2023 62nd IEEE Conference on Decision and Control (CDC)
December 13-15, 2023. Marina Bay Sands, Singapore

Collaboration as a Mechanism for More Robust Strategic Classification

Kun Jin*, Ziyuan Huang*, and Mingyan Liu

Abstract— **A conventional strategic classification problem takes on a Stackelberg form: a decision maker commits to a decision rule (e.g., in the form of a binary classifier) and agents best respond to the published decision rule by deciding on an effort level so as to maximize their chance of getting a favorable decision less the cost of the effort. This problem becomes significantly more complex when agents are allowed two types of effort: honest (improvement actions) and dishonest (or cheating/gaming). While the former improves an agent's underlying unobservable states (e.g., certain types of qualification), the latter merely improves an agent's outward observable feature, serving as input to the classifier. Under the natural assumption that honest effort is more costly than cheating, prior work has shown that the decision maker has limited ability to mitigate cheating by simply adjusting the decision rule. In this paper, we consider a *collaboration* mechanism, which the decision maker establishes at a cost and offers to the agents together with the decision rule. In this case, an agent best responds by choosing not only its effort but also whether to participate in the mechanism and if so, with which other agents it wishes to form a connection or collaboration relation. While agents outside the mechanism remain independent of each other, those inside the mechanism are connected to a group of collaborators and enjoy positive externality in the form of a boost in their observable features and consequently enhanced probability of a favorable decision outcome. We show how the collaboration mechanism can induce agents to participate and take improvement actions over gaming and how it can benefit both parties. We also discuss the social value of the system, including social welfare, social qualification status, and the mechanism surplus.**

I. INTRODUCTION

Conventional strategic classification studies the interaction between a decision maker (algorithm designer) and individuals/agents subject to the decision outcome, typically formulated as a Stackelberg game. While the former benefits from the accuracy of its decisions, the latter may have an incentive to *game* the algorithm into making favorable but erroneous decisions.

Prior works on addressing such misuses typically commences with designing a classifier that achieve robust accuracy subject to such strategic maneuvering [1]–[7]. However, as noted by [2], a robust classifier often adopts a more conservative boundary, potentially amplifying the social burden on honest agents. This insight underscores the second objective of strategic classification, which seeks classifiers that are both robust in accuracy and conducive to honest behavior among agents. For instance, [8] proposed a decision model incorporating a weighted strategic recourse term in the decision maker's optimization objective. An alternative practice is to delegate the responsibility of promoting honesty to external mechanisms [9]–[12]. The motivation of this approach is demonstrated by [13]–[15] that no classifier alone can incentivize improvement actions when gaming is a more cost-efficient strategy. To the best of our knowledge, existing literature primarily focuses on transferable, e.g., tax- or subsidy-based, mechanisms. For instance, [12] examined the social impact of subsidizing interventions by the decision maker for disadvantaged groups, and [14] studied the effects of a subsidy-based mechanism introduced by a third party. Both studies revealed occasional adverse effects of transferable mechanisms on inequality gaps. While in the same line of research, our study introduces a non-transferable *collaboration*-based mechanism.

Our proposed mechanism can be viewed as a platform/environment in which agents can form collaborations (e.g., working on a team project, co-authoring a paper), which can lead to positive changes in their underlying attributes. We will show in what sense this mechanism leads to more robust strategic classification.

Specifically, the decision maker is the first mover, committing to and publishing a classifier as well as the collaboration mechanism/platform; this is followed by the simultaneous moves of N agents best responding to the classifier and the mechanism. The classifier takes as input an agent's *observable* feature and outputs a binary decision that impacts the agent's utility. The collaboration mechanism is designed and established by the decision maker at a cost.

Participation in the mechanism is voluntary and free of charge. By opting in, agents form bi-directional connections with other opt-in agents (i.e., agents decide with whom they will collaborate/team up). The mechanism sets an upper bound on how many total connections an agent can have as well as their strength, which is a positive constant uniform across all formed connections (this could be indicative of the nature of the team projects that determines how much collaboration is required/allowed). The agents thus collectively decide the formed network. This then gives rise to a positive network effect, where an agent's observable feature is a function of not only its own attribute but those of its collaborators.

To capture the agent's ability to both game the decision rule and make real change, we assume each agent has an endowed *pre-response* attribute (endowed private information), that is causal [7] to a set of observable features as well as the agent's true label, also referred to as its *qualification status*. An agent can exert effort to improve its attribute, referred to as *improvement* (or honest effort), thereby improving its features and its underlying qualification, or employ non-

Kun Jin, Ziyuan Huang, and Mingyan Liu are with the Electrical and Computer Engineering Department, University of Michigan, Ann Arbor, MI 48109, USA; e-mail: {kunj, ziyuanh, mingyan} @umich.edu.

979-8-3503-0123-6/23/$31.00 ©2023 IEEE

causal schemes to improve only its features without changing its attribute [7], referred to as *gaming* (or cheating, or dishonest effort), or both. Both actions are costly, though gaming is generally assumed to be (much) cheaper than improvement [2].

This paper sets out to demonstrate that meaningful team work can offset the cost difference between gaming and improvement. As a motivating example, we note that the outcome of team projects (e.g., solar car racing, First Robotics competition, or research papers with multiple co-authors) is frequently used as features in real-world classification tasks such as employment, admissions, and scholarship decisions. Arguably, all participants of such a team project benefit from the project's outcome, which relies on team members' attributes and genuine effort. Accordingly, we will assume that agents, once in the mechanism, reveal their endowment and post-response attributes to other agents, and as a result their final observable features become independent of their gaming actions.

Our main contributions are:

1) We propose a collaboration mechanism formulated as a Stackelberg game, where the simultaneous second movers best respond by forming a network (Sec. II). This formulation substantially enriches the literature on both strategic classification and network games.
2) We establish the existence of a graph-formation equilibrium in the second stage, under the assumption of mutual agreement and proper tie-breaking. We identify a subset of equilibrium graphs as regular graphs and proposed a systematic procedure of constructing them, shedding lights on the presence of "local circles" in real collaboration networks (Sec. III).
3) We show that, at equilibrium, each agent's post-response attribute clears an "augmented virtual threshold". This can be viewed as a successful incentivization of improvement actions. We also provide an argument for the equilibrium graph topology using score monotonicity of generalized Katz centrality.

II. PROBLEM FORMULATION

A. Conventional Strategic Interactions (CO)

We start by introducing the *conventional strategic* (CO) setting where no incentive mechanisms are implemented [14]. Consider a Stackelberg game between a decision maker as the first mover and a set of N agents as the second movers. Each agent, indexed by $i \in [N]$ with $[N] := \{1, \cdots, N\}$, is endowed with a **pre-response attribute** $\mathbf{x}^{(i)} \in \mathbb{R}^K_{\geq 0}$, where $\mathbf{x}^{(i)} \sim p_x$ are *i.i.d.* for all $i \in [N]$; i can take action $\mathbf{a}^{(i)} = [\mathbf{a}^{(i)}_+, \mathbf{a}^{(i)}_-] \in \mathbb{R}^M_{\geq 0}$ on $M := M_+ + M_-$ dimensions, where $\mathbf{a}^{(i)}_+ \in \mathbb{R}^{M_+}$ (resp. $\mathbf{a}^{(i)}_- \in \mathbb{R}^{M_-}$) denotes the improvement (resp. gaming) action.

We define a projection matrix $P = [P_+, P_-] \in \mathbb{R}^{K \times M}$, where $P_+ \in \mathbb{R}^{K \times M_+}, P_- \in \mathbb{R}^{K \times M_-}$. Then the i-th agent's **post-response attribute** is

$$\tilde{\mathbf{x}}^{(i)} := \mathbf{x}^{(i)} + P_+ \mathbf{a}^{(i)}_+, \tag{1}$$

which represents the true state of the agent after the improvement effort. Let the agent's **pre-response label** or **qualification** $y^{(i)} \in \{0, 1\}$ (resp. **post-response label** or **qualification** $\tilde{y}^{(i)} \in \{0, 1\}$) be determined by its pre-response attribute $x^{(i)}$ (resp. post-response attribute $\tilde{x}^{(i)}$) through the following relationship similar to that used in [12], [14]:

$$\mathbb{E}\big[y^{(i)} | \mathbf{x}^{(i)}\big] = l(\theta^T \mathbf{x}^{(i)}), \quad \mathbb{E}\big[\tilde{y}^{(i)} | \tilde{\mathbf{x}}^{(i)}\big] = l(\theta^T \tilde{\mathbf{x}}^{(i)}), \tag{2}$$

where we can interpret $l : \mathbb{R} \mapsto [0, 1]$ as a likelihood function which is weakly increasing (l is a step-function in [12]). We assume $\mathbb{E}\big[\tilde{y}^{(i)} | \tilde{\mathbf{x}}^{(i)}\big] \geq \mathbb{E}\big[y^{(i)} | \mathbf{x}^{(i)}\big]$ almost surely, i.e., improvement actions do not worsen one's qualification. Both l and θ are assumed to be public knowledge.

The i-th agent's **post-response observable feature** is given by

$$\mathbf{z}^{(i)} := \mathbf{x}^{(i)} + P_+ \mathbf{a}^{(i)}_+ + P_- \mathbf{a}^{(i)}_-. \tag{3}$$

The decision maker chooses a **linear classifier (decision rule)** $f : \mathbb{R}^K \mapsto \{0, 1\}$, taking as input an agent's post-response observable feature and outputs a binary decision outcome, i.e.,

$$f(\mathbf{z}^{(i)}) := \mathbf{1}_{\{\mathbf{w}^T \mathbf{z}^{(i)} \geq \tau\}}, \tag{4}$$

where the decision maker controls both $\mathbf{w} \in \mathbb{R}^K_{\geq 0}$, the coefficient vector, and $\tau \geq 0$, the threshold. The utility function of agent i is given by:

$$u^{(i)}(\mathbf{a}^{(i)}; \mathbf{x}^{(i)}) := f(\mathbf{z}^{(i)}) - h(\mathbf{a}^{(i)}), \tag{5}$$

whereby the agent benefits from the decision outcome but pays an action cost $h(\mathbf{a}^{(i)}) := (\mathbf{c}^{(i)})^T \mathbf{a}^{(i)}$, where $\mathbf{c}^{(i)}$ is the **cost profile** of agent i.

Assumption 1: $\mathbf{c}^{(i)} = \gamma_i \mathbf{c}$, $\forall i \in [N]$, where $\gamma_i > 0$.

Assumption 1 says that the relative cost between different actions is the same for all agents. Let $\gamma_i \sim p_\gamma$ i.i.d. for all $i \in [N]$, and denote the joint distribution as $(\gamma_i, \mathbf{x}^{(i)}) \sim p_{\gamma,x}$ whose marginal distributions are denoted as p_γ and p_x.

Assume the decision maker knows the agents' action space, utility function Eq. (5), and the distribution $p_{\gamma,x}$, but does not know the realizations of γ_i and $x^{(i)}$. The decision maker anticipates the agents' best response to the classifier at a population level and uses backward induction to design the optimal classifier ($\mathbf{w}^*$, τ^*) to maximize its objective:

$$\begin{aligned} U_{CO}(f) := \ & \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{(\gamma_i, \mathbf{x}^{(i)})} \big[\mathbf{1}_{\{f(\mathbf{z}^{(i)*}) = \tilde{y}^{(i)}\}}\big] \\ = \ & \mathbb{E}_{(\gamma_i, \mathbf{x}^{(i)})} \big[\mathbf{1}_{\{f(\mathbf{z}^{(i)*}) = \tilde{y}^{(i)}\}}\big], \end{aligned} \tag{6}$$

where $\mathbf{z}^{(i)*} = \mathbf{x}^{(i)} + P\mathbf{a}^{(i)*}_{co}(\mathbf{x}^{(i)})$ is the i-th agent's post-response observable feature (3) at the second-stage equilibrium, and $\mathbf{a}^{(i)*}_{co}(\mathbf{x}^{(i)}) \in \mathbb{R}^M$ the agent's equilibrium action in the CO system when endowed with $\mathbf{x}^{(i)}$. The second equality follows from the i.i.d. assumption.

[14] shows that, in realistic situations where gaming is more cost-effective than improvement actions, there is no decision rule that can incentivize improvement and the decision maker's optimal strategy is to choose $\mathbf{w} = \theta$. Formally,

under Assumption 1, let κ_s denote the **substitutability** of action dimension $s \in [M]$ [10], [13],

$$\kappa_s := \min_{\mathbf{a} \in \mathbb{R}^M, \mathbf{a} \geq 0} \frac{\mathbf{c}^T \mathbf{a}}{c_s}, \quad \text{s.t. } P\mathbf{a} - \mathbf{p}_s \geq 0, \tag{7}$$

where $\mathbf{p}_s$ is the s-th column of P. Note that by assumption 1, κ_s is the same for all agents. The intuition of substitutability is provided in [10], [14]. Specifically, if $\kappa_s = 1$, then there exists a $\mathbf{w}$ that can incentivize action on dimension s, and such $\mathbf{w}$ can be found in polynomial time. Conversely, if $\kappa_s < 1, \forall s \leq M_+$, then there always exist linear combinations of gaming actions that weakly dominate every improvement action for any choice of $\mathbf{w}$.

Assumption 2: (Theorem 3.3 [14]) Assume $\kappa_i < 1, \forall i \leq M_+$. Then there is no f that can incentivize improvement actions, and the decision maker's optimal CO strategy f_C^* satisfies $\mathbf{w}^* = \theta$.

B. Augmented Strategic Interactions (AU)

Compared to the conventional case, the *augmented strategic* setting includes a collaboration mechanism in the first stage of the Stackelberg game. As the classifier itself cannot sufficiently incentivize improvement, we are interested in the extent to which the collaboration mechanism together with the classifier can help incentivize improvement actions.

Methodologically, we will fix the decision rule at the CO optimal solution, i.e., $\mathbf{w} \equiv \theta$, and fix τ as well, so that we can better illustrate the mechanism's impact on the system. The decision maker then specifies a mechanism given by a connection strength $\alpha > 0$, which is applied to all established connections, and a maximum number of collaborators allowed $D \in \mathbb{N}$ for each participant.

After the classifier f and the mechanism (α, D) are announced, the N agents best-respond in a simultaneous-move game by determining (1) whether to participate in the mechanism, (2) if yes, which other agents to form connection with, and (3) what effort to exert. In this second-stage game, the realizations of $\mathbf{x}^{(i)}$ and γ_i for $i \in [N]$ are known to all, i.e., in the second stage the N agents play a simultaneous-move non-cooperative game with complete information.

As indicated above, participation in the mechanism is free and voluntary. However, once agents choose to participate, they are restricted to improvement actions only. We also assume that changes (formation or deletion) to a pairwise connection follows *mutual agreement*, i.e., it has to be favored by both agents involved.

The result of the second-stage game is represented as a weighted graph (mutual agreement implies symmetry, thus directed and undirected graphs are analytically equivalent), where $G = G^T \in \mathbb{R}_{\geq 0}^{L \times L}$ denotes the corresponding adjacency matrix (G is also frequently referred to as a graph with a slight abuse of terminology), L is the number of mechanism participants, and G's entries are $g_{ij} = g_{ji} = \alpha > 0, \forall i, j \in [L]$ with diagonal entries $g_{ii} := 0, \forall i \in [L]$. We denote by $\mathscr{N}_i = \{j | j \neq i, g_{ij} > 0\}$ the set of agent indices of the i-th agent's neighbors/collaborators. We call such a graph a (α, D) graph.

We note that the parameters (α, D) is chosen by the decision maker, but the actual topology of G is decided by the participating agents. One topology of interest is one where no additional edges can be established. We call this the *maximum collaboration* property formally defined as follows.

Definition 1: The (α, D) graph G achieves **maximum collaboration** if $g_{ij} = 0 \Rightarrow \min\{D_i, D_j\} = D$, where D_i is the degree of agent i on graph G; i.e., for any two agents i, j that are not presently connected, at least one agent has already reached the maximum number of connections D.

Maximum collaboration characterizes a graph's connectivity. When $N \geq D + 1$ and $N \cdot D$ is even, a graph satisfying maximum collaboration is a D-regular graph (a graph where all nodes have degree D). This kind of graph plays a crucial role in determining the Nash equilibrium among mechanism participants, as we will show in Section III.

Assumption 3: (Bounded Externality) $\alpha D < 1$.

This assumption implies that $I + G$ is invertible, a common assumption in the network games literature [13], [16].

The mechanism affects the observable feature of each participant. Given the graph G over the set of participants $\mathscr{P}$, agent i's **in-mechanism observable feature** $\bar{\mathbf{z}}^{(i)}$ is given by

$$\bar{\mathbf{z}}^{(i)} := \tilde{\mathbf{x}}^{(i)} + \sum_{j \in \mathscr{N}_i} g_{ij} \tilde{\mathbf{x}}^{(j)}, \quad \forall i \in \mathscr{P}. \tag{8}$$

Eqn (8) indicates that each participant's observable feature benefits from its own post-response attribute as well as its neighbors' post-response attributes (positive externality). We note that one's post-response attribute is unaffected by one's neighbors. Moreover, gaming actions cannot improve any participant's in-mechanism feature. Previous works on graph neural networks [17], [18] and interdependent strategic classification [13] use similar feature aggregation methods as Eq. (8).

When participating in the mechanism, the utility function of agent i is given by:

$$\tilde{u}^{(i)}(\mathbf{a}^{(i)}, \mathbf{a}^{(-i)}; \mathbf{x}^{(i)}, \mathbf{x}^{(-i)}) := f(\bar{\mathbf{z}}^{(i)}) - h(\mathbf{a}^{(i)}), \tag{9}$$

where $\mathbf{a}^{(-i)}, \mathbf{x}^{(-i)}$ denotes other participants' actions and endowments.

In the AU system, the decision maker's utility is

$$U_{AU}(f) := \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{(\gamma_i, \mathbf{x}^{(i)})} \left[\mathbf{1}_{\{f(\bar{\mathbf{z}}^{(i)*}) = \tilde{y}^{(i)}\}} \right] - \varsigma H(G), \tag{10}$$

where $H(G)$ is the cost of implementing the mechanism and $\bar{\mathbf{z}}^{(i)*}$ is the i-th agent's in-mechanism observable feature (8) at the second-stage equilibrium $\mathbf{a}^{(i)*}$ (precisely defined below). Notice that the latter is also a function of $\mathbf{x}^{(i)}$ and γ_i. The scalar $\varsigma > 0$ describes the relative importance between the classifier accuracy and the mechanism cost. We assume

$$H(G) := \|G\|_F^2 = \alpha^2 \sum_{i \neq j} \mathbf{1}_{\{g_{ij} \neq 0\}}. \tag{11}$$

As an example, if G is D-regular, then $H(G) = ND\alpha^2$.

For the solution concept under the mechanism, we constrain ourselves to the Nash equilibrium satisfying the mutual agreement property as defined below.

Definition 2: The tuple $(G, \{\mathbf{a}^{(i)*}\}_{i\in V})$ is a **Nash equilibrium satisfying mutual agreement (NEMA)** over the agent set $V \subseteq [N]$ if: (i) G is an (α, D) graph with vertex set V; (ii) $\{\mathbf{a}^{(i)*}\}_{i\in V}$ is a Nash equilibrium given the network G; (iii) adding an edge $i \leftrightarrow j$ in G is either not allowed, or does not yield a Nash equilibrium that is strictly preferred by both i and j; (iv) deletion of an edge $i \leftrightarrow j$ in G yields no Nash equilibrium strictly preferred by either i or j.

In other words, $(G, \{\mathbf{a}^{(i)*}\}_{i\in V})$ constitutes a Nash equilibrium satisfying mutual agreement if every agent has no incentive to unilaterally deviate from its action $\mathbf{a}^{(i)*}$ and every pair of agents has no incentive to unilaterally change their mutual connection or lack thereof.

Definition 3: The tuple $(\mathscr{P}, G, \{\mathbf{a}^{(i)*}\}_{i\in\mathscr{P}}, \{\mathbf{a}_{co}^{(i)*}\}_{i\notin\mathscr{P}})$ is a **second-stage equilibrium** if the following holds: (i) $(G, \{\mathbf{a}^{(i)*}\}_{i\in\mathscr{P}})$ is an NEMA over $\mathscr{P} \subseteq [N]$; (ii) $\{\mathbf{a}_{co}^{(i)*}\}_{i\notin\mathscr{P}}$ is the CO optimal actions for each agent not in $\mathscr{P}$; (iii) agents in $\mathscr{P}$ satisfy voluntary participation, i.e., for each $i \in \mathscr{P}$, $\tilde{u}^{(i)*} \geq u^{(i)*}$;[1] (iv) agents not in $\mathscr{P}$ also satisfy *strict* voluntary participation, i.e., for each $i \notin \mathscr{P}$, participating in the mechanism yields no NEMA in which $\tilde{u}^{(i)*} \geq u^{(i)*}$.[2]

This definition reveals a fundamental difficulty of equilibrium characterization in this problem. First of all, each second-stage equilibrium may have a different set of participants $\mathscr{P}$. Secondly, there could be agents that are better off in some but not all second-stage equilibrium. Both complicate the characterization of the incentives of agents to opt in/out. We show in the next section that this can be alleviated under some appropriate tie-breaking rules.

III. Second-Stage Equilibrium Analysis

In this section, we study the agents' optimal strategies or best response in the second stage given the decision rule and the mechanism. They compare the equilibrium output between participation and unilaterally opting out, and choose whichever yields higher equilibrium utilities.

A. Equilibrium Actions with Fixed Graph

By Assumption 2 and Lemma 3.2 in [14], every agent only chooses non-zero action in the direction k_A (resp. k_C) in the AU (resp. CO) setting. Thus, we consider the following simplification of notations:

$$\begin{aligned} \bar{x}_i &:= \theta^T \mathbf{x}^{(i)}, \quad \bar{a}_i := [P^T\theta]_{k_A} \cdot \mathbf{a}_{k_A}^{(i)}, \\ \bar{c} &:= c_{k_A}, \quad \bar{c}_i := \frac{\gamma_i}{[P^T\theta]_{k_A}} \bar{c}, \end{aligned} \tag{12}$$

where $\bar{x}_i \sim \bar{p}_x$, which is derived from p_x; also denote $(\gamma_i, \bar{x}_i) \sim \bar{p}_{\gamma,\bar{x}}$. For clarity of presentation, we refer to $\bar{x}_i$ and $\bar{a}_i$ as well as their vectorized versions directly as endowments and actions, respectively, for the rest of the paper. Also, without loss of generality, we index the L participants with $1, 2, \ldots, L$ such that $\bar{x}_1 \geq \bar{x}_2 \geq \cdots \geq \bar{x}_L$ and let $\bar{\mathbf{x}} := (\bar{x}_i)_{i\in[L]}$, $\bar{\mathbf{a}} := (\bar{a}_i)_{i\in[L]}$ be the global vectors and $\bar{\mathbf{x}}_{-i} := (\bar{x}_j)_{j\in[L]\setminus\{i\}}$, $\bar{\mathbf{a}}_{-i} := (\bar{a}_j)_{j\in[L]\setminus\{i\}}$ be the vectors of all other participants. Then, $\theta^T\tilde{x}^{(i)} = \bar{x}_i + \bar{a}_i$ and an in-mechanism participant i has a utility of

$$\tilde{u}_i(\bar{a}_i, \bar{\mathbf{a}}_{-i}; \bar{\mathbf{x}}) = \mathbf{1}_{\{\theta^T\tilde{x}^{(i)} + \alpha \sum\limits_{l\in\mathscr{N}_i} \theta^T\tilde{x}^{(j)} \geq \tau\}} - \bar{c}_i \cdot \bar{a}_i. \tag{13}$$

Assumption 4: In the case of multiple best-response actions, an in-mechanism participant always selects the action with the highest post-response attribute, i.e., $\bar{x}_i + \bar{a}_i$.

The above assumption is a natural tie-breaking rule as it favors the highest real improvement with the same utility.

Lemma 3.1: Suppose Assumption 1-4 hold. Given any (α, D) graph G among participants, there exists a unique Nash equilibrium (NE) profile $\bar{\mathbf{a}}^*$ such that[3]

$$\bar{a}_i^* = \max\left\{0,\ \tau - \bar{x}_i - \alpha \sum_{j\in\mathscr{N}_i} (\bar{x}_j + \bar{a}_j^*)\right\},\ \forall i \in [L]. \tag{14}$$

Denoting by $\tilde{\tau} := \tau/(1+\alpha D)$. We partition the agents into three zones according to their pre-response attributes

$$\begin{aligned} \mathscr{X}_1 &:= \{\mathbf{x}^{(i)} \mid \bar{x}_i \geq \tau\}, \\ \mathscr{X}_2 &:= \{\mathbf{x}^{(i)} \mid \bar{x}_i \in [\tilde{\tau},\ \tau)\}, \\ \mathscr{X}_3 &:= \{\mathbf{x}^{(i)} \mid \bar{x}_i \in [0,\ \tilde{\tau})\}\ . \end{aligned} \tag{15}$$

These partitions divide the non-negative real line into 3 zones, ordered in decreasing endowment from $\mathscr{X}_1$ to $\mathscr{X}_3$. Agents in $\mathscr{X}_1$ have already met the decision rule's requirement τ and thus, they would always take zero actions, irrespective of their participation in the mechanism. In the case of participation, they are indifferent to establishing any connection with other agents as they have already achieved the maximum utility 1. Agents in $\mathscr{X}_2$ possess lower endowments than the required level τ. However, they can form a D-regular graph within their zone (if $|\mathscr{X}_2| \cdot D$ is odd, they can ask an agent in $\mathscr{X}_1$ to form a D-regular graph) so that everyone's in-mechanism equilibrium action is zero, in which case they can also achieve their maximum utility 1. On the contrary, agents in $\mathscr{X}_3$ may not be able to reach zero equilibrium actions just by collaborating within themselves.

B. Tie-breaking Rule and Second-Stage Equilibrium

Assumption 5 (Tie-breaking Rule): (i) agents in $\mathscr{X}_1$ participate only when suggested by the decision maker; (ii) agents in $\mathscr{X}_2$ can only collaborate with other agents in $\mathscr{X}_2$ or *upgraded agents* from $\mathscr{X}_3$ (see (iv)); (iii) if the number of the sum of $\mathscr{X}_2$ agents and upgraded agents is odd, the decision maker randomly selects an $\mathscr{X}_1$ agent to join the mechanism and collaborate with others; and (iv) any agent in $\mathscr{X}_3$ can become an *upgraded agent* by choosing an action no less than $\tilde{\tau} - \bar{x}_i$ i.e., $\bar{x}_i + \bar{a}_i \geq \tilde{\tau}$.

Intuitively, these are valid tie-breaking suggestions for the following reasons. (i) and (iii) are indeed tie-breaking rules since agents in $\mathscr{X}_1$ are indifferent to making connections as discussed above. (ii) works since the upgraded agents can be viewed as newly arrived $\mathscr{X}_2$ agents with minimum endowment $\tilde{\tau}$ and, as we discussed above, the participants from $\mathscr{X}_2$

[1] Without loss of ambiguity, we simplify the AU and CO equilibrium utility for the i-th agent respectively as $\tilde{u}^{(i)*}$ and $u^{(i)*}$

[2] In case of a tie, we assume the agent would always choose to opt out. That is, participation only occurs when the agent can strictly benefit from the mechanism.

[3] Please find all proofs in the online appendix [19].

can always get the positive decision outcome when having D collaborators among themselves, i.e., they are always weakly better off in the mechanism following the suggestions. Part (iv) simply defines who is considered an upgraded agent and does not influence tie-breaking; participation by agents in $\mathscr{X}_3$ is still voluntary.

This tie-breaking rule combined with our definition ensures that agents in $\mathscr{X}_2$ can always get the highest possible utility by forming D-regular graphs. It also allows some agents to be upgraded by collaborating with sufficiently competitive peers if they agree to exert a high enough effort. This is akin to rewarding hard-working agents and, intuitively, reduces free-riding of $\mathscr{X}_3$ agents, which makes the mechanism design problem tractable.

Define the following quantities

$$h_A^*(\mathbf{x}^{(i)}) := \frac{c^{(i)}_{k_A}}{(P^T\theta)_{k_A}}(\tilde{\tau}-\bar{x}_i) = \bar{c}_i(\tilde{\tau}-\bar{x}_i), \tag{16}$$

$$h_C^*(\mathbf{x}^{(i)}) := \frac{c^{(i)}_{k_C}}{(P^T\theta)_{k_C}}(\tau-\bar{x}_i), \tag{17}$$

and denote the following subset of $\mathscr{X}_3$ by

$$\mathscr{X}_3^p := \{i \in \mathscr{X}_3 \mid h_A^*(\mathbf{x}^{(i)}) < \min\{1,\ h_C^*(\mathbf{x}^{(i)})\}\}.^4 \tag{18}$$

Eq. (16) and (17) define, for an agent $i \in \mathscr{X}_3$, the equilibrium action cost of participation and opting out, respectively. Eq. (18) identifies the set of in-mechanism participants from $\mathscr{X}_3$. We summarize this point in the theorem below.

Theorem 3.2: Under Assumptions 1-5, $\mathscr{P} := \mathscr{X}_2 \cup \mathscr{X}_3^p$ defines the set of in-mechanism participants, which is invariant to specific second-stage equilibrium outcomes.[5] Moreover, the set of second-stage equilibrium, denoted by $\mathscr{S}_{NE}$, is non-empty, and

1) (Decision Outcome) for each equilibrium in $\mathscr{S}_{NE}$, all agents in $\mathscr{P}$ receive positive decision outcomes;
2) (Upgraded Agents) for each equilibrium in $\mathscr{S}_{NE}$, every agent in $\mathscr{X}_3^p$ becomes an upgraded agent;
3) (Equilibrium Graph) each D-regular graph among agents in $\mathscr{X}_2 \cup \mathscr{X}_3^p$ corresponds to some equilibrium in $\mathscr{S}_{NE}$, i.e., the set of equilibrium graphs include all D-regular graphs;
4) (Equilibrium Action) for each equilibrium in 3), $\bar{a}_i^* = 0$ for all $i \in \mathscr{X}_2$ and $\bar{a}_i^* = \tilde{\tau} - \bar{x}_i$ for all $i \in \mathscr{X}_3^p$.

Notice that D-regular graphs yield the highest utility for every agent. For technical reasons, we impose another tie-breaking rule where the decision maker favors D-regular equilibria over others.

Assumption 6: The decision maker favors an equilibrium with D-regular graphs over others, and agents are willing to break-tie in favor of the decision maker.

An interesting observation from the above theorem is that there is an observable pattern of symmetry in agents' post-response qualification status, i.e., $\bar{x}_i + \bar{a}_i^* = \tilde{\tau}$ for $\mathscr{X}_3$ agents, where $\tilde{\tau}$ can be thought of as an "augmented virtual threshold" created by the mechanism. It also indicates that in the second-stage equilibrium, every agent provides the same externality to all its collaborators. Thus, they are indifferent to the identity of collaborators and only care about their zones. This symmetry and equivalence result benefits both the agents and the mechanism designer since finding their respective optimal strategies becomes much simpler.

C. Two Examples

Example 1: Consider the following two situations.

(1) (Same Cost) $\mathbf{c}^{(i)} = \mathbf{c}$, $\forall i \in [N]$;
(2) (Same Endowment) $\bar{x}_i = \hat{x} \leq \frac{\tau}{2} < \tilde{\tau}$, $\forall i \in [N]$.

Example 1-(1) models the situation where the action costs can be standardized e.g., in time and monetary terms such as the number of credits taken in a specific field. But each agent is endowed with different skills or experiences such that some agents get positive decision outcomes easier than others. On the other hand, Example 1-(2) models the situation where agents do not possess advantage over each other (e.g., no prior experience), but their cost of action can differ.

Definition 4: A (α, D) graph is a $(D+1)$**-ordered clique graph** w.r.t. the partial order $\preceq$ over its vertex set if the first (left-most) $D+1$ vertices form a complete sub-graph, the next first $D+1$ vertices form another complete sub-graph, and so on.

Note that any $(D+1)$-ordered clique graph obviously satisfies maximum collaboration. We can show that the equilibrium graphs of the two scenarios in Example 1 can be found by simply creating $(D+1)$-ordered clique graphs w.r.t. some appropriate partial order. This provides a stable graph formation process that helps the mechanism designer from having to enumerate through the exponentially growing number of collaboration options.

Proposition 3.3: Let $\preceq_x$ and $\preceq_\gamma$ be partial orders such that $i \preceq_x j$ if $\bar{x}_i \geq \bar{x}_j$ and $i \preceq_\gamma j$ if $\gamma_i \leq \gamma_2$. Then, under Assumptions 1-6, a $(D+1)$-ordered clique graph over $\mathscr{X}_2 \cup \mathscr{X}_3^p$ w.r.t. $\preceq_x$ constitutes a second-stage equilibrium of Example 1-(1) and that w.r.t. $\preceq_\gamma$ a second-stage equilibrium of Example 1-(2).

For the first $D+1$ participants, forming a $D+1$ clique among themselves guarantees the highest participation utility for each of them. Then, by induction, after removing the first $D+1$ participants (since they have all reached the collaboration limit), the next highest endowed $D+1$ participants may form another $D+1$ clique among themselves. This describes a process of forming the $(D+1)$-ordered clique graph, the stability of which is guaranteed because every participant attains its maximum utility, given the fact that all its precedents have already been grouped. The result of this process is that the graph consists of multiple "local circles" which is reminiscent of real-world cases of research teams and collaborative course projects.

When costs are equal, agents with higher endowments are preferred; when endowments are the same, agents with

[4] With a slight abuse of notation, we will not distinguish between $i \in \mathscr{X}_3$ and $\mathbf{x}^{(i)} \in \mathscr{X}_3$ in the rest of the paper. The meaning of both should be clear from the context.

[5] For simplicity, we assume an arbitrary $\mathscr{X}_1$ agent is automatically inserted into $\mathscr{X}_2$ in case of odd number of participants. Thus, the cardinality of the set of participants $\mathscr{P}$ is always even.

lower costs are preferred. Intuitively, in both cases, the top indexed agents can generate weakly larger positive externalities to their collaborators, so everyone will weakly prefer to collaborate with top indexed agents. This is also true for the top indexed agents themselves and thus the circle effect. We observe that in the real world, it is often the case that scholars of similar traits (by whatever definition) collaborate more frequently with other similar scholars (by the same definition).

D. Maximum Collaboration from a Centrality Perspective

If all agents' endowments satisfy $\bar{x}_i \leq (1-\alpha D)\tau$, then the best-response relationship in Eq. (14) can be written in a compact form as

$$\bar{\mathbf{x}}+\bar{\mathbf{a}} = \tau(I+G)^{-1}\mathbf{1}. \tag{19}$$

For any given graph G, we can then obtain the following intermediate result that may be of independent interest.

Theorem 3.4: For a symmetric matrix G, such that $g_{ij} = g_{ji} \in \{0, \alpha\}$ for all $i \neq j$ and $g_{ii} = 0$ for all i, define a score vector

$$\mathbf{r} := (I+G)^{-1}\mathbf{1}. \tag{20}$$

Furthermore, define the companion matrix $\tilde{G} := G+\alpha(\mathbf{e}_i\mathbf{e}_j^T + \mathbf{e}_j\mathbf{e}_i^T)$, where $\mathbf{e}_k$ denotes the k^{th} standard Euclidean basis, which adds a pair of edges $(i,j),(j,i)$ to G. Define a new score vector

$$\mathbf{s} := (I+\tilde{G})^{-1}\mathbf{1}. \tag{21}$$

Define $D := \|\tilde{G}\|_\infty/\alpha$. Notice that D is fixed regardless of the value of α. If $\alpha \in \left(0, \frac{1}{1+D}\right)$, then $r_i > s_i > 0$ and $r_j > s_j > 0$.

Comparing Eqn (19) with (20), we note that the scores, i.e., entries of the score vectors $\mathbf{r}, \mathbf{s}$, measure agents' action costs, and thus agent i's utility decreases in its corresponding score. The game-theoretic interpretation of Thm. 3.4 is that for two disconnected participants in a graph G, if adding an edge between them yields an (α, D) graph $\tilde{G}$, then they will both strictly prefer the new graph $\tilde{G}$ since their equilibrium utilities will both improve by choosing smaller actions for the same positive decision outcome. However, this theorem only holds when α is less than $\frac{1}{D+1}$, slightly stronger than the $\alpha < \frac{1}{D}$ condition in Assumption 3, although they are asymptotically similar when D becomes large. We also note that this result does not require the tie-breaking rule (Assumption 5).

We can also think of Thm. 3.4 as the score monotonicity of the generalized Katz centrality measure on negative graphs with bounded node degrees (less than D) and uniform but limited edge weights (less than $\frac{1}{D+1}$). The Katz centrality measure for the binary matrix G with parameter $\alpha > 0$ is defined as $\mathbf{s}' := (I-G^T)^{-1}\mathbf{1}-\mathbf{1}$. Suppose $g_{ij} = g_{ji} = 0$, define $\tilde{G}$ the same as above, and let $\mathbf{r}' := (I-\tilde{G}^T)^{-1}\mathbf{1}-\mathbf{1}$. Then if $\alpha \in \left(0, \frac{1}{1+D}\right)$, the Katz centrality is score monotone, i.e., $r'_i > s'_i > 0$ and $r'_j > s'_j > 0$ [20], meaning that both nodes will have increased centrality after adding the edge. In Thm. 3.4, we have negative edge weights and show that adding a pair of edges will decrease the generalized Katz centrality of both nodes. We note that this result itself is a non-trivial finding of the generalized Katz centrality and its derivation is not straightforward compared to the original measure.

IV. ACKNOWLEDGMENTS

This work has been supported by the ARO under contract W911NF1810208.

REFERENCES

[1] M. Hardt, N. Megiddo, C. Papadimitriou, and M. Wootters, "Strategic classification," 01 2016, pp. 111–122.

[2] S. Milli, J. Miller, A. Dragan, and M. Hardt, "The social cost of strategic classification," 01 2019, pp. 230–239.

[3] M. Brückner and T. Scheffer, "Stackelberg games for adversarial prediction problems," 08 2011, pp. 547–555.

[4] M. Brückner, C. Kanzow, and T. Scheffer, "Static prediction games for adversarial learning problems," *The Journal of Machine Learning Research*, vol. 13, pp. 2617–2654, 09 2012.

[5] J. Dong, A. Roth, Z. Schutzman, B. Waggoner, and Z. S. Wu, "Strategic classification from revealed preferences," in *Proceedings of the 2018 ACM Conference on Economics and Computation*, 2018, pp. 55–70.

[6] M. Braverman and S. Garg, "The role of randomness and noise in strategic classification," in *1st Symposium on Foundations of Responsible Computing*, 2020.

[7] J. Miller, S. Milli, and M. Hardt, "Strategic classification is causal modeling in disguise," in *Proceedings of the 37th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, H. D. III and A. Singh, Eds., vol. 119. PMLR, 13–18 Jul 2020, pp. 6917–6926.

[8] Y. Chen, J. Wang, and Y. Liu, "Strategic recourse in linear classification," *arXiv preprint arXiv:2011.00355*, 2020.

[9] N. Haghtalab, N. Immorlica, B. Lucier, and J. Wang, "Maximizing welfare with incentive-aware evaluation mechanisms," 07 2020, pp. 160–166.

[10] J. Kleinberg and M. Raghavan, "How do classifiers induce agents to invest effort strategically?" *ACM Transactions on Economics and Computation*, vol. 8, pp. 1–23, 11 2020.

[11] Y. Shavit, B. Edelman, and B. Axelrod, "Causal strategic linear regression," 2020.

[12] L. Hu, N. Immorlica, and J. Vaughan, "The disparate effects of strategic manipulation," 01 2019, pp. 259–268.

[13] K. Jin, T. Yin, C. A. Kamhoua, and M. Liu, "Network games with strategic machine learning," in *Decision and Game Theory for Security*, B. Bošanský, C. Gonzalez, S. Rass, and A. Sinha, Eds. Cham: Springer International Publishing, 2021, pp. 118–137.

[14] K. Jin, X. Zhang, M. M. Khalili, P. Naghizadeh, and M. Liu, "Incentive mechanisms for strategic classification and regression problems," in *EC '22: The 23rd ACM Conference on Economics and Computation, Boulder, CO, USA, July 11 - 15, 2022*, D. M. Pennock, I. Segal, and S. Seuken, Eds. ACM, 2022, pp. 760–790. [Online]. Available: https://doi.org/10.1145/3490486.3538300

[15] X. Zhang, M. M. Khalili, K. Jin, P. Naghizadeh, and M. Liu, "Fairness interventions as (Dis)Incentives for strategic manipulation," in *Proceedings of the 39th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, Eds., vol. 162. PMLR, 17–23 Jul 2022, pp. 26 239–26 264. [Online]. Available: https://proceedings.mlr.press/v162/zhang22l.html

[16] A. Galeotti, B. Golub, and S. Goyal, "Targeting interventions in networks," *SSRN Electronic Journal*, 10 2017.

[17] T. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," 09 2016.

[18] W. Hamilton, R. Ying, and J. Leskovec, "Inductive representation learning on large graphs," 06 2017.

[19] K. Jin, Z. Huang, and M. Liu. (2023) Appendix for Collaboration as a Mechanism for More Robust Strategic Classification. Available at https://www.dropbox.com/s/bumc5oztyzzpw3f/cdc23apdx.pdf?dl=0.

[20] P. Boldi, F. Furia, and S. Vigna, "Monotonicity in undirected networks," 2022.