Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21)
Early Career Track

Towards Fair and Transparent Algorithmic Systems

Yair Zick
University of Massachusetts, Amherst
yzick@umass.edu

Abstract

My research in the past few years has focused on fostering trust in algorithmic systems. I often analyze scenarios where a variety of desirable trust-oriented goals must be simultaneously satisfied; for example, ensuring that an allocation mechanism is both fair and efficient, or that a model explanation framework is both effective and differentially private. This interdisciplinary approach requires tools from a variety of computer science disciplines, such as game theory, economics, ML and differential privacy.

1 Overview

Recent years have seen algorithms outperform humans in a rapidly growing list of tasks. Not only are they able to quickly process vast amounts of data, automated decision makers are often perceived as more accurate and impartial (compared to their human counterparts). However, accuracy often comes at the cost of *fairness*, *transparency*, and ultimately - trust in automated decision makers. To achieve satisfactory levels of accuracy, one must train complex models on large quantities of data. Even highly accurate algorithms can be unfair. Since accuracy is measured across the entire dataset, decision-making algorithms can maintain accuracy while consistently offering worse outcomes to some individuals (say, members of a very small minority population). What's worse, even if one were able to identify undesirable algorithmic discriminatory behavior, it would be extremely challenging to identify its underlying causes. For example, a course allocation mechanism may offer excellent course slots to 99% of the student population, but consistently fail to offer desirable classes to 1% of the student cohort (say, offering random classes to exchange students). Such a mechanism would still be considered highly accurate — after all, it is able to offer good outcomes to 99% of the students — and treat certain students unfairly. Our recent work focuses on incorporating fairness and efficiency criteria in a variety of resource allocation domains. In addition, we study how various aspects of trustworthy ML come into play in the design and implementation of algorithmic model explanations, i.e. frameworks that elucidate the internal workings of complex decision makers to stakeholders. Accuracy/fairness concerns often arise in model explanations as well: is the model explanation any good? How can we ensure that it is well behaved across different populations and domains? For example, our explanation framework might consistently offer bad explanations to certain minority groups, or inadvertantly expose such populations to security breaches (e.g. exposing their personal data to potential malicious actors). Ensuring good practices and ethical norms in AI/ML applications is a foundational problem, requiring a broad interdisciplinary effort. My research focuses on two key aspects of this problem: *norms* and *data*. In particular, my work focuses on (a) normative analysis of algorithmic decision makers; (b) ensuring the fair treatment of individuals and groups when allocating sparse resources and (c) using data to derive good solutions in game-theoretic domains.

2 Norms in Algorithmic Decision Making Systems

What makes an algorithm fair? The answer to this question largely depends on how one would define fairness; in other words, one needs to identify *norms*, or *axioms* required from the algorithmic system. This idea is not new; axiomatic analysis is a commonly used technique in theoretical economics, employed in cooperative games [Shapley, 1953], bargaining [Nash, 1950], and social choice theory [Arrow, 1950]. With this paradigm in mind, I design *fair allocation algorithms* and *black-box model explanations* that provably satisfy a set of desirable properties. In some cases, we can even show that certain frameworks *uniquely* satisfy a set of criteria; in other words, if we agree that the properties we propose make sense, then our analysis points to a single way of explaining a model/allocating resources/deciding on an outcome. The axiomatic approach is appealing for a variety of reasons.

First of all, it offers a set of provable assurances regarding one's framework: in a recent paper, we present an algorithm that allocates items to individuals in a *provably* efficient and envy-free manner; in others, we show that certain model explanation frameworks are the only ones that satisfy a certain set of properties. Secondly, not only does the axiomatic approach provide provable guarantees to users, it offers stakeholders the opportunity to shift the debate from algorithms and formulas, to norms and desiderata: while fair allocation algorithms can rely on complex foundations (such as matroids, maximum flows, or matching theory), the axioms

that they satisfy are often simple enough to explain: envy-freeness is a concept that most people intuitively grasp at a very young age; similarly, we would want our model explanations to be unbiased, i.e. not a-priori assign importance to some features (e.g. deciding that individual income is to be assigned at least 10% of the responsibility for an outcome without even observing the data). By shifting the debate to the realm of norms and axioms, we offer a broader community of stakeholders (beyond computer scientists, economists and mathematicians) an opportunity to participate in the design of systems that may greatly affect them.

2.1 An Axiomatic Approach to Algorithmic Transparency

Decision-making algorithms are often trained over vast quantities of data, and incorporate complex architectures (e.g. deep neural networks, or large-scale online auction mechanisms); as a result, they are *black boxes*: their internal workings are not understood by various stakeholders — even their own designers. In particular, it would be immensely difficult to explain how such models arrive at their decisions. The lack of algorithmic transparency raises significant concerns: it is unclear whether these algorithms are making decisions based on wrong parameters, whether they are inadvertently discriminatory in nature, or potentially expose their users' data. In a recent line of work, we propose various methods for automatically generating *model explanations*. Consider, for example, an applicant for a bank loan, whose eligibility is deteremined by an automated decision maker. If this person has not been granted a loan, they need to be offered an explanation why: what were the most important factors in making this decision? How can the user change the outcome?[1] Formally, we are given a specific datapoint $\vec{x}_0$; what made a classification algorithm c assign $\vec{x}_0$ a label of $c(\vec{x}_0)$? We generate explanations based on a given dataset, its labels, and (possibly) additional information; our focus in past work is on *feature-based explanations*, assigning a value $\phi_i(\vec{x}_0)$ to each feature of the datapoint $\vec{x}_0$. The value $\phi_i(\vec{x}_0)$ roughly corresponds to the degree to which i affects the label of $\vec{x}_0$; a great deal of our analysis deals with the question of *how* one might capture the notion of effect: should we consider only effects observed in the training data? Is it reasonable to examine what the model does on randomly sampled points? What computations are we allowed to execute on the model (e.g. access to the model's gradient or internal parameters). We take an *axiomatic approach*. Rather than arbitrarily choosing explanation frameworks, we set forth a number of natural explanation desiderata, and proceed to show that a given measure uniquely satisfies these properties. Our initial work in this space [Datta *et al.*, 2015] showed some promise — for example, it was able to identify the importance of language and location in determining ads presented to Google users — but its axioms turned out to be too restrictive: our measure was unable to attribute influence to any feature on richer datasets. Using a more robust set of axioms, we propose *Quantitative Input Influence* (QII) [Datta *et al.*, 2016], a method based on axioms from cooperative game theory [Young, 1985]. Interestingly, QII has deep connections to the formal theory of counterfactual causality [Chockler and Halpern, 2004], a connection we are still in the process of exploring.

QII satisfies several desirable properties; however, it suffers from two major issues: it is expensive to compute, and requires access queries to the underlying black-box classifier c: in order to generate explanations, QII asks questions like "what happens if we randomize the values of a set of features S?", which assesses the model's expected behavior over a randomized domain that we may not have access to. In a recent line of work [Sliwinski *et al.*, 2019], we consider a set of axioms for *data-driven explanations*, assuming no access to the classifier. Our approach, termed *Monotone Influence Measures (MIM)*, is computationally efficient, and is able to generate interpretable explanations in a variety of domains.

Model explanations are essentially a means of conveying additional information about the decision-making process to end users. This raises a potential concern: adversaries can exploit additional information encoded in the model explanations to conduct *inference attacks*; for example, recovering segments of the training data, or uncover hidden user attributes. In a recent line of work [Shokri *et al.*, 2021], we explore the vulnerability of model explanations to adversarial behavior. Our results indicate that indeed, some popular explanation methods [Koh and Liang, 2017; Ribeiro *et al.*, 2016] are highly exploitable, and could be used to reconstruct some of the model training data. In another line of work we provide provably secure model explanations [Patel *et al.*, 2020]; this work also shows the inherent tradeoffs between privacy, accuracy, and fairness in model explanations. As can be expected, private model explanations are less accurate; moreover, we show that minority populations are offered worse model explanations than more common classes, and that one needs to lower explanation accuracy further in order to protect their data.

Going beyond single-feature attribution methods, in a recent paper we provide an axiomatic treatment of a *high-dimensional* model explanation framework [Patel *et al.*, 2021]. Our characterization yields a rather intuitive model explanation framework, that is able to capture natural feature synergies, both in theory and on real-world datasets.

2.2 Fair Algorithmic Decision Making

Designing algorithms that treat users fairly is a fundamental AI problem, arising in a variety of domains, from computational resource allocation, via picking a set of candidates, to fairly dividing revenue amongst collaborative agents. I have explored a number of fairness related topics in the past few years.

Envy-Free Rent Division

A group of roommates would like to allocate rooms and divide the rent in a fair and mutually agreeable manner. An algorithm for this problem has been implemented on spliddit.org; its outputs are guaranteed to be *envy-free* (EF): the rent and the rooms are allocated so that no tenant prefers another's room, given that room's price. In [Gal *et al.*, 2016;

[1] While similar, the explanations offered to answer the two questoins are quite distinct: the first *highlights* important features, whereas the second offers users *recourse* (see [Barocas *et al.*, 2020]).

Gal *et al.*, 2017], we show that envy-freeness is neither theoreticall fair, nor is it necessarily perceived as fair. We propose choosing an EF outcome that maximizes the welfare of the least well-off tenant, called the *maximin EF outcome*. We show that the maximin EF outcome has significant welfare effects on data collected from spliddit.org, changing the rent allocation by hundreds of dollars in some cases. Moreover, in our human study, Spliddit users indicated a preference towards the maximin EF outcome over the original (arbitrary) EF outcome. This finding led to a reimplementation of the algorithm used on spliddit.org.

Fair and Diverse Allocation of Indivisible Goods

The Singapore public housing allocation system is a dominant force in the Singapore housing market: over 80% of Singaporeans live in (heavily subsidized) public housing developments. The government office in charge of housing — the Singapore Housing Development Board (HDB) — implements an *ethnic integration policy*. The three major ethnic groups in Singapore (Chinese, Malay and Indian/Others) must be adequately represented in every HDB development; this is achieved by imposing ethnic quotas: upper bounds on the number of applicants from each ethnicity that are allowed to purchase flats in a new development. We begin our study [Benabbou *et al.*, 2018] with a simple question: *what are the welfare effects of imposing ethnic quotas?* We model this setting as an assignment problem where players (prospective buyers) have different *types* and the goods (apartments) are partitioned to *blocks*. Each block can house no more than a certain number of players of each type. Naturally, limiting players by type quotas decreases social welfare. Thus, our objective is to bound the *price of diversity*: the ratio between the welfare of the optimal *unconstrained* and the optimal *type-constrained* assignment. We show several bounds on the price of diversity in terms of natural problem parameters, and apply our methods to data from the Housing Development Board. Our empirical results are encouraging: under several natural utility models, our (optimal) allocation mechanisms, and the (suboptimal) mechanisms implemented by HDB exhibit a low price of diversity.

In two recent papers, we show how the notion of envy is sensible in the group allocation setting [Benabbou *et al.*, 2019]; we treat ethnic groups as individuals, and define group envy as the capacity to which an individual ethnicity desires the apartments allotted to other ethnic groups. While envy is indeed a compelling fairness notion, it does not take into consideration group *size*; in [Chakraborty *et al.*, 2020], we introduce the notion of *weighted envy-freeness*, and offer several mechanisms for computing weighted envy-free outcomes.

Group preferences in housing allocation mechanisms are assumed to be matching-based valuations: the more homes that a group can allocate to individuals that want them, the better. This naturally induces a highly structured valuation profile; matching based valuations are known as OXS valuations in the economics literature [Leme, 2017]. Under a mild assumption (that individuals either approve or disapprove of any apartment), these valuations constitute a subclass of a large, non-trivial valuation class: *binary submodular*, or *matroid rank* valuations. For example, introducing cardinality constraints to valuations (e.g., agents may receive no more than k items each) prevents them from being matching-based valuations, while retaining submodularity. In a recent paper [Benabbou *et al.*, 2020], we provide the "ultimate solution" for this valuation class: an efficient algorithm that computes a socially optimal, envy-free (up to one good) allocation. In addition, we show that binary submodular valuations admit allocations that satisfy several other desirable properties, though these are computationally hard to find.

3 Data-Driven Solution Concepts

My PhD thesis focused primarily on the mathematical foundations of game theory; while the theory itself was interesting and compelling, I wanted to see whether it could be used in order to test — and predict — actual human behavior. More broadly, do *humans actually care about provably fair algorithms*? The answer to this question is elusive, as data on strategic decision making is hard to collect and assess. Furthermore, most game-theoretic literature assumes knowledge of underlying player preferences. This is an unrealistic assumption in practice, which led me to explore the possibility of *inferring game-theoretic solution concepts from data*.

3.1 Strategic Cooperative Behavior in the Wild

Human experiments were a key part of our work in the EF rent division domain [Gal *et al.*, 2016]. We selected users of the spliddit.org service (ones who legitemately used the website to divide rent among roommates), and asked them to reevaluate their initial allocation versus the newly proposed maximin share rent division. We tried out different questionnaire interfaces on Amazon Mechanical Turk (AMT), before settling on a simple two question format. Our results show that users prefer the maximin share method to a standard EF mechanism. This preference held true even if they were required to pay a higher rent.

We continued to analyze human collaborative behavior, this time examining reward division in a more controlled lab environment [Bachrach *et al.*, 2017; Mash *et al.*, 2020]. We asked users to complete a simple collaborative task - each user i has a weight w_i, and they need to form a coalition of weight at least 10. If they successfully do so, they receive a reward of 100 points, which they need to split amongst coalition members in an agreeable manner. Users were allowed to iteratively propose coalitions and splits; gameplay continued until an agreeable split was found, or a maximal number of rounds was reached. Using initial test runs on AMT and in the lab, we trained an agent who was able to consistently out-negotiate humans, by using cooperative solution concepts to determine proposal structure. We designed a more engaging interface to get users to play a more complex cooperative game: instead of combining their weights to achieve a single objective, users could complete multiple different tasks (e.g. there might be a task that needs a total weight of 5 which pays 40, in addition to a task whose required weight is 10 but pays 100); we established the theoretical properties of our model [Nguyen and Zick, 2018], and provided an in-depth analysis of human play [Gal *et al.*, 2020]. It turns out that humans naturally bargain their way towards (approximately) fair payoff

divisions in a large percentage of games, a result that confirms known models of cooperative bargaining [Aumann and Maschler, 1964].

3.2 A Learning Framework for Game-Theoretic Solution Concepts

Game theoretic solution concepts are often meant to be utilized in full information domains: we are fully aware of player preferences, and algorithms that output solutions rely on this assumption to work. What happens when we are not given player preferences, but rather partial observations of their realizations over sampled data?

For example, to find a fair payoff division for an unknown cooperative game v, one can use data to train an approximation of v, say v^*, and find a satisfactory payoff division for v^*. This approach is problematic. First, it may be hard to approximate v in the first place; more importantly, *fair solutions for an approximate model need not be approximately fair for the original model*. We propose an alternative approach in [Balcan *et al.*, 2015]: instead of learning an approximate model, learn an *approximate solution directly from data*. Our learned solutions are guaranteed to be *mostly* core-stable with respect to the original game, and can be efficiently learned from data, even for complex cooperative games. This last point is particularly interesting: intuitively, it means that in the domain of cooperative games, core-stable payoff divisions are fundamentally simpler mathematical objects than the games upon which they're computed. In later works, we extend our results to the more complex class of hedonic games [Igarashi *et al.*, 2019; Sliwinski and Zick, 2017]: instead of assigning a global value to every subset S, each player i assigns a value (or preference) over S, $v_i(S)$. In addition, rather than trying to identify "well-behaved" payoff divisions, we are interested in finding "well-behaved" partitions of players to disjoint groups (also known as coalition structures). In a recent paper, we propose a general framework for learning game-theoretic solution concepts from data, employing novel definitions of function class complexity [Jha and Zick, 2020]; in addition to our cooperative game-theoretic frameworks, our results can be applied in many other game-theoretic domains (e.g [Balcan *et al.*, 2018; Syrgkanis, 2017; Huang *et al.*, 2018; Zhang and Conitzer, 2019]. In a recent paper, we show how the solution-learning framework in [Jha and Zick, 2020] can be applied to find PAC market equilibria for a variety of market valuation classes [Lev *et al.*, 2021].

4 Future Directions

Algorithmic fairness/accountability/transparency is a fast growing field, which makes for exciting times ahead. There are still plenty of directions for formally defining and evaluating transparency measures: not only in developing new frameworks (namely, ones based on non-linear explanation models), but also assessing their risks. As our recent work [Shokri *et al.*, 2021] indicates, model explanations may pose a significant risk to their users. This raises a natural question: can we develop *meaningfull* model explanations that maintain *privacy guarantees*? While we offer some answers to this question in [Patel *et al.*, 2020], there is still a lot that we do not know: can we offer any meaningful privacy guarantees to other types of model explanations, such as rule-based systems or algorithmic recourse? Can we ensure that all parts of the data space receive good model explanations? What are the potential costs of such guarantees?

We are also starting to explore alternative characterizations of causal analysis in model explanations: can we formally capture the relation between formal causality and model explanations? What types of explanations are "causality-faithful"?

In the fairness domain, I am interested in the interplay of individual and group fairness: can we jointly guarantee that both individuals and groups are treated fairly? Can we come up with algorithms that ensure that group and individual fairness guarantees both hold? In addition, I continue to explore how structural restrictions on preference profiles affect fairness guarantees; in particular, whether we can guarantee the existence of fair and efficient outcomes for preference profiles that conform to certain combinatorial structures.

Acknowledgments

I would like to thank my coauthors and collaborators throughout the years; in particular, I would like to thank my PhD advisor, Edith Elkind, whose great advice and support I still rely upon today. I would also like to thank my postdoc advisors and mentors, Anupam Datta and Ariel Procaccia, who helped me become a better researcher. I would like to thank my group members: Nawal, Mithun, Vinh, Ridi, TJ, Urban, Wei, Duy, Neel, Justin, Jakub, Martin, Alan, and Vignesh. My work has been generously supported by several grants throughout the years, in particular the Singapore NRF Research Fellowship R-252-000-750-733 and the NRF AI Research award R-252-000-A20-490.

References

[Arrow, 1950] Kenneth J. Arrow. A difficulty in the concept of social welfare. *Jour. of Political Economy*, 58(4):328–346, 1950.

[Aumann and Maschler, 1964] Robert J. Aumann and Michael Maschler. The bargaining set for cooperative games. *Advances in game theory*, 52:443–476, 1964.

[Bachrach *et al.*, 2017] Yoram Bachrach, Kobi Gal, Moshe Mash, and Yair Zick. How to form winning coalitions in mixed human-computer settings. In *Proc.of the 26th IJCAI*, pages 465–471, 2017.

[Balcan *et al.*, 2015] Maria F. Balcan, Ariel D. Procaccia, and Yair Zick. Learning cooperative games. In *Proc.of the 24th IJCAI*, pages 475–481, 2015.

[Balcan *et al.*, 2018] Maria F. Balcan, Travis Dick, Ritesh Noothigattu, and Ariel D. Procaccia. Envy-free classification. *CoRR*, abs/1809.08700, 2018.

[Barocas *et al.*, 2020] Solon Barocas, Andrew D. Selbst, and Manish Raghavan. The hidden assumptions behind counterfactual explanations and principal reasons. In *Proc.of the 3rd FAT**, pages 80–89, 2020.

[Benabbou *et al.*, 2018] Nawal Benabbou, Mithun Chakraborty, Vinh Ho Xuan, Jakub Sliwinski, and Yair Zick. Diversity constraints in public housing allocation. *Proc.of the 17th AAMAS*, pages 973–981, 2018.

[Benabbou *et al.*, 2019] Nawal Benabbou, Mithun Chakraborty, Edith Elkind, and Yair Zick. Fairness towards groups of agents in the allocation of indivisible items. In *Proc.of the 28th IJCAI*, pages 95–101, 2019.

[Benabbou *et al.*, 2020] Nawal Benabbou, Mithun Chakraborty, Ayumi Igarashi, and Yair Zick. Finding fair and efficient allocations when valuations don't add up. In *Proc.of the 13th SAGT*, pages 32–46, 2020.

[Chakraborty *et al.*, 2020] Mithun Chakraborty, Ayumi Igarashi, Warut Suksompong, and Yair Zick. Weighted envy-freeness in indivisible item allocation. In *Proc.of the 19th AAMAS*, pages 231–239, 2020.

[Chockler and Halpern, 2004] Hannah Chockler and Joseph Y. Halpern. Responsibility and blame: A structural-model approach. *Jour. of AI Research*, 22:93–115, 2004.

[Datta *et al.*, 2015] Amit Datta, Anupam Datta, Ariel D Procaccia, and Yair Zick. Influence in classification via cooperative game theory. In *Proc.of the 24th IJCAI*, 2015.

[Datta *et al.*, 2016] Anupam Datta, Shayak Sen, and Yair Zick. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *Proc.of the 37th Oakland*, pages 598–617. IEEE, 2016.

[Gal *et al.*, 2016] Ya'akov (Kobi) Gal, Moshe Mash, Ariel D. Procaccia, and Yair Zick. Which is the fairest (rent division) of them all? In *Proc.of the 17th EC*, pages 67–84, 2016.

[Gal *et al.*, 2017] Ya'akov (Kobi) Gal, Moshe Mash, Ariel D. Procaccia, and Yair Zick. Which is the fairest (rent division) of them all? *Jour. of the ACM*, 64(6:39), 2017.

[Gal *et al.*, 2020] Kobi Gal, Ta Duy Nguyen, Quang Nhat Tran, and Yair Zick. Threshold task games: Theory, platform and experiments. In *Proc.of the 19th AAMAS*, pages 393–401, 2020.

[Huang *et al.*, 2018] Zhiyi Huang, Yishay Mansour, and Tim Roughgarden. Making the most of your samples. *SIAM J. Comput.*, 47:651–674, 2018.

[Igarashi *et al.*, 2019] Ayumi Igarashi, Jakub Sliwinski, and Yair Zick. Forming probably stable communities with limited interactions. In *Proc.of the 33rd AAAI*, pages 2053–2060, 2019.

[Jha and Zick, 2020] Tushant Jha and Yair Zick. A learning framework for distribution-based game-theoretic solution concepts. In *Proc.of the 21st EC*, 2020.

[Koh and Liang, 2017] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proc.of the 34th ICML*, pages 1885–1894, 2017.

[Leme, 2017] Renato Paes Leme. Gross substitutability: An algorithmic survey. *Games and Economic Behavior*, 106:294 – 316, 2017.

[Lev *et al.*, 2021] Omer Lev, Vignesh Viswanathan, Neel Patel, and Yair Zick. The price is (probably) right: Learning market equilibria from samples. In *Proc.of the 20th AAMAS*, pages 755–763, 2021.

[Mash *et al.*, 2020] Moshe Mash, Roy Fairstein, Yoram Bachrach, Kobi Gal, and Yair Zick. Human-computer coalition formation in weighted voting games. *ACM Trans. Intell. Syst. Technol.*, 11(6):73:1–73:20, 2020.

[Nash, 1950] John F. Nash. The bargaining problem. *Econometrica*, 18(2):155–162, 1950.

[Nguyen and Zick, 2018] Ta Duy Nguyen and Yair Zick. Resource based cooperative games: Optimization, fairness and stability. In *Proc.of the 11th SAGT*, 2018.

[Patel *et al.*, 2020] Neel Patel, Reza Shokri, and Yair Zick. Model explanations with differential privacy. *CoRR*, abs/2006.09129, 2020.

[Patel *et al.*, 2021] Neel Patel, Martin Strobel, and Yair Zick. High dimensional model explanations: An axiomatic approach. In *Proc.of the 4th FAccT*, pages 401–411, 2021.

[Ribeiro *et al.*, 2016] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *Proc.of the 22nd KDD*, pages 1135–1144, 2016.

[Shapley, 1953] Lloyd S. Shapley. A value for n-person games. In *Contributions to the Theory of Games, vol. 2*, Annals of Mathematics Studies, no. 28, pages 307–317. Princeton University Press, 1953.

[Shokri *et al.*, 2021] Reza Shokri, Martin Strobel, and Yair Zick. On the privacy risks of model explanations. In *Proc.of the 4th AIES*, 2021.

[Sliwinski and Zick, 2017] Jakub Sliwinski and Yair Zick. Learning hedonic games. In *Proc.of the 26th IJCAI*, pages 2730–2736, 2017.

[Sliwinski *et al.*, 2019] Jakub Sliwinski, Martin Strobel, and Yair Zick. A characterization of monotone influence measures for data classification. In *Proc.of the 33rd AAAI*, pages 718–725, 2019.

[Syrgkanis, 2017] Vasilis Syrgkanis. A sample complexity measure with applications to learning optimal auctions. In *Proc.of the 30th NIPS*, pages 5352–5359, 2017.

[Young, 1985] Hobart P. Young. Monotonic solutions of cooperative games. *International Jour. of Game Theory*, 14(2):65–72, 1985.

[Zhang and Conitzer, 2019] Hanrui Zhang and Vincent Conitzer. A PAC framework for aggregating agents' judgments. In *Proc.of the 33rd AAAI*, pages 2237–2244, 2019.