Theoretical and Empirical results in Strategyproof Conference Review

https://j-b-z.github.io/

Jeremy Zhang , Advisor: Nihar Shah
PhD Students: Yichong Xu, Han Zhao, Xiaofei Shi

May 2019

1 Introduction

1.1 Motivation

Peer review in conferences provides a way to efficiently evaluate quality of submitted papers. However, despite its scalability, peer review is vulnerable to strategic behavior from its reviewers. By assigning certain papers low rankings, a reviewer may increase the overall rankings of their own paper, thereby increasing their chance of being accepted to the conferences. Due to this concern of integrity in paper submissions, various properties of conferences are studied to determine the impact a reviewer can and cannot have on their own papers.

1.2 Related Works

There are currently many interpretations on how to quantify strategyproofness in a conference. The primary source of difference arises from the method a reviewer uses to rank their papers. In general, papers are either judged with a *cardinal* ranking (where papers are assigned scores), or with an *ordinal* ranking (where papers are ordered relative to their goodness). There are various trade-offs with both approaches: *cardinal* rankings provide more information about papers with respect to another, especially papers not reviewed by

the same reviewer; however, then rankings are heavily affected by bias. On the other hand, *ordinal* rankings suffer less from reviewer bias, but are harder to join together to create an overall rankings of the papers.

In both cases, *strategyproofness* is often defined as a reviewer's inability to change overall rankings of their own papers by changing their own reviews or judgment. *Strateyproofness* by itself, however, is not a sufficient desirable property of a conference review structure; by fixing an arbitrary overall ranking of the papers, we throw away all reviewers' impact on paper rankings. Indeed, an individual reviewer cannot change the position of their own paper, but then the overall rankings are entirely independent of the opinions of the reviewers which is contrary to the purpose of peer review. Additional properties in the peer review setting are also introduced and studied to ensure that reviewers have some impact in the overall rankings of the papers.

One such desirable property is *unanimity*. At a high level, this dictates that if all reviewers agree that one set of papers should rank above another set of papers, then the overall ranking of the papers should respect this by ranking all the papers from the first over those in the second set. *Unanimity* itself is often formally defined from two perspectives: a social choice perspective and a statistical perspective. The social choice perspective treats *unanimity* as a property that is either completely satisfied or not satisfied, whereas the statistical perspective empirically quantifies unanimity based off of the sum of the individual reviewer's agreements with the overall rankings. The benefit of the social choice perspective is that it does not introduce an arbitrary metric, which itself needs to be analyzed. Furthermore, *unanimity* can be classified based off the number of papers agreed. *Pairwise unanimity* is satisfied if *unanimity* holds for all pairs of papers, whereas *Group unanimity* is satisfied if *unanimity* holds for any two sets which partition the entire paper set.

1.3 Our work

Previous work on conferences employing *ordinal* rankings suggest that conferences with a connected reviewer/paper conflict graph cannot simultaneously satisfy *strategyproofness* and *group unanimity* from a social choice perspective. However, this has never been directly show, and remains as a conjecture. I demonstrate a family of connected conflict graphs which cannot satisfy *strategyproofness* and *group unanimity*.

Though this negative result may seem grim, Xu et al. [1] propose an algorithm to partition the conflict graph in a way such that the remaining conflict graph can satisfy both positive properties. Unfortunately, this algorithm as well as several others in strategyproof conference review suffer from a lack of empirical results. This issue arises from the fact that conferences rely on the Toronto Paper Matching System (TPMS) to do their reviewer/paper assignments. However, since this algorithm is done in a closed box and primarily for conferences only, researchers find it difficult to measure the impact of strategyproof algorithms on TPMS assignment and quality. Therefore to tackle this, we have created our own dataset simulating the reviewer/paper conflict graph of the ICLR 2018 conferences as well as the similarity scores between reviewer paper pairs.

1.4 Personal Contributions

- With guidance from Nihar Shah, I found the families of connected conflict graphs for which the negative results were shown.
- In the creation of the ICLR 2018 simliarity matrix, I created the pipeline for scraping author information from arXiv.

2 Design

Our research was primarily split into two components: the theory portion dedicated to showing the negative results for some families of connected conflict graphs, and the empirical portion dedicated to constructing the similarity matrix dataset using author/paper submission information from the ICLR 2018 conference.

2.1 Theory Design

I adapted my notions of formal conference settings, *strategyproofness*, and *group unanimity* from the problem setting described by Xu et al.[1] Under their setting, a review process specified by two bipartite graphs, the first being a conflict graph, and the second being a review graph. Both graphs are bipartite graphs between the set of reviewers R and set of papers P. Then, a review process is *strategyproof* with respect to conflict graph C if for every pair r_i, p_j such that the edge (r_i, p_j) is present in the conflict graph the following conditions holds: for every pair of voting profiles that differ only in the rankings given by reviewer r_i, the position of p_j is fixed in the overall rankings.

Furthermore, a review process is *group unanimous* with respect to conflict graph C if for two nonempty disjoint paper sets P_1, P_2 which partition P, then if every reviewer ranks papers from P_1 higher than papers from P_2, then the overall rankings of papers must rank papers in P_1 over those in P_2.

My motivation for pursuing negative results for connected conflict graphs was inspired by Proposition 5.3 by Xu et al. [1], where they demonstrate and prove a specific example of a review graph which cannot be simultaneously *strategyproof* and *group unanimous*. By specifying several voting profiles, they are able to directly show this negative result. Using this strategy, I was able to show Result 1 for a general infinite family of review graphs.

After proving Result 1, I was encouraged by Nihar to generalize this findings for a larger family of graphs. I was able to do so by reducing another family of graphs to the original family of graphs in Result 1, thereby showing Result 2. I also tried to employ graph induction to construct other families of connected conflict graphs, but I was unable to show negative results for these families.

2.2 Empirical Design

The original purpose of this project was to provide a method for measuring the impact of the graph partitioning method proposed in Xu et al.[1] on the parameters maximized in the review graph as a result of matching. To do this, we created a similarity matrix emulating the strategies utilized by TPMS, and then maximize the scores of the similarities. On the surface level, the difficulty of this project is due to the fact that reviewer/paper information for conferences is not generally public information, and that the specific behavior of TPMS is unknown and algorithm itself is not readily available for us to use.

To resolve the issue of reviewer/paper information, we scraped submissions data from ICLR conferences, which releases all their submission data publicly. From this list, we have a list of papers, and then we approximate the set of reviewers by taking an arbitrary but large subset of authors who have submitted to the conference.

To replicate the behavior of TPMS, we assume we can represent an author and a paper as a corpus of text. By viewing each corpus as a bag of words, we can essentially normalize the frequencies of words and then take a dot product of the two vectors to yield a similarity score. To construct a corpus representative of an author, we modified a scraper script provided by the TPMS public code to scrape up to ten recent papers from an author and then convert the papers into text.

3 Evaluation

3.1 Theory Results

Below I have reproduced the theory results shown this semester.

Proposition 1. *Consider any $n \geq 4$ and suppose $\mathcal{P} = \bigcup_{i \in [n]} P_i$ where papers P_i are pairwise disjoint nonempty sets of papers. Consider a conflict graph G with $m = n$ reviewers where each reviewer i authors the papers in sets $P_{i+2 \pmod n}, P_{i+3 \pmod n}, ...P_{i-1 \pmod n}$. Then the review graph where each reviewer reviews 2 disjoint sets of papers cannot simultaneously satisfy group unanimity and strategyproofness.*

Proof. Fix some ranking of papers within each individual set P_i for $i \in [n]$ (e.g., according to the natural order of their indices). For the remainder of the proof, rankings of all papers considers this fixed rankings within these individual sets. With this in place, we consider the rankings in terms of the n sets of papers P_i.

By the structure of the conflict graph, reviewer i must review papers $i, i+1 \pmod n$. Let $P_i \succ_i P_j$ denote that reviewer i ranks all papers in P_i above papers in P_j. Suppose there is some f that satisfies group unanimity and strategyproofness for G. Then, consider the following 3 profiles:

1. $P_1 \succ_1 P_2 \succ_2 P_3 ... \succ_{n-1} P_n$ and $P_1 \succ_n P_n$
2. $P_1 \prec_1 P_2 \succ_2 P_3 \succ_3 P_4 ... \succ_{n-1} P_n$ and $P_1 \succ_n P_n$
3. $P_1 \prec_1 P_2 \succ_2 P_3 \succ_3 P_4 ... \succ_{n-1} P_n$ and $P_1 \prec_n P_n$

In the first profile, for every partition $\mathcal{P}_1 = \{P_1, P_2, ...P_i\}, \mathcal{P}_2 = \{P_{i+1}, ...P_n\}$ of $\mathcal{P}$, we have that $P_1 \succ P_2$. By group unanimity, this leads to the output $P_1 \succ P_2 \succ ...P_n$.

The first profile and second profile only differs in reviewer 1's rankings. Since f satisfies strategyproofness, the output must fix the position of sets $P_3, P_4, ...P_n$ in the output. Furthermore, since $\mathcal{P}_1 = \{P_2\}, \mathcal{P}_2 = \{P_1, P_3, ...P_n\}$ partitions $\mathcal{P}$ and $\mathcal{P}_1 \succ \mathcal{P}_2$, then by group unanimity the output must be $P_2 \succ P_1 \succ P_3 ... \succ P_n$.

The third profile changes reviwer n's rankings relative to the second profile. Then by strategyproofness, the rankings of $P_2, P_3, ...P_{n-1}$ are fixed relative to the output of the second profile. Partitioning the set into $\mathcal{P}_1 = \{P_2, P_3, P_4, ...P_{n-1}\}, \mathcal{P}_2 = \{P_1, P_n\}$, we have that $\mathcal{P}_1 \succ \mathcal{P}_2$. By group unanimity, P_1, P_2 must be the lowest ranking sets of papers in the output. Yet, since the rankings of the rest of the sets are fixed, one of the two sets must be ranked second in the output. This yields a contradiction.

□

Proposition 2. *Consider any $n \geq 8$ even and suppose $\mathcal{P} = \bigcup_{i \in [n]} P_i$ where papers P_i are pairwise disjoint nonempty sets of papers. Consider a conflict graph G with $m = n$ reviewers where each reviewer i authors the papers in sets $P_{i+3 \pmod n}, P_{i+4 \pmod n}, ...P_{i-1 \pmod n}$. Then the review graph where each reviewer reviews 3 disjoint sets of papers cannot simultaneously satisfy group unanimity and strategyproofness.*

Proof. Fix some ranking of papers within each individual set P_i for $i \in [n]$ (e.g., according to the natural order of their indices). For the remainder of the proof, rankings of all papers considers this fixed rankings within these individual sets. With this in place, we consider the rankings in terms of the n sets of papers P_i.

By the structure of the conflict graph, reviewer i must review papers $P_i, P_{i+1}, P_{i+2} \pmod n$. Let $P_i \succ_i P_j$ denote that reviewer i ranks all papers in P_i above papers in P_j. For $n \geq 8$, observe that reviewer i is the only reviewer that simultaneously reviews papers P_i and $P_{i+2 \pmod n}$, since reviewers must review 3 consecutive papers.

Fix some ranking of even papers $P_2, P_4, P_6, ...P_n$ (e.g., according to the natural order of these indices). For the remainder of the proof, rankings of papers considers this fixed rankings between these sets. Furthermore, rankings between P_i where i is odd and P_j where j is even ranks $P_i \succ P_j$. By group unanimity, then even indexed sets P_j must be the lowest papers in any output ranking. In conjunction with the fixed rankings of even indexed P_j, then these sets also have a fixed ranking in the output rankings.

With the rankings of the even indexed sets fixed, it remains to consider the rankings on the odd indexed papers $P_1, P_3, P_5, ...P_{n-1}$. Each reviewer i reviews two odd-indexed sets P_i, P_{i+2}, and an even-indexed set P_{i+1}. Since we have fixed even-indexed papers to be ranked below odd-indexed papers, then each reviewer i ranks P_{i+1} third in their profile, and has to choose a ranking between P_i, P_{i+2}. This review process then reduces to $\frac{n}{2}$ reviewers reviewing two papers each, in a manner analogous to that in the statement of **Proposition 1**. Furthermore, since $\frac{n}{2} \geq 4$, then by **Proposition 1**, this review graph cannot simultaneously satisfy strategyproofness and group unanimity.

□

These results are all proven under the problem setting described by Xu et al.[1], that is under the lens of social choice and *ordinal* rankings. However, the above analysis does not hold for different frameworks, since definitions would differ completely. Though the second result does generalize to a larger set of conflict graphs (compared to the family constructed in the first result), this family is still a very small subset of connected conflict graphs in general.

3.2 Empirical Results

We compiled a list of submitted authors and papers from the ICLR 2018 conference, which is publicly available data. Since there is not an inherent way to distinguish authors from a name search alone, we filtered scraped papers by their associated tags on arXiv. Specifically, I developed a script which, using `BeautifulSoup` python library, scraped arXiv for the author's five most recent papers pertaining to Machine Learning and several other related fields. Then, I used `pdftotext` to generate a body of text for each author in the author list.

With a corpus of text representing a papers and authors, Han generated and normalized bag of words representations for these texts, and then computed a `numpy` array representing the similarity matrix between pairs of papers and authors.

This dataset is available online at: `https://github.com/KeiraZhao/PeerReview/`.

4 Takeaways

Though we presume that all connected conflict graphs cannot satisfy both *strategyproofness* and *group unanimity*, Result 2 only shows this for a very small fraction of connected conflict graphs. This suggests that perhaps finding a minimum sufficient condition for this negative result requires additional insight or a different proof strategy.

In general, for theory research, I've found that there isn't quite a "right" way to approach an open problem; of course, if some one knew of a "right" way, then the problem would not be open. In contrast to this, there only exists recommendations, inspirations, and intuitions. My Result 1 was a direct application of a previous proposition to a more general case. Later on, my Result 2 was a result of hours of scratch work and trial-and-error. Furthermore, for this specific area of theory research, sometimes formulating the problem itself is as important as trying to solve it. Choosing the appropriate definitions of *strategyproofness* and *unanimity* was crucial for some strategies I used in my proofs.

Regarding the empirical results, the dataset is an approximation of the methods used in TPMS, due to the fact that we estimate its reviewer set by using all submitted authors. However, on top of this, the dataset may not contain entirely accurate similarities. Though majority of the cells should be objectively accurate, a small fraction of the authors do not have their papers uploaded arXiv and are thus unable to be properly represented with our methods. Furthermore, authors with name conflicts may be improperly represented, although we tried to ameliorate this by searching only for papers with topics related to the ICLR conference topics itself.

In general, for empirical research, I've found that many of my issues were issues that came along during the project, rather than issues we expected when planning the project. For example, the base code for the modified scraper was originally designed for python 2. Making the code compatible with python 3 generated a host of different issues with the libraries used by the script, creating a bunch of unexpected errors. Furthermore, initially, we were planning to scrape Google Scholar for a more accurate depiction of the authors themselves. However, after multiple failed attempts to scrape Google Scholar (including attempts with variable timers and AWS cloud machines), we decided to scrap arXiv instead. Even then, arXiv doesn't tolerate mass scraping, although we found that putting a variable delay long enough will ensure (for the most part) that arXiv will permit our scraping.

For the better or worse, I've found research in general to be fairly flexible, yet unpredictable. Research is flexible in the sense that for most things, one can approach the problem in any way they see fit, even redesigning and re-framing their research if they need to. On the other hand, I've found that planning for research is pretty difficult. Creating a time line in particular is extremely difficult, since the challenges one runs into in research is spontaneous and hard to predict. Furthermore, research (especially empirical research) involves a lot of coordination from the involved parties, otherwise mis-communications and delays inevitably occur.

5 Conclusion

As a result of the theory results, we have some families of connected conflict graphs which cannot satisfy both *strategyproofness* and *group unanimity*, further supporting the conjecture that the negative result holds for all connected conflict graphs. Future studies may take inspirations from the proof techniques I've used to make broader generalizations to conflict graphs. Specifically, future work may involve relaxing some conditions on the conflict graph, or extending it a bit further and look for familiar structures to again prove the negative result. In general, the families I've exhibited may demonstrate specific properties that might give insight into the minimum sufficient conditions on conflict graphs to derive the negative results.

Due to the empirical project, we have published a dataset for similarities between authors and papers of the ICLR 2018 conference. This enables future works to empirically asses the impact of an algorithm on the total similarity scores of an assignment.

References

[1] Yichong Xu, Han Zhao, Xiaofei Shi, and Nihar B. Shah. On Strategyproof Conference Peer Review. *arXiv e-prints*, page arXiv:1806.06266, Jun 2018.