP-Pivot(V, x)
A recursive algorithm for rounding the LP for weighted Fas-Tournament. *Given an LP solution $x = \{x_{ij}\}_{i,j\in V}$, returns an ordering on the vertices.*

```
Set V_R = ∅, V_L = ∅.
Pick random pivot i ∈ V.

For all j ∈ V, j ≠ i:
    With probability x_ji
        Add j to V_L.
    Else (With remaining probability x_ij = 1 − x_ji)
        Add j to V_R.

Return order FasLP-Pivot(V_L, x), i, FasLP-Pivot(V_R, x)
```
(concatenation of left recursion, i, and right recursion)

CCLP-Pivot(V, x^+, x^-)
A recursive algorithm for rounding the LP for weighted Correlation-Clustering. *Given an LP solution $x^+ = \{x^+_{ij}\}_{i<j}$, $x^- = \{x^-_{ij}\}_{i<j}$, returns a clustering of the vertices.*

```
Set C_i = ∅, V' = ∅.
Pick random pivot i ∈ V.

For all j ∈ V, j ≠ i :
    With probability x^+_ij
        Add j to C_i.
    Else (With remaining probability x^-_ij = 1 − x^+_ij)
        Add j to V'.

Return clustering C_i, CCLP-Pivot(V', x^+, x^-).
```

Figure 1: Standard LP relaxations and their corresponding rounding algorithms.

So $F_{LP} = F^{PIV}_{LP}$. We have the following lemma, the proof of which can be found in [1]:

Lemma 13. *If the weight system satisfies the probability constraints (resp. probability constraints and triangle inequality constraints), then for any $t \in T$,*

$$y(t) \le \tau \sum_{\{i,j\}\subseteq t} (p^t_{ij} + p^t_{ji})c^*_{ij},$$

where $\tau = 5/2$ (resp. $\tau = 2$).
Therefore, in this case, $B^{PIV}_{LP} \le \tau B_{LP}$.

Although this just gives yet another 2 approximation algorithm for the rank aggregation problem, we can do better there. We couple FasLP-Piv with Pick-A-Perm. The expected value of the Pick-A-Perm algorithm is

$$E[C^{PAP}] = \sum_{i<j} z_{ij},$$

where $z_{ij} = 2w_{ij}(1-w_{ij})$. We rearrange this sum as follows:

$$E[C^{PAP}] = B^{PAP}_{LP} + F^{PAP}_{LP},$$

where

$$B^{PAP}_{LP} = \sum_t p_t \sum_{\{i,j\}\subseteq t} (p^t_{ij} + p^t_{ji})z_{ij}$$

$$F^{PAP}_{LP} = \sum_{i<j} \left(1 - \sum_{t:\{i,j\}\subseteq t} p_t(p^t_{ij} + p^t_{ji})\right) z_{ij} \ .$$

It is easy to see that $F^{PAP}_{LP} \le 2F_{LP}$ (this is because $z_{ij} \le 2c^*_{ij}$, and $\sum_{t:i,j\in t} p_t(p^t_{ij} + p^t_{ji}) \le 1$). We have the following lemma, (proof in [1]):

Lemma 14. *For all $t = \{i,j,k\}$,*

$$\frac{2}{3}y(t) + \frac{1}{3}\sum_{\{i,j\}\subseteq t}(p^t_{ij} + p^t_{ji})z_{ij} \le \frac{4}{3}\sum_{\{i,j\}\subseteq t}(p^t_{ij} + p^t_{ji})c^*_{ij}.$$

As a consequence, $\frac{2}{3}B^{PIV}_{LP} + \frac{1}{3}B^{PAP}_{LP} \le \frac{4}{3}B_{LP}$.

Clearly we have $\frac{2}{3}F^{PIV}_{LP} + \frac{1}{3}F^{PAP}_{LP} \le \frac{4}{3}F_{LP}$, and we conclude that the minimum between FasLP-Pivot and Pick-A-Perm has an expected approximation ratio of at most $\frac{4}{3}$ with respect to the LP cost.

We now consider the analysis of CCLP-Pivot on Correlation-Clustering and Consensus-Clustering. Define $c^*_{ij} = x^+_{ij}w^-_{ij} + x^-_{ij}w^+_{ij}$ (the LP contribution as well as the expected charge of the safe pairs, which are defined as above).

For a triplet $t = (i,j,k)$, let $B^t_{\{i\}j}$ denote the event that i,j was dangerously charged to t, because k is the pivot, i is taken in k's cluster and j is placed aside (the charge is w^+_{ij}). The probability of $B^t_{\{i\}j}$ conditioned on A_t is $p^t_{\{i\}j} = \frac{1}{3}x^+_{ki}x^-_{kj}$. Let $B^t_{\{ij\}}$ denote the event that i,j was dangerously charged to t, because k is the pivot, and both i and j are taken in k's cluster (the charge is w^-_{ij}). The probability of $B^t_{\{ij\}}$ conditioned on A_t is $p^t_{\{ij\}} = \frac{1}{3}x^+_{ki}x^+_{kj}$.

Define $y(t) = \sum_{\{i,j\}\subseteq t}(p^t_{\{i\}j} + p^t_{\{j\}i})w^+_{ij} + p^t_{\{ij\}}w^-_{ij}$.

For all $i \ne j$, $\sum_{t:\{i,j\}\subseteq t} p_t(p^t_{\{i\}j} + p^t_{\{j\}i} + p^t_{\{ij\}}) \le 1$ (disjointness of events). As before, we decompose $E[C^{PIV}_{LP}] = B^{PIV}_{LP} + F^{PIV}_{LP}$, $C_{LP} = F_{LP} + B_{LP}$, where

$$B_{LP}^{PIV} = \sum_t p_t y(t)$$

$$F_{LP}^{PIV} = \sum_{i<j} \left(1 - \sum_{t:\{i,j\}\subseteq t} p_t(p^t_{\{i\}j} + p^t_{\{j\}i} + p^t_{\{ij\}})\right) c^*_{ij}.$$

$$B_{LP} = \sum_t p_t \sum_{\{i,j\}\subseteq t} (p^t_{\{i\}j} + p^t_{\{j\}i} + p^t_{\{ij\}}) c^*_{ij}$$

$$F_{LP} = \sum_{i<j} \left(1 - \sum_{t:\{i,j\}\subseteq t} p_t(p^t_{\{i\}j} + p^t_{\{j\}i} + p^t_{\{ij\}})\right) c^*_{ij} \ .$$

LEMMA 15. *If the weight system satisfies the probability constraints (resp. probability constraints and triangle inequality constraints), then for any $t \in T$,*

$$y(t) \le \tau \sum_{\{i,j\}\subseteq t} (p^t_{\{i\}j} + p^t_{\{j\}i} + p^t_{\{ji\}}) c^*_{ij} \ ,$$

where $\tau = 5/2$ (resp. $\tau = 2$).

As a result, we get a 5/2 approximation for the probability constraints case, and a 2 approximation for the probability and triangle inequality constraints case.

For Consensus clustering: Let C_{LP}^{PAC} denote the value of PICK-A-CLUSTER. So $E[C_{LP}^{PAC}] = B_{LP}^{PAC} + F_{LP}^{PAC}$, where

$$B_{LP}^{PAC} = \sum_t p_t \sum_{\{i,j\}\subseteq t} (p^t_{\{i\}j} + p^t_{\{j\}i} + p^t_{\{ij\}}) z_{ij}$$

$$F_{LP}^{PAC} = \sum_{i<j} \left(1 - \sum_{t:\{i,j\}\subseteq t} p_t(p^t_{\{i\}j} + p^t_{\{j\}i} + p^t_{\{ij\}})\right) z_{ij} \ge 0.$$

$$z_{ij} = 2w^+_{ij}w^-_{ij}$$

LEMMA 16. *For all $t = \{i,j,k\}$,*

$$\frac{2}{3}y(t) + \frac{1}{3}\sum_{\{i,j\}\subseteq t} (p^t_{\{i\}j} + p^t_{\{j\}i} + p^t_{\{ij\}}) z_{ij} \le \frac{4}{3}\sum_{\{i,j\}\subseteq t} (p^t_{\{i\}j} + p^t_{\{j\}i} + p^t_{\{ij\}}) c^*_{ij} \ .$$

Therefore, we get a 4/3 approximation algorithm for consensus clustering. The proofs of Lemmas 15 and 16 can be found in [1]. This completes the proof of Theorem 12. □

8. NP-HARDNESS OF FEEDBACK ARC SET ON TOURNAMENTS

All the problems referred to in Table 1 in Section 1.2 were previously known to be NP-hard except for FAS-TOURNAMENT. In this section we show:

THEOREM 17. *Unless NP⊆BPP,* FAS-TOURNAMENT *has no polynomial time algorithm.*

PROOF. We reduce to FAS-TOURNAMENT from FAS-DIGRAPH, which is the problem of finding a minimum feedback arc set in a general directed graph. FAS-DIGRAPH is NP-hard [22] (in fact, it is MAX-SNP-hard, see [21, 25, 26]).

Let $G = (V, A)$ (with $|V| = n$) be an instance of FAS-DIGRAPH. Suppose we could add a set of edges A_R to G such that $(V, A \cup A_R)$ is a tournament, and such that exactly half of A_R are backward in any ordering π of V. Then by solving FAS-TOURNAMENT we would be able to recover the feedback arc set of G. This is generally impossible. However, if we add the edges A_R randomly (i.e. for every i, j such the neither (i,j) nor (j,i) are in A add (i,j) or (j,i) to A_R with equal probability) then for any π the expected number of backward edges is half $|R|$. The variance makes this approach fail. By blowing up G and using a concentration property of the random variable counting the number of backward edges in A_R, we can use this construction (see similar random digraph constructions in [25, 26]).

We pick an integer $k = \text{poly}(n)$ (chosen later). The blow-up digraph $G^k = (V^k, A^k)$ is defined as follows:

$$V^k = \bigcup_{v\in V} \{v_1, \ldots, v_k\}$$

$$A^k = \{(u_i, v_j) | (u,v) \in A, i, j \in \{1, \ldots, k\}\} \ .$$

We observe that the minimum feedback arc set of G^k is exactly k^2 times the minimum feedback arc set of G. Indeed, it suffices to consider only rankings π on V^k that rank the vertices $v_1, \ldots, v_k$ as one block for all $v \in V$ (as explained in [2], if $v_i <_\pi v_j$ are not adjacent in the ranking, then either moving v_i immediately to the left of v_j or moving v_j immediately to the right of v_i will result in a ranking inducing no more feedback edges than π).

Now we turn G^k into a tournament $T^k = \{V^k, A^k \cup A^k_R\}$ using the construction defined above. For a ranking π of V^k, let $f_R(\pi)$ denote the number of feedback edges in A^k_R with respect to π. Denote by μ the expected value of $f_R(\pi)$, which is the same for all π, and can be efficiently computed. We claim that for $k = \text{poly}(n)$, with probability at least 2/3, all rankings π satisfy $|f_R(\pi) - \mu| = O((nk)^{3/2}\sqrt{\log(nk)})$. This would imply, using the above observation, that for big enough $k = \text{poly}(n)$ the size of the minimum feedback arc set of T^k can be used to efficiently recover the size of the minimum feedback arc set of G, because $(nk)^{3/2}\sqrt{\log(nk)} = o(k^2)$. To prove the claim, for any fixed ranking π, set a random indicator variable X^π_{wz} for every non-edge $\{w, z\}$ of G^k which equals 1 iff the edge between w and z in A^k_R is backward w.r.t. π. So $f_R(\pi) = \sum X^\pi_{wz}$. A simple application of Chernoff bounds [3] and union bound (over all possible $(nk)!$ rankings) completes the proof of the claim. It follows that unless FAS-DIGRAPH $\in BPP$, we cannot solve FAS-TOURNAMENT in polynomial time. □

We wish to thank Noga Alon for ideas significantly simplifying the proof [2]. Our initial hardness result was via max-SNP hardness of FAS-DIGRAPH, and Noga Alon pointed out that the same idea also works with the weaker NP-hardness.

9. OPEN PROBLEMS AND FUTURE WORK

We propose the following directions of research.

- FAS-PIVOT can be thought of as a "quicksort" heuristic for FAS-TOURNAMENT. Can we use other heuristics, such as mergesort?
- Can FAS-PIVOT and CC-PIVOT and their LP rounding analogues be derandomized?
- Is RANK-AGGREGATION NP-Hard for 3 permutations [14, 15]?
- Is CONSENSUS-CLUSTERING NP-Hard for a constant number of clusters [31, 18]?

- Can we approximate weighted CORRELATION-CLUSTERING with triangle inequalities, but no probability constraints?

10. ACKNOWLEDGEMENTS

We would like to thank Ravi Kumar and D. Sivakumar for several discussions on these problems. Thanks also to Shuchi Chawla and Tony Wirth for extensive discussions on consensus clustering, to Aristides Gionis for sending us a preprint of their paper [20], and to Noga Alon for discussions on the hardness result.

11. REFERENCES

[1] N. Ailon, M. Charikar, and A. Newman. Proofs of conjectures in 'aggregating inconsistent information: Ranking and clustering'. Technical Report TR-719-05, Princeton University, 2005.

[2] N. Alon. Ranking tournaments (draft). *Personal communication*, 2004.

[3] N. Alon and J. H. Spencer. *The Probabilistic Method*. Wiley, New York, 1992.

[4] S. Arora, A. Frieze, and H. Kaplan. A new rounding procedure for the assignment problem with applications to dense graph arrangement problems. In *Proceedings of the 37th Annual Symposium on the Foundations of Computer Science (FOCS)*, pages 24–33, Burlington, VT, 1996.

[5] J. Bang-Jensen and C. Thomassen. A polynomial algorithm for the 2-path problem in semicomplete graphs. *SIAM Journal of Discrete Mathematics*, 5(3):366–376, 1992.

[6] N. Bansal, A. Blum, and S. Chawla. Correlation clustering. *Machine Learning Journal (Special Issue on Theoretical Advances in Data Clustering)*, 56(1–3):89–113, 2004. Extended abstract appeared in FOCS 2002, pages 238–247.

[7] J. Bartholdi, C. A. Tovey, and M. Trick. Voting schemes for which it can be difficult to tell who won the election. *Social Choice and Welfare*, 6(2):157–165, 1989.

[8] J. C. Borda. Mémoire sur les élections au scrutin. *Histoire de l'Académie Royale des Sciences*, 1781.

[9] M.-C. Cai, X. Deng, and W. Zang. An approximation algorithm for feedback vertex sets in tournaments. *SIAM Journal on Computing*, 30(6):1993–2007, 2001.

[10] M. Charikar, V. Guruswami, and A. Wirth. Clustering with qualitative information. In *Proceedings of the 44th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 524–533, Boston, 2003.

[11] M.-J. Condorcet. Éssai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix. 1785.

[12] P. Diaconis and R. Graham. Spearman's footrule as a measure of disarray. *Journal of the Royal Statistical Society, Series B*, 39(2):262–268, 1977.

[13] I. Dinur and S. Safra. On the importance of being biased. In *Proceedings of the 34th Annual Symposium on the Theory of Compututing (STOC)*, pages 33–42, 2002.

[14] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar. Rank aggregation methods for the web. In *Proceedings of the Tenth International Conference on the World Wide Web (WWW10)*, pages 613–622, Hong Kong, 2001.

[15] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar. Rank aggregation revisited. 2001. Manuscript.

[16] G. Even, J. S. Naor, M. Sudan, and B. Schieber. Approximating minimum feedback sets and multicuts in directed graphs. *Algorithmica*, 20(2):151–174, 1998.

[17] R. Fagin, R. Kumar, and D. Sivakumar. Efficient similarity search and classification via rank aggregation. In *Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data*, pages 301–312, San Diego, 2003.

[18] V. Filkov and S. Skiena. Integrating microarray data by consensus clustering. In *Proceedings of International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 418–425, Sacramento, 2003.

[19] A. Frieze and R. Kannan. Quick approximations to matrices and applications. *Combinatorica*, 19(2):175–220, 1999.

[20] A. Gionis, H. Mannila, and P. Tsaparas. Clustering aggregation. In *Proceedings of the 21st International Conference on Data Engineering (ICDE)*, Tokyo, 2005. To appear.

[21] J. Håstad. Some optimal inapproximability results. *Journal of the ACM*, 48:798–859, 2001.

[22] R. M. Karp. Reducibility among combinatorial problems. In *Complexity of Computer Computations*, pages 85–104. Plenum Press, New York, 1972.

[23] J. Kemeny and J. Snell. *Mathematical Models in the Social Sciences*. Blaisdell, New York, 1962. Reprinted by MIT Press, Cambridge, 1972.

[24] J. G. Kemeny. Mathematics without numbers. *Daedalus*, 88:571–591, 1959.

[25] A. Newman. Approximating the maximum acyclic subgraph. Master's thesis, Massachusetts Institute of Technology, Cambridge, MA, June 2000.

[26] A. Newman and S. Vempala. Fences are futile: On relaxations for the linear ordering problem. In *Proceedings of the Eighth Conference on Integer Programming and Combinatorial Optimization (IPCO)*, pages 333–347, 2001.

[27] C. N. Potts. An algorithm for the single machine sequencing problem with precedence constraints. *Mathematical Programming*, 13:78–87, 1980.

[28] P. Seymour. Packing directed circuits fractionally. *Combinatorica*, 15:281–288, 1995.

[29] E. Speckenmeyer. On feedback problems in digraphs. *Graph Theoretic Concepts in Computer Science, Lecture Notes in Computer Science*, 411:218–231, 1989.

[30] A. Strehl. *Relationship-based Clustering and Cluster Ensembles for High-dimensional Data Mining.* PhD dissertation, University of Texas at Austin, May 2002.

[31] Y. Wakabayashi. The complexity of computing medians of relations. *Resenhas*, 3(3):323–349, 1998.