On Strategyproof Conference Peer Review

Yichong Xu[1*], **Han Zhao**[1*], **Xiaofei Shi**[2] and **Nihar B. Shah**[1]

[1]Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA, USA
[2]Department of Mathematical Sciences, Carnegie Mellon University, Pittsburgh, PA, USA
{yichongx, han.zhao, nihars}@cs.cmu.edu, xiaofeis@andrew.cmu.edu

Abstract

We consider peer review under a conference setting where there are conflicts between the reviewers and the submissions. Under such conflicts, reviewers can manipulate their reviews in a strategic manner to influence the final rankings of their own papers. Present-day peer-review systems are not designed to guard against such strategic behavior, beyond minimal (and insufficient) checks such as not assigning a paper to a conflicted reviewer. In this work, we address this problem through the lens of social choice, and present a theoretical framework for strategyproof and efficient peer review. Given the conflict graph which satisfies a simple property, we first present and analyze a flexible framework for reviewer-assignment and aggregation for the reviews that guarantees not only strategyproofness but also a natural efficiency property (unanimity). Our framework is based on the so-called partitioning method, and can be treated as a generalization of this type of method to conference peer review settings. We then empirically show that the requisite property on the (authorship) conflict graph is indeed satisfied in the ICLR-17 submissions data, and further demonstrate a simple trick to make the partitioning method more practically appealing under conference peer-review settings. Finally, we complement our positive results with negative theoretical results where we prove that under slightly stronger requirements, it is impossible for any algorithm to be both strategyproof and efficient.

1 Introduction

Peer review serves as an effective solution for quality evaluation in reviewing processes, especially in academic paper review [Dörfler *et al.*, 2017, Shah *et al.*, 2017] and massive open online courses (MOOCs) [Díez Peláez *et al.*, 2013, Piech *et al.*, 2013, Shah *et al.*, 2013]. However, despite its scalability, competitive peer review faces the serious challenge of being vulnerable to strategic manipulations [Anderson *et al.*, 2007, Thurner and Hanel, 2011, Alon *et al.*, 2011, Kurokawa *et al.*, 2015, Kahng *et al.*, 2018]. By manipulating the ratings or rankings provided reviewers may be able to increase the chance that their own submissions get accepted. As noted by Thurner *et al.* [2011], even a small number of selfish, strategic reviewers can drastically reduce the quality of scientific standard. Thus the importance of peer review in academia and its considerable influence over the careers of researchers significantly underscores the need to design peer review systems that are insulated from strategic manipulations. Present-day peer-review systems are, however, ill-equipped to handle strategic behavior, and do not have any rigorous framework in place except some basic checks like not assigning a paper to a conflicted reviewer. It is easy to show that these checks are insufficient, that is, reviewers can manipulate the final ranking of their conflicted papers by strategically manipulating their reports in current peer-review systems.

In this work, we present a framework for conference peer review that addresses the problem of strategic behavior. Our problem setup comprises a set of submitted papers and a set of reviewers. We are given a graph which we term as "conflict graph". The conflict graph is a bipartite graph with the reviewers and papers as the two partitions of vertices, and an edge if there exists a conflict between the reviewer and the paper. Conflicts may arise due to authorship or other reasons such as institutional conflicts. Given this, there are two design steps in the peer review procedure: (i) assigning each paper to a subset of reviewers, and (ii) aggregating the review results provided by the reviewers to give a final evaluation of each paper.

We focus on ordinal preferences where each reviewer is asked to give a total ranking of the assigned papers, as ordinal data avoids biases and miscalibrations and provides a more direct comparison between papers [Barnett, 2003, Stewart *et al.*, 2005, Douceur, 2009, Tsukida and Gupta, 2011, Shah *et al.*, 2013, Shah *et al.*, 2016]. Some modern large conferences (e.g., NeurIPS [Shah *et al.*, 2017]) have also collected ordinal preferences for experimental purposes. Also, our methods can be extended easily to a score-based review system (cf. Section 4.1). Under the setting, we require our mechanism to output a total ranking over all papers, since this automated output in practice would be a guideline for program chairs to make their decisions (e.g., orals, posters, best papers, etc.).

We consider two important goals for designing a good conference review procedure – strategyproofness and efficiency. In our context, strategyproofness means that no reviewer can change the outcome of any papers which she/he has a conflict with, and efficiency means that the final output of the procedure reflects the opinion of most reviewers. Here, we consider

*Equal contribution.

efficiency in terms of the notion of unanimity in social choice theory: an agreement among all reviewers must be reflected in the final aggregation.

In addition to the conceptual contributions, we make several technical contributions towards this important problem. We first design a peer review framework which theoretically guarantees strategyproofness along with a notion of efficiency that we term as "group unanimity". With only a mild and realistic assumption on the conflict graph, we establish our positive results for the peer review design task. Importantly, our framework and the specific algorithms are quite flexible in that it guarantees strategyproofness and unanimity, while also leaving open a significant room for program chairs to implement their choice of decision-making strategies.

On the practical front, we validate that the aforementioned assumption indeed holds in practice via an empirical analysis of the submissions made to the International Conference on Learning Representations 2017 (ICLR-17). Furthermore, we demonstrate a simple trick to make the partitioning method more practically appealing for conference peer review and present it on the ICLR-17 data.

We also complement our positive results with negative results showing that one cannot expect to meet requirements that are stronger than that provided by our framework. First, we show that under mild assumptions on the conflict graph, there is no algorithm that satisfies "pairwise unanimity", which is a stronger notion of efficiency than group unanimity, and is also known as Pareto efficiency in the literature of social choice [Brandt *et al.*, 2016]. Second, we provide a conjecture and insightful results to show that if we require the assignment to satisfy a simple "connectivity" condition, our negative result continues to hold even when the notion of strategyproofness is made extremely weak. These negative results highlight the intrinsic hardness in designing strategyproof conference review systems.

2 Related Work

As early as in the 1970s, Gibbard and Satterthwaite had already been aware of the importance of a healthy voting rule that guarantees strategyproofness under the setting of social choice [Gibbard, 1973, Satterthwaite, 1975]. Recently, the fact that prominent peer review mechanisms are manipulable, has further called for strategyproof peer review mechanisms [Merrifield and Saari, 2009, Hazelrigg, 2013]. Our work is most closely related to a series of works on strategyproof peer selection [Alon *et al.*, 2011, Holzman and Moulin, 2013, Fischer and Klimm, 2015, Kurokawa *et al.*, 2015, Aziz *et al.*, 2016, Kahng *et al.*, 2018], where agents cannot benefit from misreporting their preferences over other agents. In all these works, each agent is essentially required to evaluate *all the other agents* except herself. This is impractical for conference peer review, where each reviewer only has to review a small subset of submissions. In light of such constraints, Kurokawa *et al.* [2015] propose a strategyproof (impartial) mechanism and provide associated approximation guarantees in which each agent is only required to review a few other agents, but their algorithm might return an empty set. Based on their work, Aziz *et al.* [2016] then propose an improved mechanism for peer selection which is strategyproof and satisfies a natural monotonicity property. However, even if the target output size is k, it may return a subset of size strictly larger than k. Past works focus on applications of peer-grading and grant proposal review, and hence consider only the case where every reviewer is conflicted with exactly one paper. In contrast, our setting of conference peer review is more challenging: We allow each reviewer to have conflicts with multiple papers and each paper to have conflicts with multiple reviewers. Formally, the conflict graph $\mathcal{C}$ under conference peer review is a more general bipartite graph, where conflicts between reviewers and papers can arise not only because of authorships, but also advisor-advisee relationships, institutional conflicts, etc.

3 Problem Setting

We define $\mathcal{R} := \{r_1, \ldots, r_m\}$ to be the set of m reviewers and $\mathcal{P} := \{p_1, \ldots, p_n\}$ to be the set of n submitted papers. To characterize conflicts of interest, we use a bipartite graph $\mathcal{C}$ with vertices $(\mathcal{R}, \mathcal{P})$, where an edge is connected between a reviewer r and a paper p if there exists some conflict of interests between the r and p. Given the set of submitted papers and reviewers, the conflict graph is fixed and cannot be controlled. Note that the conflict graph $\mathcal{C}$ defined above can be viewed as a realistic generalization of the authorship graph in the previously-studied settings of peer grading and grant proposal review, in which each reviewer (paper) is connected to at most one paper (reviewer).

The review process is modeled by a second bipartite graph $\mathcal{G}$, termed as *review graph*, that also has the reviewers and submissions $(\mathcal{R}, \mathcal{P})$ as its vertices. This review graph has an edge between a reviewer and a paper if that reviewer reviews that submission. For each reviewer r_i $(i \in [m])$, we let $\mathcal{P}_i \subseteq \mathcal{P}$ denote the set of papers assigned to this reviewer for review, or in other words, the neighborhood of node r_i in the bipartite graph $\mathcal{G}$. To ensure balanced workloads across reviewers, we require that every reviewer is assigned at most μ papers for some integer $1 \leqslant \mu \leqslant n$. In other words, every node in $\mathcal{R}$ has at most μ neighbors (in $\mathcal{P}$) in graph $\mathcal{G}$. Additionally, each paper must be reviewed by at least λ reviewers for some integer $1 \leqslant \lambda \leqslant m$. Thus every node in the set $\mathcal{P}$ must have at least λ neighbors (in $\mathcal{R}$) in the graph $\mathcal{G}$. For any (directed or undirected) graph $\mathcal{H}$, we let the notation $E_{\mathcal{H}}$ denote the set of edges in graph $\mathcal{H}$.

At the end of the reviewing period, each reviewer provides a total ranking of the papers that she/he reviewed. For any set of papers $\mathcal{P}' \subseteq \mathcal{P}$, we let $\Pi(\mathcal{P}')$ denote the set of all permutations of papers in $\mathcal{P}'$. Furthermore, for any paper $p_j \in \mathcal{P}'$ and any permutation $\pi(\mathcal{P}') \in \Pi(\mathcal{P}')$, we let $\pi_j(\mathcal{P}')$ denote the position of paper p_j in the permutation $\pi(\mathcal{P}')$. At the end of the reviewing period, each reviewer r_i submits a total ranking $\pi^{(i)}(\mathcal{P}_i) \in \Pi(\mathcal{P}_i)$ of the papers in $\mathcal{P}_i$. We define a (partial) ranking profile $\boldsymbol{\pi} := (\pi^{(1)}(\mathcal{P}_1), \ldots, \pi^{(m)}(\mathcal{P}_m))$ as the collection of rankings from all the reviewers. When the assignment $\mathcal{P}_1, \ldots, \mathcal{P}_m$ of papers to reviewers is fixed, we use the shorthand $(\pi^{(1)}, \ldots, \pi^{(m)})$ for profile $\boldsymbol{\pi}$. For any subset of papers $\mathcal{P}' \subseteq \mathcal{P}$, we let $\boldsymbol{\pi}_{\mathcal{P}'}$ denote the restriction of $\boldsymbol{\pi}$ to only the induced rankings on $\mathcal{P}'$. Finally, when the ranking under consideration is clear from context, we use the notation $p \succ p'$ to say that paper p is ranked higher than paper p'.

Under this problem setup, the goal is to jointly design: (a) a paper-reviewer assignment scheme, that is, edges of the graph $\mathcal{G}$, and (b) an associated review aggregation rule $f : \prod_{i=1}^{m} \Pi(\mathcal{P}_i) \rightarrow \Pi(\mathcal{P})$ which maps from the ranking profile to an aggregate total ranking of all papers. For any aggregation function f, we let $f_j(\boldsymbol{\pi})$ be the position of paper p_j when the input to f is the profile $\boldsymbol{\pi}$.

Although we assume ordinal feedback from the reviewers, our results continue to hold if we have review scores as our input instead of rankings; our framework is flexible enough to take the scores into account (cf. Section 4.1). In what follows we define strategyproofness and efficiency that any conference review mechanism f should satisfy under our setting. Due to space limit, we omit all the proofs and refer interested readers to an online version [Xu *et al.*, 2018] for all the further details.

3.1 Strategyproofness

In the context of conference review, strategyproofness is defined with respect to a given conflict graph $\mathcal{C}$. It means that a reviewer cannot change the position of her conflicting papers, by manipulating the ranking she provides.

Definition 1 (Strategyproofness, SP)**.** A review process $(\mathcal{G}, f)$ is called strategyproof with respect to a conflict graph $\mathcal{C}$ if for every reviewer $r_i \in \mathcal{R}$ and every paper $p_j \in \mathcal{P}$ with $(r_i, p_j) \in E_{\mathcal{C}}$, under assignment $\mathcal{G}$, for every pair of profiles that differ only in the ranking given by reviewer r_i, the position of p_j is unchanged. Formally, for every $\boldsymbol{\pi} = (\pi^{(1)}, \ldots, \pi^{(i-1)}, \pi^{(i)}, \pi^{(i+1)}, \ldots, \pi^{(m)}))$ and $\boldsymbol{\pi}' = (\pi^{(1)}, \ldots, \pi^{(i-1)}, {\pi^{(i)}}', \pi^{(i+1)}, \ldots, \pi^{(m)})$, the results remain unchanged as $f_j(\boldsymbol{\pi}) = f_j(\boldsymbol{\pi}')$.

A strategyproof peer review procedure alone is inadequate with respect to any practical requirements – simply giving out a fixed, arbitrary evaluation makes the peer review procedure strategyproof. We therefore consider efficiency of the procedure in the next section, to ensure that the authors receive meaningful and helpful feedback for their work.

3.2 Efficiency (Unanimity)

We consider efficiency of a peer review process in the notion of "unanimity", which is one of the most prevalent and classic properties to measure the efficiency of a voting system in the literature of social choice [Fishburn, 2015]. At a colloquial level, unanimity states that when there is a common agreement among all reviewers, the aggregation of the opinions must also respect this agreement. We discuss two kinds of unanimity, termed as group unanimity (GU) and pairwise unanimity (PU). Both kinds of unanimity impose requirements on the aggregation function for given paper-reviewer assignment. The group unanimity is defined as follows:

Definition 2 (Group Unanimity, GU)**.** The pair $(\mathcal{G}, f)$ is said to be group unanimous (GU) if the following condition holds for every possible profile $\boldsymbol{\pi}$: For every set of papers $\mathcal{P}' \subset \mathcal{P}$ such that every reviewer ranks the papers she reviewed from $\mathcal{P}'$ higher than those she reviewed from $\mathcal{P} \backslash \mathcal{P}'$, the output $f(\boldsymbol{\pi})$ must satisfy $p_x \succ p_y$ for every pair of papers $p_x \in \mathcal{P}'$ and $p_y \in \mathcal{P} \backslash \mathcal{P}'$ such that at least one reviewer has reviewed both p_x and p_y.

Intuitively, group unanimity says that if papers can be partitioned into two sets such that every reviewer who has reviewed papers from both sets agrees that the papers she has reviewed from the first set are better than what she reviewed from the second set, then the final output ranking should respect this agreement. Our second notion of unanimity, termed pairwise unanimity, is a local refinement of group unanimity. This notion is identical to the classical notion of unanimity stated in Arrow's impossibility theorem [Arrow, 1950]. Notice that the classical unanimity considers every reviewer to review all papers (that is, $\mathcal{P}_i = \mathcal{P}, \forall i \in [m]$), whereas our notion is generalized to settings where reviewers may review only a subset of papers.

Definition 3 (Pairwise Unanimity, PU)**.** We define $(\mathcal{G}, f)$ to be pairwise unanimous (PU) if the following condition holds for every possible profile $\boldsymbol{\pi}$ and every pair of papers $p_{j_1}, p_{j_2} \in \mathcal{P}$: If at least one reviewer has reviewed both p_{j_1} and p_{j_2} and all the reviewers that have reviewed p_{j_1} and p_{j_2} agree on $p_{j_1} \succ p_{j_2}$, then $f_{j_1}(\boldsymbol{\pi}) \succ f_{j_2}(\boldsymbol{\pi})$.

An important property is that pairwise unanimity is stronger than group unanimity.

Proposition 1. If $(\mathcal{G}, f)$ is pairwise unanimous, then $(\mathcal{G}, f)$ is also group unanimous.

4 Positive Results

In this section we consider the design of paper-reviewer assignments and aggregation rules for strategyproofness and group unanimity (efficiency). Prior works on this topic consider a specific and restricted class of conflict graphs - - those with one-to-one relations between papers and reviewers – which do not capture conference peer review settings. We consider a more general class of conflict graphs and present a framework based on the partitioning-based method [Alon *et al.*, 2011], which we show can achieve group unanimous and strategyproofness. The key idea is to assign a paper to a reviewer only if there is no path between this paper and reviewer in the conflict graph $\mathcal{C}$. We then empirically demonstrate, using submission data from the ICLR-17 conference, that this class of conflict graphs is indeed representative of peer review settings. In addition to the feasibility, we present a simple trick to improve the practical appeal of our framework (and more generally the partitioning method) to conference peer review.

4.1 The Divide-and-Rank Framework

We now present our "Divide-and-Rank" framework consisting of the reviewer assignment algorithm (Algorithm 1) and the rank aggregation algorithm (Algorithm 2). At a higher level, our framework performs a partition of the reviewers and papers for assignment, and aggregates the reviews by computing a ranking which is consistent with any group agreements. Divide-and-Rank works for a general conflict graph $\mathcal{C}$ as long as the conflict graph can be divided into two reasonably-sized disconnected components.

Importantly, the framework is simple yet flexible in that the assignment within each partition and the aggregation among certain groups of papers can leverage any existing algorithm for assignment and aggregation respectively, which is useful as it allows to further optimize various other metrics in addition to strategyproofness and unanimity.

The Divide-and-Rank assignment algorithm begins by partitioning the conflict graph into two disconnected components

Algorithm 1 Divide-and-Rank assignment

Input: conflict graph $\mathcal{C}$, parameters λ, μ, assignment algorithm $\mathfrak{A}$
Output: an assignment of reviewers to papers
1: $(\mathcal{R}_C, \mathcal{P}_C), (\mathcal{R}_{\bar{C}}, \mathcal{P}_{\bar{C}}) \leftarrow \text{Partition}(\mathcal{C}, \lambda, \mu)$
2: use algorithm $\mathfrak{A}$ to assign papers $\mathcal{P}_{\bar{C}}$ to reviewers $\mathcal{R}_C$
3: use algorithm $\mathfrak{A}$ to assign papers $\mathcal{P}_C$ to reviewers $\mathcal{R}_{\bar{C}}$
4: **return** the union of assignments from step 2 and 3
5: ---
6: **procedure** PARTITION(conflict graph $\mathcal{C}$, parameters λ, μ)
7: run a BFS on $\mathcal{C}$ to get connected K components $\{(\mathcal{R}_k, \mathcal{P}_k)\}_{k=1}^K$
8: let $r_k = |\mathcal{R}_k|, p_k = |\mathcal{P}_k|, \forall k \in [K]$
9: initialize a table $T[\cdot,\cdot,\cdot] \in \{0,1\}^{K\times(m+1)\times(n+1)}$ so that $T[1, r_1, p_1] = T[1, 0, 0] = 1$, otherwise 0
10: **for** $k = 2$ to K **do**
11: $T[k, r, p] = T[k-1, r, p] \vee T[k-1, r-r_k, p-p_k]$, $\forall 0 \leqslant r \leqslant m, 0 \leqslant p \leqslant n$
12: **end for**
13: for $0 \leqslant r \leqslant m, 0 \leqslant p \leqslant n$, if there is no $T[K, r, p] = 1$ such that $\max\{\frac{p}{m-r}, \frac{n-p}{r}\} \leqslant \frac{\mu}{\lambda}$, return ERROR
14: use the standard backtracking in the table $T[\cdot,\cdot,\cdot]$ to return $(\mathcal{R}_C, \mathcal{P}_C)$ and $(\mathcal{R}_{\bar{C}}, \mathcal{P}_{\bar{C}})$
15: **end procedure**

that meet the requirements specified by μ and λ. Although dividing into more groups can lead to similar unanimity and strategyproof properties, we use two groups for simplicity and computational efficiency. The subroutine Partition first runs a breadth-first-search (BFS) algorithm to partition the original conflict graph into K connected components, where the kth connected component contains $r_k \geqslant 0$ reviewers and $p_k \geqslant 0$ papers. Next, the algorithm performs a dynamic programming to compute all the possible subset sums achievable by the K connected components. Here $T[k, r, p] = 1$ means that there exists a partition of the first k components such that one side of the partition has r reviewers and p papers, and 0 otherwise. The last step is to check whether there exists a subset C satisfying the requirement, and if so, runs a standard backtracking algorithm along the table to find the actual subset C. Clearly the Partition runs in $O(Knm)$, and since $K \leqslant n + m$, it runs in polynomial time in the size of the input conflict graph $\mathcal{C}$.

In the next step, the algorithm assigns papers to reviewers in a fashion that guarantees each paper is going to be reviewed by at least λ reviewers and each reviewer reviews at most μ papers. The assignment of papers in any individual component (to reviewers in the other component) can be done using any assignment algorithm (taken as an input $\mathfrak{A}$) as long as the algorithm can satisfy the (μ, λ)-requirements. Possible choices for the algorithm $\mathfrak{A}$ include the popular Toronto paper matching system [Charlin and Zemel, 2013] and others [Hartvigsen *et al.*, 1999, Garg *et al.*, 2010, Stelmakh *et al.*, 2018].

We then introduce the aggregation procedure in Algorithm 2. Generally speaking, the papers in each component are aggregated separately using the subroutine Contract-and-Sort. This aggregation in Contract-and-Sort is performed by a topological ordering of all strongly connected components (SCCs) according to the reviews, and then ranking the papers within each set using any arbitrary aggregation algorithm (taken as an input $\mathfrak{B}$)[1]. Possible choices for the algorithm $\mathfrak{B}$ include the modified Borda count [Emerson, 2013], Plackett-Luce aggregation [Hajek *et al.*, 2014], or others [Caragiannis *et al.*, 2017]. Moving back to Algorithm 2, the two rankings returned by Contract-and-Sort respectively for the two components are simply interlaced to obtain a total ranking over all the papers: the slots for C are reserved in set I, and $[n]\backslash I$ contain the slots for the remaining papers. In our extended version of the paper we also show that the interleaving only causes a small change w.r.t an underlying optimal ranking. The following theorem now shows that Divide-and-Rank satisfies group unanimity and is strategyproof.

Algorithm 2 Divide-and-Rank aggregation

Input: profile $\boldsymbol{\pi} = (\pi^{(1)}(\mathcal{P}_1), \ldots, \pi^{(m)}(\mathcal{P}_m))$, groups $(\mathcal{R}_C, \mathcal{P}_C), (\mathcal{R}_{\bar{C}}, \mathcal{P}_{\bar{C}})$ with $|\mathcal{P}_C| \geqslant |\mathcal{P}_{\bar{C}}|$, aggregation algorithm $\mathfrak{B}$
Output: total ranking of all papers
1: compute $\boldsymbol{\pi}_C$ as the restriction of profile $\boldsymbol{\pi}$ to only papers in $\mathcal{P}_C$, and $\boldsymbol{\pi}_{\bar{C}}$ as the restriction of profile $\boldsymbol{\pi}$ to only papers in $\mathcal{P}_{\bar{C}}$
2: $\pi_C \leftarrow \text{Contract-and-Sort}(\mathfrak{B}, \boldsymbol{\pi}_C)$
3: $\pi_{\bar{C}} \leftarrow \text{Contract-and-Sort}(\mathfrak{B}, \boldsymbol{\pi}_{\bar{C}})$
4: define $I = \left(\left\lfloor \frac{n}{|\mathcal{P}_C|} \right\rfloor, \left\lfloor \frac{2n}{|\mathcal{P}_C|} \right\rfloor, ..., n\right)$
5: **return** total ranking obtained by filling papers in $\mathcal{P}_C$ into positions in I in order given by π_C, and papers in $\mathcal{P}_{\bar{C}}$ into positions in $[n]\backslash I$ in order given by $\pi_{\bar{C}}$
6: ---
7: **procedure** CONTRACT-AND-SORT(aggregation algorithm $\mathfrak{B}$, profile $\widetilde{\boldsymbol{\pi}} = (\pi^{(1)}, \ldots, \pi^{(m')})$)
8: build a directed graph $G_{\widetilde{\boldsymbol{\pi}}}$ with the papers in $\widetilde{\boldsymbol{\pi}}$ as its vertices and no edges
9: **for** each $i \in [m']$ **do**
10: Suppose $\pi^{(i)} = (p_{j_1} \succ \ldots \succ p_{j_{t_i}})$, add a directed edge from p_{j_k} to $p_{j_{k+1}}$ in $G_{\widetilde{\boldsymbol{\pi}}}$, $\forall k \in [t_i - 1]$
11: **end for**
12: compute a topological ordering of the strongly connected components (SCCs) in $G_{\widetilde{\boldsymbol{\pi}}}$
13: for every SCC in $G_{\widetilde{\boldsymbol{\pi}}}$, compute a permutation of the papers in the component using algorithm $\mathfrak{B}$
14: **return** the permutation of all papers that is consistent with the topological ordering of the SCCs and the permutations within the SCCs
15: **end procedure**

Theorem 1. Suppose $\mathcal{C}$ can be partitioned into two groups $(\mathcal{R}_C, \mathcal{P}_C)$ and $(\mathcal{R}_{\bar{C}}, \mathcal{P}_{\bar{C}})$ such that there are no edges in $\mathcal{C}$ across the groups and that $\max\left\{\frac{|\mathcal{P}_C|}{|\mathcal{R}_{\bar{C}}|}, \frac{|\mathcal{P}_{\bar{C}}|}{|\mathcal{R}_C|}\right\} \leqslant \frac{\mu}{\lambda}$. Then Divide-and-Rank is group unanimous and strategyproof.

Remark. Our Divide-and-Rank framework aptly handles the various nuances of real-world conferences peer review, which render other existing methods inapplicable. This includes the facts that each reviewer can have conflicts with

[1] In the case where there are multiple topological orderings, any one of them suffices.

multiple papers and each paper can have conflicts with multiple reviewers, and furthermore that each reviewer may review only a subset of papers. Even under this challenging setting, our framework guarantees that no reviewer can influence the ranking of her own paper via strategic behavior, and that it is efficient from a social choice perspective.

Extension to review scores. Our framework can easily extend to a score-based setting, wherein each reviewer r_i provides a score s_{ij} for every paper $p_j \in \mathcal{P}_i$. The assignment algorithm remains the same in this setting; for aggregation, we can use the same procedure with the ranking induced by the review scores. The only difference is that in step 10 of Contract-and-Sort, we add an edge between every pair of papers $p_{j_1} \rightarrow p_{j_2}$ if $s_{ij_1} > s_{ij_2}$. This makes sure that the graph fully reflects the opinion of the reviewer and do not impose constraints on papers that are equally rated. On the other hand, the aggregation algorithm $\mathfrak{B}$ can also leverage the review scores for a more granularized ranking (e.g., average scores). Our definition of unanimity and strategyproof can also be straightforwardly extended to the score setting, and our framework still preserves these properties under these definitions. See our extended version of this paper [Xu *et al.*, 2018] for further details.

4.2 Analysis of ICLR-17 Submissions

Our Divide-and-Rank framework is based on a partitioning method that relies on a partition of the set of reviewers and papers so that there is no conflict across the partition. In this subsection we restrict attention to the authorship conflict graph, where we empirically verify that the partitioning conditions indeed hold in a conference peer review setting using data from the ICLR-17 conference. We then demonstrate how to make the partitioning method more appealing for conference peer review. In particular, we show that removing only a small number of reviewers can drastically reduce the size of the largest component in the conflict graph, thus providing great flexibility towards partitioning the papers and authors.

We analyzed all papers submitted to ICLR-17 with the given authorship relationship as the conflict graph. ICLR-17 received 489 submissions by 1,417 authors; we believe it is a good representative of a medium-sized modern conference. It is important to note that we consider only the set of authors as the entire reviewer pool (since we do not have access to the actual reviewer identities). Adding reviewers from outside the set of authors would only improve the results since these additional reviewers will have no edges in the (authorship) conflict graph. We first investigate the existence of (moderately sized) components in the conflict graph,which shows that the authorship graph is not only disconnected, but also has more than 250 components. The largest connected component(CC) contains 133 (that is, about 27%) of all papers, and the 2nd largest CC is much smaller, hence empirically verify our assumptions in Theorem 1.

The partitioning method is previously considered for the problem of peer grading [Kahng *et al.*, 2018], where the setting is quite homogeneous in that each reviewer (student) goes through the same course and hence any paper (homework) can be assigned to any reviewer. In peer review, however, different reviewers typically have different areas of expertise and hence their abilities to review any paper varies by the area of the paper. In order to accommodate this diversity in area of expertise in peer review, one must have a greater flexibility in terms of paper assignment. In our analysis, the largest CC contains 372 authors and 133 papers. It is reasonable to expect that a large number of reviewers with expertise required to review these 133 papers would also be in the same CC, thus a naïve application of Divide-and-Rank would assign these 133 papers to reviewers who may have a much less expertise. This is indeed a concern, and in what follows, we discuss a simple yet effective way to ameliorate this problem.

	#Authors removed from reviewer pool						
	0	5	10	15	20	50	100
#Components	253	268	278	292	302	334	389
1st #Authors	371	313	304	228	205	55	28
1st #Papers	133	114	110	82	74	18	8

Table 1: Statistics of the conflict graph on removing a small number ($< 7\%$) of authors from the reviewer pool of 1,417 authors.

We show empirically using the ICLR-17 data that by removing only a small number of authors from the reviewer pool of the ICLR-17 data, the conflict graph can be much more sparse, allowing for more flexible application of Divide-and-Rank. Specifically, we remove a small fraction of authors from the reviewer pool using the simple heuristic of removing the authors with the maximum degree in the (authorship) conflict graph. We then study the statistics of the resulting conflict graph (with all papers but only the remaining reviewers) in terms of the numbers and sizes of the CC. The results are shown in Table 1. After removing a small fraction – 100 authors which is only about 7% – the number of papers in the largest CC reduces by 94% to just 8. Likewise, the number of authors in the largest CC reduces to as small as 28 from 371 originally. It demonstrates that despite all the idiosyncrasies of conference peer review, the Divide-and-Rank and partitioning can be made practically applicable.

5 Negative Results

The positive results in the previous section focus on group unanimity, which is weaker than the conventional notion of unanimity, also known as pairwise unanimity. Moreover, the framework induces a disconnected review graph whereas the review graphs of conferences today are typically connected [Shah *et al.*, 2017]. It is thus natural to ask the following: Can a peer review system with a connected reviewer graph satisfy these properties? Can a strategyproof peer review system be pairwise unanimous? In this section we present some negative results toward these questions, thereby highlighting the critical impediments towards stronger results. Before stating our results, we give another notion of strategyproofness, which is significantly weaker than the notion of strategyproofness (Definition 1), and is hence termed as weak strategyproofness.

Definition 4 (Weak Strategyproofness, WSP)**.** A review process $(\mathcal{G}, f)$ is called *weakly strategyproof*, if for every reviewer r_i, there **exists** some paper $p_j \in \mathcal{P}$ such that for every pair of distinct profiles (under assignment $\mathcal{G}$) $\boldsymbol{\pi} = (\pi^{(1)}, \ldots, \pi^{(i-1)}, \pi^{(i)}, \pi^{(i+1)}, \ldots, \pi^{(m)})$ and $\boldsymbol{\pi}' = (\pi^{(1)}, \ldots, \pi^{(i-1)}, {\pi^{(i)}}', \pi^{(i+1)}, \ldots, \pi^{(m)})$, it is guaranteed that $f_j(\boldsymbol{\pi}) = f_j(\boldsymbol{\pi}')$.

Unanimity	Strategyproof	Requirement on $\mathcal{G}$	Possible?	Reference
Pairwise	None	Mild (see Corollary 1)	No	Theorem 2
Group	Weak	Mild (Connected $\mathcal{G}$)	Conjecture: No	Proposition 2
Group	Yes	None	Yes	Theorem 1

Table 2: Summary of our negative results (first two rows of the table), and a comparison to our positive result (third row).

In other words, weak strategyproofness requires for each reviewer there is at least *one* paper (not necessarily authored by this reviewer) whose ranking cannot be influenced by the reviewer. As the name suggests, strategyproofness is strictly stronger than weak strategyproofness, when each reviewer has at least one paper of conflict. We define the notion of weak strategyproofness mainly for our theoretical purposes of negative results, since WSP is too weak to be practical. However, even this extremely weak requirement is impossible to satisfy.

We summarize our results in Table 2. We show the property of group unanimity and strategyproofness for Divide-and-Rank; as the first direction of possible extension, we show in Theorem 2 that the slightly stronger notion of pairwise unanimity is impossible to satisfy under mild assumptions, even *without* strategyproof constraints. Then we explore the second direction of extension, by requiring a connected $\mathcal{G}$, and give conjectures that group unanimity and weak strategyproofness is impossible under this setting.

5.1 Impossibility of Pairwise Unanimity

In order to precisely state our result, we first introduce the notion of a *review-relation graph* $\mathcal{H}$. Given a paper-review assignment $\{\mathcal{P}_i\}_{i=1}^m$, the review-relation graph $\mathcal{H}$ is an undirected graph with $[n]$ as its vertices and where any two papers p_{j_1} and p_{j_2} are connected iff there exists at least one reviewer who reviews both the papers. With this preliminary in place, we are now ready to state our results:

Theorem 2. If $\mathcal{H}$ has a cycle of length 3 or more and there is no single reviewer reviews all the papers in the cycle, then there is no review process $(\mathcal{G}, f)$ that is pairwise unanimous.

The proof of Theorem 2 is similar to a Condorcet cycle proof. In the corollary below we give some direct implications of the condition in Theorem 2 when $|\mathcal{P}_1| = \cdots = |\mathcal{P}_m| = \mu$, that is, when every reviewer ranks a same number of papers.

Corollary 1. Suppose $|\mathcal{P}_1| = \cdots = |\mathcal{P}_m| = \mu \geqslant 2$. If $(\mathcal{G}, f)$ is pairwise unanimous, the following conditions hold:

(i) $\mathcal{H}$ does not contain any cycles of length $\mu + 1$ or more.

(ii) The set of papers reviewed by any pair of reviewers r_{i_1} and r_{i_2} must satisfy the condition $|\mathcal{P}_{i_1} \cap \mathcal{P}_{i_2}| \in \{0, 1, \mu\}$. In words, if a pair of reviewers review more than one common papers, they must review exactly the same set.

(iii) The number of distinct sets in $\mathcal{P}_i, \ldots, \mathcal{P}_m$ is at most $\frac{n-1}{\mu-1}$.

Remark. In modern conferences [Shah *et al.*, 2017], each reviewer usually reviews 3 to 6 papers. If the review process is pairwise unanimous, by Corollary 1(iii) the number of distinct review sets is much smaller than the number of reviewers; this severely limits the design of review sets, since many reviewers would be necessitated to review identical sets of papers. (ii) is also a relatively strong requirement, since the specialization of reviewers might not allow for such limit of the intersection of review sets. For instance, there is a large number of pairs of reviewers who review more than one common paper but none with the same set of papers [Shah *et al.*, 2017]. In summary, Theorem 2 and Corollary 1 show the difficulty to satisfy pairwise unanimity, even without strategyproofness.

5.2 Group Unanimity and Strategyproofness for a Connected Review Graph

Having shown that pairwise unanimity is too strong a requirement to satisfy, we now consider another direction for extension – conditions on the review graph $\mathcal{G}$. A natural question follows: Under what condition on the review graph $\mathcal{G}$ are both group unanimity and strategyproofness possible? Although we will leave the question of finding the exact condition open, we conjecture that if we require $\mathcal{G}$ to be connected, then group unanimity and strategyproofness cannot be simultaneously satisfied. To show our insights, we analyze an extremely simplified review setting and present a negative result under this setting.

Proposition 2. Consider any $n \geqslant 4$ and suppose $\mathcal{P} = P_1 \cup P_2 \cup P_3 \cup P_4$, where P_1, P_2, P_3, P_4 are disjoint nonempty sets of papers. Consider a review graph $\mathcal{G}$ with $m = 3$ reviewers, where reviewer r_1 reviews $\{P_1, P_2\}$, r_2 reviews $\{P_2, P_3\}$, and r_3 reviews $\{P_3, P_4\}$. Then there is no aggregation function f that is both weakly strategyproof and group unanimous.

Proposition 2 thus shows for simple review graph considered in the statement, group unanimity and weak strategyproofness cannot hold at the same time. We conjecture that such a negative result may hold for more general connected review graphs, which could be shown by identifying a component of the general review graph satisfying the condition of Proposition 2. This shows that our design process of the review graph in Section 4 is quite essential.

6 Discussion

We provide a framework and associated algorithms to impart strategyproofness to conference peer review. Our framework, besides guaranteeing strategyproofness, is importantly very flexible in allowing the program chairs to use the decision-making criteria of their choice. We complement these positive results with negative results showing that it is impossible for any algorithm to remain strategyproof and satisfy the stronger notion of pairwise unanimity. Future work includes considering efficiency from a statistical perspective and characterizing the precise set of conflict-of-interest graphs that permit (or not) strategyproofness.

Acknowledgements

YX was partially supported by DARPA award FA8750-17-2-0130. HZ would like to acknowledge the support from DARPA XAI project, contract FA87501720152. NBS was partially supported by NSF 1755656 and 1763734.

References

[Alon *et al.*, 2011] Noga Alon, Felix Fischer, Ariel Procaccia, and Moshe Tennenholtz. Sum of us: Strategyproof selection from the selectors. In *TARK*. ACM, 2011.

[Anderson *et al.*, 2007] Melissa S Anderson, Emily A Ronning, Raymond De Vries, and Brian C Martinson. The perverse effects of competition on scientists' work and relationships. *Science and engineering ethics*, 2007.

[Arrow, 1950] Kenneth J Arrow. A difficulty in the concept of social welfare. *Journal of political economy*, 1950.

[Aziz *et al.*, 2016] Haris Aziz, Omer Lev, Nicholas Mattei, Jeffrey S Rosenschein, and Toby Walsh. Strategyproof peer selection: Mechanisms, analyses, and experiments. In *AAAI*, pages 397–403, 2016.

[Barnett, 2003] William Barnett. The modern theory of consumer behavior: Ordinal or cardinal? *The Quarterly Journal of Austrian Economics*, 6(1):41–65, 2003.

[Brandt *et al.*, 2016] Felix Brandt, Vincent Conitzer, Ulle Endriss, Ariel D Procaccia, and Jérôme Lang. *Handbook of computational social choice*. 2016.

[Caragiannis *et al.*, 2017] Ioannis Caragiannis, Xenophon Chatzigeorgiou, George A Krimpas, and Alexandros A Voudouris. Optimizing positional scoring rules for rank aggregation. In *AAAI*, pages 430–436, 2017.

[Charlin and Zemel, 2013] L. Charlin and R. S. Zemel. The Toronto Paper Matching System: An automated paper-reviewer assignment system, 2013.

[Díez Peláez *et al.*, 2013] Jorge Díez Peláez, Óscar Luaces Rodríguez, Amparo Alonso Betanzos, Alicia Troncoso, and Antonio Bahamonde Rionda. Peer assessment in moocs using preference learning via matrix factorization. In *NIPS Workshop on Data Driven Education*, 2013.

[Dörfler *et al.*, 2017] Florian Dörfler, Yuanzhang Xiao, and Mihaela van der Schaar. Incentive design in peer review: Rating and repeated endogenous matching. *IEEE Transactions on Network Science and Engineering*, 2017.

[Douceur, 2009] John R Douceur. Paper rating vs. paper ranking. *ACM SIGOPS Operating Systems Review*, 2009.

[Emerson, 2013] Peter Emerson. The original Borda count and partial voting. *Social Choice and Welfare*, 2013.

[Fischer and Klimm, 2015] Felix Fischer and Max Klimm. Optimal impartial selection. *SIAM Journal on Computing*, 44(5):1263–1285, 2015.

[Fishburn, 2015] Peter C Fishburn. *The theory of social choice*. Princeton University Press, 2015.

[Garg *et al.*, 2010] N. Garg, T. Kavitha, A. Kumar, K. Mehlhorn, and J. Mestre. Assigning papers to referees. *Algorithmica*, 58(1):119–136, Sep 2010.

[Gibbard, 1973] Allan Gibbard. Manipulation of voting schemes: a general result. *Econometrica: journal of the Econometric Society*, pages 587–601, 1973.

[Hajek *et al.*, 2014] Bruce Hajek, Sewoong Oh, and Jiaming Xu. Minimax-optimal inference from partial rankings. In *Advances in Neural Information Processing Systems*, pages 1475–1483, 2014.

[Hartvigsen *et al.*, 1999] David Hartvigsen, Jerry C. Wei, and Richard Czuchlewski. The conference paper-reviewer assignment problem. *Decision Sciences*, 1999.

[Hazelrigg, 2013] GA Hazelrigg. Dear colleague letter: Information to principal investigators (PIs) planning to submit proposals to the Sensors and Sensing Systems (SSS) program October 1, 2013, deadline. 2013.

[Holzman and Moulin, 2013] Ron Holzman and Hervé Moulin. Impartial nominations for a prize. *Econometrica*, 81(1):173–196, 2013.

[Kahng *et al.*, 2018] Anson Kahng, Yasmine Kotturi, Chinmay Kulkarni, David Kurokawa, and Ariel D Procaccia. Ranking wily people who rank each other. In *AAAI*, 2018.

[Kurokawa *et al.*, 2015] David Kurokawa, Omer Lev, Jamie Morgenstern, and Ariel D Procaccia. Impartial peer review. In *IJCAI*, pages 582–588, 2015.

[Merrifield and Saari, 2009] Michael R Merrifield and Donald G Saari. Telescope time without tears: a distributed approach to peer review. *Astronomy & Geophysics*, 2009.

[Piech *et al.*, 2013] Chris Piech, Jonathan Huang, Zhenghao Chen, Chuong Do, Andrew Ng, and Daphne Koller. Tuned models of peer assessment in moocs. *arXiv preprint arXiv:1307.2579*, 2013.

[Satterthwaite, 1975] Mark Allen Satterthwaite. Strategy-proofness and arrow's conditions: Existence and correspondence theorems for voting procedures and social welfare functions. *Journal of economic theory*, 1975.

[Shah *et al.*, 2013] Nihar B Shah, Joseph K Bradley, Abhay Parekh, Martin Wainwright, and Kannan Ramchandran. A case for ordinal peer-evaluation in moocs. In *NIPS Workshop on Data Driven Education*, 2013.

[Shah *et al.*, 2016] Nihar B Shah, Sivaraman Balakrishnan, Joseph Bradley, Abhay Parekh, Kannan Ramchandran, and Martin J Wainwright. Estimation from pairwise comparisons: Sharp minimax bounds with topology dependence. *The Journal of Machine Learning Research*, 2016.

[Shah *et al.*, 2017] Nihar B Shah, Behzad Tabibian, Krikamol Muandet, Isabelle Guyon, and Ulrike von Luxburg. Design and Analysis of the NIPS 2016 Review Process. *arXiv preprint arXiv:1708.09794*, 2017.

[Stelmakh *et al.*, 2018] Ivan Stelmakh, Nihar B Shah, and Aarti Singh. PeerReview4All: Fair and accurate reviewer assignment in peer review. 2018.

[Stewart *et al.*, 2005] Neil Stewart, Gordon DA Brown, and Nick Chater. Absolute identification by relative judgment. *Psychological review*, 112(4):881, 2005.

[Thurner and Hanel, 2011] Stefan Thurner and Rudolf Hanel. Peer-review in a world with rational scientists: Toward selection of the average. *The European Physical Journal B*, 84(4):707–711, 2011.

[Tsukida and Gupta, 2011] Kristi Tsukida and Maya R Gupta. How to analyze paired comparison data. Technical report, DTIC Document, 2011.

[Xu *et al.*, 2018] Yichong Xu, Han Zhao, Xiaofei Shi, and Nihar B Shah. On Strategyproof Conference Peer Review. *arXiv preprint arXiv:1806.06266*, 2018.