Incentive-Compatible Experimental Design

PANOS TOULIS, Harvard University, Department of Statistics
DAVID C. PARKES, Harvard University, SEAS
ELERY PFEFFER, Harvard University, SEAS
JAMES ZOU, Microsoft Research

We consider the design of experiments to evaluate treatments that are administered by self-interested agents, each seeking to achieve the highest evaluation and win the experiment. For example, in an advertising experiment, a company wishes to evaluate two marketing agents in terms of their efficacy in viral marketing, and assign a contract to the winner agent. Contrary to traditional experimental design, this problem has two new implications. First, the experiment induces a game among agents, where each agent can select from multiple versions of the treatment it administers. Second, the action of one agent – selection of *treatment version* – may affect the actions of another agent, with the resulting *strategic interference* complicating the evaluation of agents. An *incentive-compatible experiment design* is one with an equilibrium where each agent selects its *natural action*, which is the action that would maximize the performance of the agent if there was no competition (e.g., expected number of conversions if agent was assigned the contract).

Under a general formulation of experimental design, we identify sufficient conditions that guarantee incentive-compatible experiments. These conditions rely on the existence of statistics that can estimate how agents would perform without competition, and their use in constructing score functions to evaluate the agents. In the setting with no strategic interference, we also study the *power* of the design, i.e., the probability that the best agent wins, and show how to improve the power of incentive-compatible designs. From the technical side, our theory uses a range of statistical methods such as hypothesis testing, variance-stabilizing transformations and the Delta method, all of which rely on asymptotics.

Categories and Subject Descriptors: G.3 [**Experimental design**]; I.2.1 [**Games**]

General Terms: Experimental design, incentive compatibility, strategic interference, hypothesis testing, variance stabilization, viral marketing.

ACM Reference Format:
Panos Toulis, David C. Parkes, Elery Pfeffer and James Zou. Incentive-compatible experimental design. In EC 2015. ACM X, X, Article X (February 2015), 20 pages.
DOI = 10.1145/2764468.2764525 http://doi.acm.org/10.1145/2764468.2764525

1. INTRODUCTION

Experiments are the gold-standard for evaluating the effects of different treatments. The design of experiments is crucial in order to avoid systematic biases and to minimize random errors in the statistical evaluation of treatment effects [Cox and Reid 2000]. There are three fundamental concepts in any experiment design. The *treatment* is a well-defined prescription or set of rules, e.g., a pharmaceutical drug, a marketing campaign, or a new material. The goal of the experiment is to evaluate the effects of different treatments. The *experimental unit* is the indivisible entity that will receive a treatment within the experiment, e.g., a patient, a potential customer, or a factory process. Typically, every unit receives only one treatment, but there are important exceptions as well. The treatment is assigned according to a *treatment assignment rule*

EC'15, June 15–19, 2015, Portland, OR, USA.
http://dx.doi.org/10.1145/2764468.2764525

specified by the design and necessarily involves randomization in order to avoid systematic biases. When a unit receives the treatment it exhibits a measurable *outcome*, e.g., a health assessment, a product purchase or not, or a material failure rate.

Statistical analysis of unit outcomes is necessary for the evaluation of treatments because it accounts for the errors that are inherent to randomization of treatment and the measurement process. A key idea in experimental design is *blocking*. Background information on units is almost always available, e.g., age, gender, socioeconomic status, health status, and so on. If an experimenter believes that units' outcomes vary systematically with respect to such *covariate* information, then it is necessary to block units with respect to the available covariates. Blocking helps to avoid systematic bias and variability that is not of scientific interest. The unofficial mantra in experimental design is "block what you can and randomize what you cannot" [Box et al. 1978]).

To illustrate, consider the example of a new flu shot. A pharmaceutical company, the *experimenter*, wants to compare between the new flu shot and a baseline that is currently in the market. The treatments are the two flu shots. The experimenter has a set of volunteer patients who form the set of experimental units. When a unit receives a treatment the outcome is whether the unit got flu or not for the three months following the treatment. As a treatment assignment rule, the experimenter could simply give the new flu shot to half of the patients at random, and give the baseline to the other half. However, the outcomes could be confounded with factors such as age (older people are more vulnerable to flu), geography (urban areas are more crowded and possibly more contagious), occupation, and so on. In a blocking design, the experimenter could block the population based on age and occupation, and perform the randomization within blocks.

There are two crucial assumptions in experimental design and the related topic of *causal inference*, collectively known as the *stable unit treatment value assumption* (SUTVA) [Rubin 1980]. First, there are *no hidden versions* of a treatment. In the previous example, this means that there are no strong or weak versions of the new flu shot. Otherwise, the outcomes would be confounded with the hidden version of the treatment. This is an important problem, especially in social science studies. For example, in an educational study a new treatment could be a new type of curriculum, however a possible hidden version of the treatment is the delivery method by each teacher. A second crucial assumption is that of *no interference* among experimental units. Interference is present when the treatment assignment on one unit affects the outcome of another unit. In the flu shot example, a unit that is not vaccinated is still protected when the friends of the unit are vaccinated. Neither of these assumptions hold in our setting.

We introduce the idea of *incentive-compatible experimental design* in the context of viral marketing.[1] Imagine a company that designs a test to determine which of two vendors has the best algorithm for running an advertising campaign. The firm uses randomization to prevent systematic bias, and defines a criterion for success; e.g., the number of conversions over a two week period. The winning vendor is promised a one-year contract with the firm running the test. One challenge in this setting is that the vendors might deviate from how they would normally run a campaign, trying to win the test. For example, a lower quality vendor may try to follow a more aggressive strategy, hoping to get lucky. This is a problem for the firm designing the test, who wants to get an unbiased estimate of the usual performance of the vendor. Another challenge comes from interference between the participants. In viral marketing, for

[1]An early extended abstract of this paper was presented in the Conference on Digital Experimentation at MIT [Toulis et al. 2014].

example, one vendor may try to free-ride on word-of-mouth effects that come from another vendor.

1.1. Results

A first contribution of the present paper is to formalize this problem of incentive-compatible experimental design. The difference with traditional experimental design is that, in our framework, strategic agents administer the treatments to be evaluated, and each agent can select from multiple treatment versions. In this way, the experiment induces a non-cooperative game. The action available to an agent in the resulting *treatment selection game* is the version of the treatment that the agent will administer to its assigned units. The experimenter has a *performance metric* to evaluate each treatment version. This is the quantity of interest to the experimenter. Each agent has a *natural action*, which is the action that maximizes its performance, and is assumed to be the way the agent would act if not competing in the game. The *quality* of an agent is the maximum value of the performance metric, achieved when the agent plays the natural action without competition from other agents. The goal of the experimenter is to design an experiment to estimate the agent of highest quality. An *incentive-compatible experiment design* is one with an equilibrium in which each agent's best response is to select the treatment version corresponding to its natural action. We will focus on dominant-strategy equilibrium in this paper.

We show that incentive-compatible designs are possible when an *identifying statistic* exists that can estimate the quality difference between agents (Theorem 3.2). Critically, the variance of such a statistic has to be less sensitive to agent actions than its expected value, otherwise an agent can take advantage of the variance of the statistic. Under a no interference assumption, a class of incentive-compatible designs can be constructed through a *variance-stabilizing* transformation (Theorem 4.2), which makes the variance of the identifying statistic insensitive to agent actions; a worse agent cannot hope to increase its chances by being more aggressive. This leads to results that may sound counter-intuitive. For example, in a viral marketing application where performance is the expected number of conversions, and where higher expected conversions also correspond to increasingly higher risks, it is not incentive-compatible to select as the winner the agent with the highest average performance; rather, it is incentive-compatible to select as the winner the agent with the lowest reciprocal of average performance (see Example 2(d)).

Identifying statistics and incentive-compatible designs are generally harder to obtain under strategic interference. However, under specific modeling assumptions about the interference, better designs can yield more information about the agent performances, and thus produce identifying statistics. We illustrate this idea in a viral marketing example, which we reuse throughout this paper.

2. PRELIMINARIES

In this section we introduce notation for the operational and statistical components of incentive-compatible experimental design. The operational components include the *treatment assignment*, the *treatment selection game* and the experiment *outcomes*. The statistical components include the *estimand*– the quantity of interest to the experimenter –and the *estimators*, i.e., the data statistics used to estimate the estimand.

2.1. Treatment assignment

Let $\mathcal{U} = \{1, 2, \ldots, m\}$ denote the set of *experimental units*, indexed by u, and $\mathcal{I} = \{1, 2, \ldots, n\}$ denote the set of *agents*, indexed by i. Each agent, for example, a marketing firm or a drug company, represents a treatment to be evaluated. An experimenter needs to design the experiment that will evaluate the agents. Relative to traditional

experimental design, the new aspect is that each agent is associated with a set of *treatment versions* and each agent has a strategic choice about which version to administer in the experiment. We make this precise in Section 2.2.

For each unit $u \in \mathcal{U}$ there is covariate information that is common knowledge to agents and the experimenter. We assume the experimenter uses covariates to split units into blocks, such that units within one block are similar in terms of covariates, e.g., similar age, gender, income, etc. Without loss of generality, we will assume there is just a single block. In Appendix A of this paper, we discuss how the theory can be extended to multiple blocks.

A *treatment assignment rule* ψ assigns each unit to a single agent. Let $\mathbf{Z} = (Z_u)$ denote the $m \times 1$ assignment vector, such that $Z_u = i$ indicates that unit u is assigned to agent i. The assignment rule ψ is a probability distribution over all possible assignments $\mathbf{Z}$. Without loss of generality, we assume that the number of units m is a multiple of the number of agents n. We will also assume complete randomization, such that $Z_u = i$, for exactly $k \stackrel{\text{def}}{=} m/n$ units, for each agent i.

2.2. Treatment selection game

The set of actions $\mathcal{A}_i \subseteq \mathcal{A}$ denotes the feasible action space for agent i, where $\mathcal{A}$ is the set of all possible actions. Subsequent to treatment assignment, every agent i simultaneously selects an action $A_i \in \mathcal{A}_i$, which corresponds to a version of the treatment administered by agent i. The same version is applied to all units assigned to agent i.[2] Let $\mathbf{A} = (A_1, \ldots A_n)$ denote the joint action profile, and $\mathbf{A}_{-i} = (A_1, \ldots, A_{i-1}, A_{i+1}, \ldots, A_n)$ denote the action profile without i's action.

We refer to this stage of the process as the *treatment selection game* in order to emphasize that agents (i.e., the treatments) can be strategic in selecting the treatment version they administer to units. This differentiates our setting from traditional experimental design, because it allows multiple versions of the same treatment to be available, hidden to the experimenter, and subject to selection by strategic agents. The traditional setting of experimental design is recovered if all action spaces of all agents are singletons, i.e., there is only one treatment version for each agent.[3]

2.3. Outcomes

Subsequent to the treatment selection game, an *outcome* is measured on each experimental unit u. Generally, the *potential outcome* of unit u, denoted by $Y_u(\mathbf{Z}, \mathbf{A})$, is the outcome that will be observed under assignment $\mathbf{Z}$ and agent actions $\mathbf{A}$. We assume that outcomes are numerical values; e.g., expenditure in dollars, number of product purchases, etc.

However, only one potential outcome can be observed at any given experiment, depending on the realized assignment $\mathbf{Z}$ and actions $\mathbf{A}$, while the rest will be missing. To emphasize the difference between potential outcomes and *observed outcomes*, we use additional notation. Let Y_{ui}^{obs} denote the observed outcome on unit u that was assigned to agent i. The notation Y_{ui}^{obs} implies that u was assigned to i (i.e., $Z_u = i$), and it is undefined if $Z_u \neq i$, i.e., u was not assigned to i. Following a "dot-notation," $Y_{\cdot i}^{\text{obs}}$ denotes the $k \times 1$ vector of observed outcomes of units assigned to agent i, and $Y_{\cdot\cdot}^{\text{obs}}$ denotes the $m \times 1$ vector of observed outcomes of all units.

---

[2] In Appendix A, we introduce multiple blocks and allow an agent to pick a different action for each block. All units within a block receive the same treatment version, but versions might differ across blocks.

[3] Dealing with multiple hidden treatments remains an open problem in traditional experimental design and causal inference, although not in a game theoretic setting as ours, and it is typically assumed away, for example, through SUTVA [Rubin 1980].

Note the dependence of potential outcomes on the complete assignment vector $\mathbf{Z}$; this allows the outcome of unit u to depend on assignment $Z_{u'}$ of some other unit u', even when agent actions $\mathbf{A}$ are held fixed. This situation is reasonable, for example, when units form social networks and influence each other, and is generally known as *social network interference* [Toulis and Kao 2013]. In our setting, interference between units affects the actions agents take (treatment versions), which then affect the interference on units, and so on. We collectively refer to this situation as *strategic interference*.[4]

We now illustrate the notation with an example application in viral marketing, which we will reuse throughout this paper.

Example 1. Assume four units $\mathcal{U} = \{1, 2, 3, 4\}$ in a single block, say, undergraduate students, and two marketing agents $\mathcal{I} = \{1, 2\}$. Further assume that 1 and 2 are close friends and 3 and 4 are close friends. The experimenter wants to understand which agent is better at advertising to students. Assume a treatment assignment $\mathbf{Z} = (1, 2, 1, 2)^\intercal$, i.e., units $1, 3$ are assigned to agent 1, and units $2, 4$ to agent 2. Each agent has two actions (treatment versions): advertise through phone or through social media. The action sets are thus $\mathcal{A}_1 = \mathcal{A}_2 = \{\text{phone, social}\}$, and a possible action profile is $\mathbf{A} = (\text{phone, social})^\intercal$ with $A_1 =$ phone (agent 1 uses phone to reach units 1 and 3) and $A_2 =$ social (agent 2 uses social media to reach units 2 and 4.)

The potential outcome $Y_u(\mathbf{Z}, \mathbf{A})$ could denote the number of product purchases (integer outcome) made by unit u, or the net profit from advertising to unit u (continuous outcome). Dependence on the assignment and treatment versions of both agents is reasonable because there could be word-of-mouth effects between students.

Consider observed data $Y^{\text{obs}}_{\cdot\cdot} = (0, 1, 4, 1)^\intercal$; for example, $Y^{\text{obs}}_{31} = 4$, which indicates that unit 3 was assigned to agent 1 and purchased four product items; Y^{obs}_{32} is undefined because the outcome of unit 3 when assigned to agent 2 is not observed. To illustrate the dot-notation, $Y^{\text{obs}}_{\cdot 1} = (0, 4)^\intercal$ indicates the outcomes of units assigned to agent 1, and $Y^{\text{obs}}_{\cdot 2} = (1, 1)^\intercal$ indicates the outcomes for agent 2.

In Example 1, the experimenter might be tempted to declare agent 1 as the winner, because it achieves $\overline{Y^{\text{obs}}_{\cdot 1}} = 2.0$ purchases/unit, as opposed to $\overline{Y^{\text{obs}}_{\cdot 2}} = 1.0$ purchases/unit for agent 2. However, these sample averages are subject to random variability from the randomization in the experiment, and may result from actions that are not the natural actions of the agents. Therefore, it is unclear whether the sample averages actually estimate how agents would do if they were selecting treatments without competition.

2.4. Estimand and estimators

A principled approach is to define the quantity of interest to the experimenter, the *estimand*, and then devise appropriate estimators for that quantity. The estimand is the agent with best possible performance, and thus we need a concrete notion of performance. For this, we want to estimate how good an agent's action would be if it was played without competition and thus without strategic interference. This is important

[4]There exists work in experimental design with between-unit interference [David and Kempton 1996], although not under a strategic interference setting as ours. In this paper, we will not be concerned with such forms of interference, but it will be the focus of future work. There is also related work in estimation of treatment effects in the context of strategic agents. For example, Athey et al. [2008] and Toulis and Parkes [2015] evaluate mechanisms in terms of their revenue, under the causal framework of potential outcomes. In both papers, the treatments are two different mechanism formats, and the units are the agents competing in the mechanism. The present work differs because, under our framework, the treatments are in fact strategic agents that are evaluated through an experiment, whereas the units passively exhibit treatment outcomes. See, also, the discussion by Dash and Druzdzel [2001] on the challenges of causal inference in dynamical systems within a different causal framework, namely causal graphs [Pearl 2000].

because, ultimately, the experimenter wants to assign a contract (e.g., an advertising campaign) to the winner agent, after which the winner will act by itself.

Let's define the *performance of agent i with respect to its action α_i*, denoted by $\chi(\alpha_i)$, as

$$\chi(\alpha_i) = \mathbb{E}\left(Y_u(\mathbf{Z}, \mathbf{A}) | \mathbf{A} = \alpha_i \mathbf{1}, Z_u = i\right); \tag{1}$$

notation $\mathbf{A} = \alpha_i \mathbf{1}$ denotes the hypothetical situation where all agents other than agent i are replaced by "replicates" of i, and each replicate plays action α_i. The dependence of $\chi(\alpha_i)$ on agent index i will be implicit in the notation. Given assignment vector $\mathbf{Z}$ and actions $\mathbf{A}$, we assume that the distribution of potential outcomes is known to all agents.

The expectation in Eq. (1) is taken with respect to this distribution, and defines the quantity of interest to the experimenter because it captures how agent i would do, on average, if the agent was acting alone without competition.[5] We also refer to χ as the *performance function*, and define $\boldsymbol{\chi}(\mathbf{A}) = (\chi(A_1), \chi(A_2), \ldots, \chi(A_n))^\intercal$. For brevity, all following definitions for an agent, e.g., natural action, quality, etc., will be implicitly assumed to be stated with respect to a particular performance function χ.

The *natural action* of agent i is the action that maximizes the quantity of interest to the experimenter in a system where agent i acts alone without competition. In particular, the *natural action* of agent i, denoted by $A_i^\star$, is defined as the action that maximizes its performance, i.e.,

$$A_i^\star \stackrel{\text{def}}{=} \arg\max_{\alpha_i \in \mathcal{A}_i} \{\chi(\alpha_i)\}. \tag{2}$$

The natural action profile is denoted by $\mathbf{A}^\star = (A_1^\star, A_2^\star, \ldots, A_n^\star)$. The *quality* of agent i, denoted by $\chi_i^\star \in \mathbb{R}$, is the maximum performance that the agent can achieve, i.e., $\chi_i^\star = \chi(A_i^\star)$. The *estimand*, denoted by τ, is the agent of highest quality, i.e.,

$$\tau = \arg\max_{i \in \mathcal{I}} \{\chi_i^\star\}. \tag{3}$$

To estimate the agent of highest quality the experimenter needs to use the observed outcomes $Y_{\cdot\cdot}^{\text{obs}}$. We will assume that the experimenter uses a *score function* $\phi : \mathbb{R}^m \to \mathbb{R}^n$, mapping all outcomes to a $n \times 1$ vector of scores for each agent, denoted by ϕ_i for agent i. For convenience, we will write $\boldsymbol{\phi}(Y_{\cdot\cdot}^{\text{obs}}) = (\phi_1(Y_{\cdot\cdot}^{\text{obs}}), \phi_2(Y_{\cdot\cdot}^{\text{obs}}), \ldots, \phi_n(Y_{\cdot\cdot}^{\text{obs}}))^\intercal$.

In the experiment, agents will be evaluated according to their scores, and the winner is the agent with the highest score. Several options for the score functions are possible. For example, $\phi_i(Y_{\cdot\cdot}^{\text{obs}}) = \overline{Y_{\cdot i}^{\text{obs}}}$, the sample mean of outcomes of units assigned to agent i, is one choice for the score function; other choices are possible, e.g., the sample Sharpe ratio, the sample median, etc.

The key challenge in incentive-compatible experimental design is to align maximizing the probability of winning the experiment, as induced in part by the score function ϕ, with selecting the action with maximum performance, i.e., the natural action.

---

[5] In causal inference, Eq. (1) is a *superpopulation estimand*, where the experimental units are assumed to be a random sample from a superpopulation of units, which is the target of statistical inference. The expectation in Eq. (1) is thus over all units in the superpopulation and all treatment assignments, for fixed agent actions. Other estimands in that superpopulation are possible; for example, the experimenter might be interested in the median outcomes, $\text{med}(Y_u(\mathbf{Z}, \mathbf{A}))$, or the Sharpe ratio, $\mathbb{E}(Y_u(\mathbf{Z}, \mathbf{A}))/\text{SD}(Y_u(\mathbf{Z}, \mathbf{A}))$, all conditional on fixed actions as in Eq. (1). In this paper, we work under the estimand of Eq. (1), mainly for simplicity, however our theory applies to all aforementioned estimands as well.

2.5. Incentive-compatible experiment designs

Let's first define an experiment design using the concepts of estimand and estimators from Section 2.4.

Definition 2.1. An *experiment design* $\mathcal{D} = (\psi, \phi)$ operates in the following steps:

(1) Receives units $\mathcal{U}$ and agents $\mathcal{I}$, as input.
(2) Samples a treatment assignment $\mathbf{Z}$ according to ψ.
(3) Each agent i picks a treatment version A_i, and administers the treatment to the set of its assigned units, $\{u \in \mathcal{U} : Z_u = i\}$.
(4) Outcomes on units $Y_{\cdot\cdot}^{\text{obs}}$ are observed.
(5) The winner agent $\hat{\tau}$ is declared according to the rule

$$\hat{\tau}(Y_{\cdot\cdot}^{\text{obs}}) = \arg\max_{i \in \mathcal{I}} \left\{ \phi_i(Y_{\cdot\cdot}^{\text{obs}}) \right\}. \tag{4}$$

Given experiment design $\mathcal{D}$ and action profile $\mathbf{A}$, the probability $P_i(\mathbf{A}|\mathcal{D})$ that agent i wins the experiment is given by:

$$\Pr\left(\hat{\tau}(Y_{\cdot\cdot}^{\text{obs}}) = i | \mathbf{A}, \mathcal{D}\right) \stackrel{\text{def}}{=} P_i(\mathbf{A}|\mathcal{D}) = P_i(\alpha_i, \mathbf{A}_{-i}|\mathcal{D}). \tag{5}$$

The randomness in Eq. (5) comes from the randomness of observed data $Y_{\cdot\cdot}^{\text{obs}}$, and the randomization in the treatment assignment. The winning probability $P_i(\cdot|\mathcal{D})$ in Eq. (5) is the *expected utility* of agent i under action profile $\mathbf{A}$, because agents care only about winning the experiment.

Definition 2.2 (*Incentive-compatible experiment design*). An experiment design $\mathcal{D} = (\psi, \phi)$ is *incentive-compatible* if the natural action $A_i^\star$ is a dominant strategy for each agent i, i.e., it maximizes the probability (5) of winning the experiment regardless of other agents' actions, such that

$$\arg\max_{\alpha_i \in \mathcal{A}_i} \{P_i(\alpha_i, \mathbf{A}_{-i}|\mathcal{D})\} = A_i^\star, \tag{6}$$

for all actions $\mathbf{A}_{-i}$, and every agent i.

Remark. In an incentive-compatible experiment, the score function ϕ induces a probability of winning (5) that is monotonically increasing with the performance function χ that the experimenter cares about. If this monotonicity holds, an agent will prefer to play the action that maximizes its performance (i.e., the natural action), because this will also maximize the winning probability.

The notation is summarized in Table I. We now return to the viral marketing problem that was introduced in Example 1. Examples 2(a)-(c) deal with Normally-distributed outcomes, whereas Examples 3(a)-(g) deal with Poisson-distributed outcomes. Examples 3(c)-(g) deal specifically with the problem of interference, and work with a more realistic form of the viral marketing problem.

Example 2(a). – Normal outcomes[6]. Consider the viral marketing problem of Example 1, with multiple units and two agents, where the outcomes of interest are the profit achieved from advertising to each unit. We assume that an agent action

[6]This two-agent example (low-quality agent vs. high-quality agent) is different from the example in the original paper published at EC'2015. The example was edited to illustrate a scenario where the low-quality agent prefers to play an action that is not its natural action and also reduces the winning chances of the high-quality agent. In the example of the original paper, the deviation from the low-quality agent actually increased the chances of the high-quality agent.

Table I. Notation for incentive-compatible experimental design

Symbol	Description	Value/Domain
$\mathcal{U}$	Set of m units	$\{1, 2, \ldots, m\}$
$\mathcal{I}$	Set of n agents	$\{1, 2, \ldots, n\}$
Z_u	Treatment assignment of unit u	$Z_u \in \mathcal{I}$
$\mathbf{Z}$	Vector of treatment assignment ($m \times 1$)	$(Z_1, \ldots, Z_m)^\intercal$
k	Units per agent	$k = m/n$
$\mathcal{A}$	Generic action space	
$\mathcal{A}_i$	Action space of agent i	$\mathcal{A}_i \subseteq \mathcal{A}$
A_i	Action of agent i	$A_i \in \mathcal{A}_i$
$\mathbf{A}$	Complete action profile ($n \times 1$)	$(A_1, \ldots, A_n)^\intercal$
$Y_u(\mathbf{Z}, \mathbf{A})$	Potential outcome of unit u under assignment $\mathbf{Z}$, actions $\mathbf{A}$	$Y_u(\mathbf{Z}, \mathbf{A}) \in \mathbb{R}$
Y_{ui}^{obs}	Observed outcome for unit u assigned to agent i	
$Y_{\cdot i}^{\text{obs}}$	Vector of observed outcomes of units assigned to agent i ($k \times 1$)	$Y_{\cdot i}^{\text{obs}} \in \mathbb{R}^k$
$Y_{\cdot\cdot}^{\text{obs}}$	Vector of observed outcomes of all units ($m \times 1$)	$Y_{\cdot\cdot}^{\text{obs}} \in \mathbb{R}^m$
$\chi(\alpha_i)$	Performance of agent i playing action α_i	$\chi(\alpha_i) \in \mathbb{R}$
$\boldsymbol{\chi}(\mathbf{A})$	Vector of performances ($n \times 1$)	$(\chi(A_1), \ldots, \chi(A_n))^\intercal$
$A_i^\star$	Natural action of agent i – maximizes performance	$A_i^\star \in \mathcal{A}_i$
$\chi_i^\star$	Quality of agent – performance at natural action	$\chi_i^\star \in \mathbb{R}$
τ	Agent of highest quality	$\tau \in \mathcal{I}$
$\phi_i(Y_{\cdot\cdot}^{\text{obs}})$	Score of agent i	$\phi_i(Y_{\cdot\cdot}^{\text{obs}}) \in \mathbb{R}$
$\boldsymbol{\phi}(Y_{\cdot\cdot}^{\text{obs}})$	Vector of agent scores ($n \times 1$)	$\big(\phi_1(Y_{\cdot\cdot}^{\text{obs}}), \ldots, \phi_n(Y_{\cdot\cdot}^{\text{obs}})\big)^\intercal$
$\hat{\tau}(Y_{\cdot\cdot}^{\text{obs}})$	Estimated agent of highest quality – agent with maximum score	$\hat{\tau}(Y_{\cdot\cdot}^{\text{obs}}) \in \mathcal{I}$
$P_i(\mathbf{A}\|\mathcal{D})$	Probability agent i wins under design $\mathcal{D}$, given fixed actions $\mathbf{A}$	

$\alpha_i = (\mu_i, \sigma_i^2) \in \mathbb{R} \times \mathbb{R}^+$, determines the mean and variance of the profit from advertising to unit u, such that, given assignment $\mathbf{Z}$, actions $\mathbf{A}$,

$$Y_u(\mathbf{Z}, \mathbf{A}) \sim \mathcal{N}(\mu_i, \sigma_i^2), \text{ if } A_i = \alpha_i, Z_u = i. \tag{7}$$

Note that Eq. (7) implies there is no interference between units, and no strategic interference between agent actions. We will make this precise in Section 3.

The experimenter is interested only in expected profit, ignoring the risk. Thus, the performance of action $\alpha_i = (\mu_i, \sigma_i^2)$ of agent i is

$$\chi(\alpha_i) \stackrel{\text{def}}{=} \mathbb{E}\left(Y_u(\mathbf{Z}, \mathbf{A}) | \mathbf{A} = \alpha_i \mathbf{1}, Z_u = i\right) = \mu_i. \tag{8}$$

Hence, the quality $\chi_i^\star$ of agent i is the maximum μ_i the agent can achieve over its action space $\mathcal{A}_i$. Now, consider an experiment design $\mathcal{D} = (\psi, \phi)$, where the score function ϕ is defined as $\phi_i(Y_{\cdot\cdot}^{\text{obs}}) = \overline{Y_{\cdot i}^{\text{obs}}}$, i.e., the score of agent i is the sample mean profit from all units assigned to agent i. Ignoring ties, the winning agent is given using Eq. (4):

$$\hat{\tau}(Y_{\cdot\cdot}^{\text{obs}}) = \begin{cases} 1, & \text{if } \overline{Y_{\cdot 1}^{\text{obs}}} > \overline{Y_{\cdot 2}^{\text{obs}}}, \\ 2, & \text{if } \overline{Y_{\cdot 1}^{\text{obs}}} < \overline{Y_{\cdot 2}^{\text{obs}}}. \end{cases} \tag{9}$$

By Eq. (7), $\overline{Y_{\cdot i}^{\text{obs}}} \sim \mathcal{N}(\mu_i, \sigma_i^2/k)$, where k is the number of units per agent. Hence, the probability that agent 1 wins is

$$P_1(\mathbf{A}|\mathcal{D}) \stackrel{\text{def}}{=} \Pr\left(\hat{\tau}(Y_{\cdot\cdot}^{\text{obs}}) = 1 | \mathbf{A}, \mathcal{D}\right) = P(\overline{Y_{\cdot 1}^{\text{obs}}} > \overline{Y_{\cdot 2}^{\text{obs}}}) = \Phi(\sqrt{k}\frac{\mu_1 - \mu_2}{\sqrt{\sigma_1^2 + \sigma_2^2}}), \tag{10}$$

where Φ is the normal cumulative distribution function (CDF). This design is not incentive-compatible because the winning probability $P_1(\mathbf{A}|\mathcal{D})$ is not monotone with

performance $\chi(\alpha_1) = \mu_1$ for action $\alpha_1 = (\mu_1, \sigma_1^2)$. For example, an increase in μ_1 may be associated with an increase in the risk σ_1^2, such that the probability of winning is reduced.

To see this, assume there are only two actions for agent 1, which induce mean and variance $\mathcal{A}_1 = \{(1.5, 100), (2, 20)\}$, and only one action for agent 2, $\mathcal{A}_2 = \{(9, 1)\}$. The quality of agent 1 is $\chi_1^\star \stackrel{\text{def}}{=} \max\{\mu : (\mu, \sigma^2) \in \mathcal{A}_1\} = 2$ and thus $(2, 20)$ is agent 1's natural action. However, when agent 1 plays the natural action, its winning probability is approximately equal to 0.12, whereas action $(1.5, 100)$ yields winnining probability 0.364, approximately. When agent 1 does not play the natural action, the expected value of its outcomes are reduced but their variance is increased, thus overall increasing agent 1's chances to win the experiment. Therefore, this experiment is not incentive compatible since agent 1 prefers not to play the natural action.

Example 2(b). – Normal outcomes – High risk/reward. Continuing Example 2(a), let's suppose that the variance of the unit's outcome satisfies $\sigma_i^2 = \mu_i^4$, indicating a delicate trade-off between expected return and risk. The probability that agent 1 wins is easily obtained from (10) as,

$$P_1(\mathbf{A}|\mathcal{D}) = P(\overline{Y_{\cdot 1}^{\text{obs}}} > \overline{Y_{\cdot 2}^{\text{obs}}}) = \Phi(\sqrt{k}\frac{\mu_1 - \mu_2}{\sqrt{\mu_1^4 + \mu_2^4}}). \tag{11}$$

The experiment design is still not incentive-compatible because (11) is not increasing monotonically with μ_1. As before, the better agent will choose to be more conservative, and will not reveal its quality (maximum possible μ_1). However, we will show in Section 3 that an incentive-compatible design can be achieved through the score function $\phi_i(Y_{\cdot\cdot}^{\text{obs}}) = -1/\overline{Y_{\cdot i}^{\text{obs}}}$, i.e., the negative reciprocal of the sample mean profit. We will show that, with this score function, the risk-reward trade-off in (11) disappears, which allows the experimenter to estimate agents' qualities.

Example 3(a) – Poisson outcomes. Now suppose the outcomes are integer-valued, e.g., representing the number of purchases. In this case, we assume that an agent's action $\alpha_i = (\lambda_i) \in \mathbb{R}^+$ determines the purchase rate by unit u, such that, given assignment $\mathbf{Z}$, actions $\mathbf{A}$,

$$Y_u(\mathbf{Z}, \mathbf{A}) \sim \text{Pois}(\lambda_i), \text{ if } A_i = \alpha_i, Z_u = i. \tag{12}$$

As in Eq. (7) of Example 2(a), Eq. (12) implies no interference. Let's suppose the experimenter is interested in performance that is the expected purchase rate. Thus, using Eq. (1), the experimenter measures performance of action $\alpha_i = (\lambda_i)$ of agent i, through

$$\chi(\alpha_i) \stackrel{\text{def}}{=} \mathbb{E}\left(Y_u(\mathbf{Z}, \mathbf{A})|\mathbf{A} = \alpha_i \mathbf{1}, Z_u = i\right) = \lambda_i. \tag{13}$$

Hence, the quality $\chi_i^\star$ of agent i is the maximum purchase rate λ_i that the agent can achieve over its action space $\mathcal{A}_i$. Now, consider the experiment design $\mathcal{D} = (\psi, \phi)$, where the score function ϕ is defined as $\phi_i(Y_{\cdot\cdot}^{\text{obs}}) = \overline{Y_{\cdot i}^{\text{obs}}}$, i.e., the score of agent i is the sample mean purchase rate from all units assigned to agent i. Ignoring ties, the winning agent $\hat{\tau}(Y_{\cdot\cdot}^{\text{obs}})$ is given using Eq. (9). By the central limit theorem, $\overline{Y_{\cdot i}^{\text{obs}}} \xrightarrow{D} \mathcal{N}(\lambda_i, \lambda_i/k)$, where "$\xrightarrow{D}$" denotes convergence in distribution, and k is the number of units per agent. The

probability that agent 1 wins is, asymptotically,

$$P_1(\mathbf{A}|\mathcal{D}) = P(\overline{Y_{\cdot 1}^{\text{obs}}} > \overline{Y_{\cdot 2}^{\text{obs}}}) = \Phi(\sqrt{k}\frac{\lambda_1 - \lambda_2}{\sqrt{\lambda_1 + \lambda_2}}). \tag{14}$$

This design is incentive-compatible because the winning probability $P_1(\mathbf{A}|\mathcal{D})$ is monotone with the agent performance; for example, an increase in λ_1 incurs a larger increase in the nominator of Eq. (14) than in the denominator. By symmetry, the winning probability for agent i is maximized at its natural action.

In Section 4.1, we will show that a more powerful design is possible, i.e., there exists an experiment design $\mathcal{D}'$ that is incentive-compatible and also guarantees higher winning chances to the better agent.

The examples highlight the challenges in incentive-compatible experimental design that arise because the experimenter is interested in some quality of an agent (e.g., expected return) but cannot find a design that incentivizes agents to play in a way that reveals their qualities. The problem that can arise is because of a mismatch between the score function ϕ that is used to declare the winner, and its effect in inducing a non-cooperative game, and the performance function χ that is of interest to the experimenter.

Compared with classical *mechanism design theory*, incentive-compatible experimental design differs in that:

— In mechanism design, the private information is an agent's preferences, whereas here the private information is an agent's quality (i.e., the performance of its natural action).
— In mechanism design, there may be side payments that can be made, whereas here the incentives are winner-take-all and depend on the outcome of the experiment.
— In mechanism design, it is standard to appeal to the *revelation principle* and design a direct-revelation mechanism, in which agents report their preference type to the mechanism. In comparison, the agents in our setting select an action and the designer observes the effect of this action, but not the action itself.

3. THEORY OF INCENTIVE-COMPATIBLE EXPERIMENTAL DESIGN

In this section we prove our main result, which provides a construction of score functions to design incentive-compatible experiments. The proof relies on the existence of statistics that can estimate the individual agent performances $\chi(A_i)$, as the number of units grows large.

Definition 3.1 (*Identifiable performance, identifying statistic*). An experiment design $\mathcal{D} = (\psi, \phi)$ has *identifiable performance* χ, if for every fixed action profile $\mathbf{A}$, there exists a statistic $T : \mathbb{R}^m \to \mathbb{R}^n$ calculated over data $Y_{\cdot\cdot}^{\text{obs}}$, such that

$$\sqrt{k}\left(T(Y_{\cdot\cdot}^{\text{obs}}) - \boldsymbol{\chi}(\mathbf{A})\right) \xrightarrow{D} \mathcal{N}(0, \Sigma(\mathbf{A})), \tag{15}$$

as the number of units per agent k grows large; $\mathcal{N}$ is the n-variate standard normal, and $\Sigma(\mathbf{A})$ is the $n \times n$ covariance matrix of T that can depend on $\mathbf{A}$. The statistic T is an *identifying statistic* for experiment design $\mathcal{D}$.

An identifying statistic is important because it estimates the individual performances $\chi(A_i)$, which are the quantities of interest to the experimenter. Although finding such a statistic is not an easy task, one simple strategy is to use sample quantities, such as averages, and then appeal to the central limit theorem, or other large-sample asymptotic results. We use this strategy extensively in this paper.

However, an identifying statistic T calculated over data $Y_{\cdot\cdot}^{\text{obs}}$ need not be sufficient for incentive alignment in our winner-take-all experiments. Thus, we consider score functions defined as $\phi_i(Y_{\cdot\cdot}^{\text{obs}}) = f(T_i)$, for an appropriate transformation $f : \mathbb{R} \to \mathbb{R}$. The transformation is used to add flexibility in the design of the score function. Agents will be evaluated according to the score vector $\boldsymbol{\phi}(Y_{\cdot\cdot}^{\text{obs}})$. The covariance matrix of the score vector $\boldsymbol{\phi}(Y_{\cdot\cdot}^{\text{obs}})$ is, asymptotically, equal to

$$V_f(\mathbf{A}) = \mathcal{J}_\phi \Sigma(\mathbf{A}) \mathcal{J}_\phi^\intercal, \tag{16}$$

where $\mathcal{J}_\phi$ is the Jacobian of ϕ calculated at $\boldsymbol{\chi}(\mathbf{A})$, actually a diagonal matrix with elements $f'(\chi(A_i))$. Whether an experiment design (ψ, ϕ) is incentive-compatible or not, depends crucially on the matrix $V_f(\mathbf{A})$ because this matrix defines the variances of the scores used to evaluate the agents.

THEOREM 3.2. *Fix agent actions* $\mathbf{A}$*, and consider design* $\mathcal{D} = (\psi, \phi)$ *that has an identifying statistic* T *with covariance matrix* $\Sigma(\mathbf{A})$*. Define the score function as* $\phi_i(Y_{\cdot\cdot}^{\text{obs}}) = f(T_i)$*, for some function* $f : \mathbb{R} \to \mathbb{R}$*, and let* $v_{ij}(\mathbf{A})$ *be the* ij*th element of* $V(\mathbf{A})$ *defined in Eq.* (16)*. Also define,*

$$v_f^{ij}(\alpha | \mathbf{A}_{-i}) = v_{ii}(\alpha, \mathbf{A}_{-i}) + v_{jj}(\alpha, \mathbf{A}_{-i}) - v_{ij}(\alpha, \mathbf{A}_{-i}) - v_{ji}(\alpha, \mathbf{A}_{-i}). \tag{17}$$

Design $\mathcal{D}$ *is incentive-compatible, if, for every agent* i*,*

$$\arg\max_{\alpha_i \in \mathcal{A}_i} \left\{ \frac{f(\chi(\alpha_i))}{v_f^{ij}(\alpha_i | \mathbf{A}_{-i})^{1/2}} \right\} = \arg\max_{\alpha_i \in \mathcal{A}_i} \{\chi(\alpha_i)\} \stackrel{\text{def}}{=} A_i^\star, \tag{18}$$

for every agent $j \neq i$*, and all actions* $\mathbf{A}_{-i}$*.*

For a fixed action profile $\mathbf{A}$, the element v_f^{ij} in Eq. (18), is the variance of the difference between the scores of agents i and j, $\phi_i(Y_{\cdot\cdot}^{\text{obs}}) - \phi_j(Y_{\cdot\cdot}^{\text{obs}})$, as defined in Theorem 3.2. Thus, Eq. (18) is the probability that agent i has a larger score than agent j, and implies that this probability is maximized at the natural action.

Theorem 3.2 suggests a recipe to construct incentive-compatible experiments, as we illustrate through examples in the following sections.

— First, one needs to find an identifying statistic to estimate the performances of agents, i.e., their outcomes without competition. A parametric model for the unit outcomes together with known asymptotic results, such as the central limit theorem, or the asymptotic normality of the maximum-likelihood estimator, can provide such an identifying statistic with known covariance matrix $\Sigma(\mathbf{A})$; see also Appendix D for a relevant discussion.
— Second, given the identifying statistic, one then needs to find an appropriate transformation f to satisfy Eq. (18). This transformation can be as simple as the identity function, as in Example 3(g), or the reciprocal function, as in Example 2(c). Intuitively, the design goal for f is to make the denominator of (18) less sensitive to agent actions than the nominator.

Theorem 3.2 makes no assumption about interference. In the following sections, we will specialize and apply Theorem 3.2 on the viral marketing example, both with and without interference.

4. INCENTIVE-COMPATIBLE EXPERIMENTS WITHOUT INTERFERENCE

The setting without interference is formally defined through the following assumption.

ASSUMPTION 1 (NO INTERFERENCE). *There is no strategic interference among agents and no interference between units, i.e., for all assignments* $\mathbf{Z}$ *and all agent actions* $\mathbf{A}$*,*

$$Y_u(\mathbf{Z}, \mathbf{A}) \equiv Y_u(A_i), \text{ where } Z_u = i. \tag{19}$$

Assumption 1 postulates that the potential outcome $Y_u(\mathbf{Z}, \mathbf{A})$ of a unit u assigned to agent i, remains constant as long as agent i's action and unit u's assignment to agent i are held fixed. Under no interference, the distribution of a score function defined through an identifying statistic is a univariate normal, as shown in the following proposition.

PROPOSITION 4.1. *Consider design* $\mathcal{D} = (\psi, \phi)$ *with an identifying statistic* T *with covariance matrix* $\Sigma(\mathbf{A})$*. Let* $\phi_i(Y^{\text{obs}}_{\cdot\cdot}) = f(T_i)$*, for some function* $f : \mathbb{R} \to \mathbb{R}$*, and suppose Assumption 1 holds. Then, for fixed actions* $\mathbf{A}$*,*

$$\sqrt{k}\left(\phi_i(Y^{\text{obs}}_{\cdot\cdot}) - f(\chi(A_i)\right) \xrightarrow{D} \mathcal{N}(0, \sigma^2(A_i)), \tag{20}$$

where $\sigma^2(A_i) = f'(\chi(A_i))^2\sigma^2_{ii}$*, with* σ^2_{ii} *being the* i*th diagonal element of* $\Sigma(\mathbf{A})$*.*

PROOF. By Assumption 1 (no interference), the covariance matrix $\Sigma(\mathbf{A})$ of T is diagonal with elements σ^2_{ii}. Thus, by definition of the identifying statistic,

$$\sqrt{k}(T_i - \chi(A_i)) \xrightarrow{D} \mathcal{N}(0, \sigma^2_{ii}).$$

Since, $\phi_i(Y^{\text{obs}}_{\cdot\cdot}) = f(T_i)$, Eq. (20) follows from a simple application of the Delta theorem; see, for example, Bickel and Doksum [2001, Chapter 5], or Cox [1998]. □

Proposition 4.1 provides the asymptotic distribution of the score function, given an identifying statistic and a known transformation f, when there is no interference. This will be useful to derive the winning probabilities for agents in the experiment. We first illustrate Proposition 4.1, and then show how it can be used to simplify the conditions of the more general Theorem 3.2.

Example 2(c). We continue from Example 2(b), where agent i's action is $A_i = (\mu_i)$, and $\overline{Y^{\text{obs}}_{\cdot i}} \sim \mathcal{N}(\mu_i, \mu_i^4/k)$, where k is the number of units per agent. The statistic $T(Y^{\text{obs}}_{\cdot\cdot}) = (\overline{Y^{\text{obs}}_{\cdot 1}}, \overline{Y^{\text{obs}}_{\cdot 2}}, \ldots, \overline{Y^{\text{obs}}_{\cdot n}})^\intercal \equiv T$, is an identifying statistic, since $\boldsymbol{\chi}(\mathbf{A}) = (\mu_1, \mu_2, \ldots, \mu_n)^\intercal \stackrel{\text{def}}{=} \boldsymbol{\mu}$, and

$$\sqrt{k}(T - \boldsymbol{\mu}) \xrightarrow{D} \mathcal{N}(0, \Sigma), \tag{21}$$

where $\Sigma = \text{diag}(\mu_1^4, \ldots, \mu_n^4)$, is the diagonal matrix with elements μ_i^4.

Consider the score functions $\phi_i(Y^{\text{obs}}_{\cdot\cdot}) = 1/T_i = 1/\overline{Y^{\text{obs}}_{\cdot i}}$, i.e., $f(x) = 1/x$, in the notation of Proposition 4.1. Using the result in Proposition 4.1, $\sigma^2(A_i) = f'(\mu_i)^2\mu_i^4 = 1$, and thus

$$\sqrt{k}(\phi_i(Y^{\text{obs}}_{\cdot\cdot}) - 1/\mu_i) \xrightarrow{D} \mathcal{N}(0, 1). \tag{22}$$

The variance of the score function in Eq. (22) is stabilized. The following theorem shows that such variance stabilization can lead to incentive-compatible designs, when there is no interference.

THEOREM 4.2. *Consider design $\mathcal{D} = (\psi, \phi)$ with an identifying statistic T with covariance matrix $\Sigma(\mathbf{A})$. Suppose Assumption 1 holds. If, for every agent i,*

$$\phi_i(Y_{\cdot\cdot}^{\text{obs}}) = f(T_i), \text{ where } f : \mathbb{R} \to \mathbb{R}, \tag{23}$$

$$\mathbb{V}\text{ar}(\phi_i(Y_{\cdot\cdot}^{\text{obs}})) = \text{const.}, \tag{24}$$

$$\arg\max_{\alpha_i \in \mathcal{A}_i} f(\chi(\alpha_i)) = \arg\max_{\alpha_i \in \mathcal{A}_i} \{\chi(\alpha_i)\} \stackrel{\text{def}}{=} A_i^\star, \tag{25}$$

then design $\mathcal{D}$ is incentive-compatible.

Condition (24) is related to variance-stabilizing transformations in statistics, which also play an important role in hypothesis testing; we discuss this relationship in Appendix C.

Example 2(d). – Normal outcomes – High risk/reward. Continuing from Example 2(c), we consider the high risk-reward setting of the viral marketing problem, where an agent's action is to pick an expected return, i.e., $A_i = (\mu_i)$, and the winning probability is given by

$$P_1(\mathbf{A}|\mathcal{D}) = \Phi(\sqrt{k}\frac{\mu_1 - \mu_2}{\sqrt{\mu_1^4 + \mu_2^4}}). \tag{26}$$

The performance function is $\chi(\alpha_i) = \mu_i$, and thus the natural action is $A_i^\star = \arg\max_{\alpha_i \in \mathcal{A}_i}\{\alpha_i\}$. It was shown that design $\mathcal{D}$ in Example 2(b) –using the sample mean as the score function– is not incentive-compatible. Consider instead a design $\mathcal{D}'$ with score function $\phi_i(Y_{\cdot\cdot}^{\text{obs}}) = -1/\overline{Y_{\cdot i}^{\text{obs}}}$. Using the result of Example 2(c),

$$\sqrt{k}\left(\phi_i(Y_{\cdot\cdot}^{\text{obs}}) - (-1/\mu_i)\right) \xrightarrow{D} \mathcal{N}(0, 1). \tag{27}$$

Condition (23) is satisfied by definition of ϕ_i. Condition (24) is also satisfied, because the variance of $\phi_i(Y_{\cdot\cdot}^{\text{obs}})$ in Eq. (27) is constant. Furthermore,

$$\arg\max_{\alpha_i \in \mathcal{A}_i} \{f(\chi(\alpha_i))\} = \arg\max_{\alpha_i \in \mathcal{A}_i} \{-1/\alpha_i\} = \arg\max_{\alpha_i \in \mathcal{A}_i} \{\alpha_i\} = A_i^\star,$$

which satisfies Condition (25). Thus, all conditions of Theorem (4.2) are fulfilled. It follows that the new design $\mathcal{D}'$ is incentive-compatible.

By construction of the probabilistic model in Example 2(b), there is a very delicate trade-off between expected return (agent performance) and risk; for example, if an agent doubles its performance, then the risk will quadruple. In such situations, it is a bad idea to adopt the sample mean as the score statistic. Intuitively, Eq. (26) shows that the higher-quality agent will try more conservative actions, thus hiding its true quality. However, if agents are scored according to the negated reciprocal of their sample mean, the probability that an agent wins increases monotonically with an agent's performance. Thus, agents have the incentive to select actions that maximize their performance, and thus it is a dominant strategy to select their natural action.

4.1. Powerful incentive-compatible experiment designs

Given the choice of two incentive-compatible designs, it is natural to prefer the design in which the highest-quality agent has the highest probability of winning. We formalize this intuition through the following definition.

Definition 4.3 (*Powerful incentive-compatible design*). Consider two experiment designs $\mathcal{D}$ and $\mathcal{D}'$ that are both incentive-compatible and operate on the same set of units $\mathcal{U}$. Let τ be the agent of highest quality. Design $\mathcal{D}'$ is (weakly) *more powerful* than

design $\mathcal{D}$ if the probability that agent τ wins in the dominant strategy equilibrium is higher in $\mathcal{D}'$ than $\mathcal{D}$; i.e.,

$$P_\tau(\mathbf{A}^\star|\mathcal{D}') \geq P_\tau(\mathbf{A}^\star|\mathcal{D}), \tag{28}$$

where $\mathbf{A}^\star$ is the natural action profile, which is the same in both designs.

In the following theorem, we give a simple case where we can transform an incentive-compatible design into a more powerful one.

THEOREM 4.4. *Consider an incentive-compatible design $\mathcal{D} = (\psi, \phi)$, where action sets $\mathcal{A}_i \subseteq \mathbb{R}$ are compact, and performance χ is one-to-one and continuous. Let,*

$$\sqrt{k}\left(\phi_i(Y_{\cdot\cdot}^{\text{obs}}) - \chi(A_i)\right) \xrightarrow{D} \mathcal{N}(0, \sigma^2(A_i)), \tag{29}$$

where function $\sigma^2 : \mathcal{A} \to \mathbb{R}^+$ satisfies

$$\chi(\alpha_i') \geq \chi(\alpha_i) \Rightarrow \sigma^2(\alpha_i') \geq \sigma^2(\alpha_i), \tag{30}$$

for every agent i, and all actions $\alpha_i', \alpha_i \in \mathcal{A}_i$.[7]

Consider a design $\mathcal{D}' = (\psi, \phi')$, where $\phi_i'(Y_{\cdot\cdot}^{\text{obs}}) = \nu(\phi_i(Y_{\cdot\cdot}^{\text{obs}}))$, for each agent i, with $\nu(\cdot)$ defined by

$$\nu(y) = \int^y \frac{1}{\sqrt{\sigma(\chi^{-1}(z))}} dz. \tag{31}$$

Then, design $\mathcal{D}'$ is incentive-compatible and more powerful than $\mathcal{D}$, if $\nu(\cdot)$ is convex, or $1/\sqrt{\sigma^2(\chi^{-1}(\cdot))}$ and $\sigma^2(\chi^{-1}(\cdot))$ are both convex.

The variance of the new score function, $\mathbb{V}\text{ar}(\phi_i'(Y_{\cdot\cdot}^{\text{obs}}))$, is constant, because function ν defined in Eq. (29) is a variance-stabilizing transformation [Cox 1998]. This fulfills Condition (24) of Theorem 4.2, while the monotonicity (30) of $\sigma(\cdot)$ maintains the monotonicity Condition (25). The new design $\mathcal{D}'$ is thus incentive-compatible.

Example 3(b) – Poisson outcomes. Continuing from Example 3(a), the actions are $A_i = (\lambda_i) \in \mathbb{R}^+$ with performance $\chi(A_i) = \lambda_i$, while the score statistic is $\phi_i(Y_{\cdot\cdot}^{\text{obs}}) = \overline{Y_{\cdot i}^{\text{obs}}}$; thus, $\sqrt{k}\left(\phi_i(Y_{\cdot\cdot}^{\text{obs}}) - \lambda_i\right) \xrightarrow{D} \mathcal{N}(0, \lambda_i)$. Let agent 1 be the best agent. Consider a new design $\mathcal{D}'$ with the transformation

$$\nu(y) = \int^y \frac{1}{\sqrt{\sigma(\chi^{-1}(z))}} dz = \int^y \frac{1}{\sqrt{z}} dz = 2\sqrt{z},$$

and score function $\phi_i'(Y_{\cdot\cdot}^{\text{obs}}) = \nu(\phi_i(Y_{\cdot\cdot}^{\text{obs}})) = 2\sqrt{\overline{Y_{\cdot i}^{\text{obs}}}}$. Design $\mathcal{D}'$ is incentive-compatible and more powerful than design $\mathcal{D}$ of Example 3(a) by Theorem 4.4, since $1/\sqrt{\sigma^2(\chi^{-1}(z))} = 1/\sqrt{z}$ and $\sigma^2(\chi^{-1}(z)) = z$, are both convex. Another way to see this is through Proposition 4.1, which implies $\sqrt{k}\left(\phi_i'(Y_{\cdot\cdot}^{\text{obs}}) - 2\sqrt{\lambda_i}\right) \xrightarrow{D} \mathcal{N}(0, 1)$. Thus, the probability that agent 1 wins is

$$P_1(\mathbf{A}|\mathcal{D}') = \Phi(\sqrt{2k}(\sqrt{\lambda_1} - \sqrt{\lambda_2})). \tag{32}$$

[7] Condition (30) posits that an agent cannot increase its expected score without increasing the variance of the score. This is a reasonable assumption in practice because actions that do increase the expected score without increasing the variance, are strongly preferred.

We can verify $P_1(\mathbf{A}|\mathcal{D}') > P_1(\mathbf{A}|\mathcal{D})$ by comparing Eq. (32) with Eq. (14):

$$\Phi\left(\sqrt{2k}(\sqrt{\lambda_1}-\sqrt{\lambda_2})\right) > \Phi\left(\sqrt{k}\frac{\lambda_1-\lambda_2}{\sqrt{\lambda_1+\lambda_2}}\right) \Leftrightarrow \sqrt{2}(\sqrt{\lambda_1}-\sqrt{\lambda_2}) > \frac{\lambda_1-\lambda_2}{\sqrt{\lambda_1+\lambda_2}}.$$

The last inequality always holds because it reduces to $(\sqrt{\lambda_1}-\sqrt{\lambda_2})^2 > 0$.

In Example 3(b), the better agent (agent 1) has higher chances of winning in the new design $\mathcal{D}'$. Since $\mathcal{D}'$ is also incentive-compatible, it follows that $\mathcal{D}'$ is more powerful than $\mathcal{D}$. Intuitively, the square root transformation in the new design stabilizes the variance – there is no denominator in Eq. (32) – which achieves incentive-compatibility through Theorem 4.2.

5. INCENTIVE-COMPATIBLE EXPERIMENTS WITH INTERFERENCE

We now consider strategic interference, whereby an action of an agent can affect the outcomes of units assigned to another agent. Therefore, agent scores calculated on individual agent outcomes are confounded with the entire action profile.

Example 3(c) – Poisson outcomes with interference. Building upon Example 3(b), we now introduce a more realistic model of the viral marketing experiment, which we assume operates as follows.

As before, units are assigned to agent 1 or agent 2. We refer to the units assigned to agent i, i.e., the set $\{u \in \mathcal{U} : Z_u = i\}$, as the *test set* of agent i. In addition, each agent is free to pick a *seed set*; each seed set is in a separate population that is disjoint from the test sets. The seed set i corresponds to treatment version –agent action– A_i. The seed set will be targeted with a promotional campaign, and outcomes will be measured on units only in the test sets, say, number of purchases for each unit. The rationale is that the experimenter is interested in the viral marketing efficacy of the agents, i.e., their ability to select influential seed sets.

Under interference, the treatment version (seed set) selected by agent i induces a rate λ_i on units assigned to i, and a rate $\gamma\lambda_i^{'}$, where $0 \leq \gamma \leq 1$, on units assigned the other agent. The parameter γ models the amount of interference; if $\gamma = 0$ there is no interference, whereas $\gamma = 1$ indicates maximum interference. For the rest of this paper we will consider γ known to agents and the designer, but this is without loss of generality. Rate $\lambda_i^{'}$ can be interpreted as the rate that agent i would achieve if the units that are targeted were its own units. Parameter γ represents a discount because the targeted units are in the test set of another agent.

The setting with interference is depicted in Figure 1. The labels on the edges correspond to the effects from the seed sets, including interference effects. For example, the purchase rate in test set 2 (units assigned to agent 2) is equal to $\gamma\lambda_1^{'} + \lambda_2$; the first term is the discounted influence from the seed set of agent 1, and the second term is the influence from the seed set of agent 2. Agents are scored based on outcomes of units in their respective test sets. Therefore, an agent can also "free-ride" on the conversion rate that comes from the action of the other agent.

Example 3(d) – Poisson outcomes with interference. Given the interference model of Example 3(c), the actions are $A_1 = (\lambda_1, \lambda_1^{'})$, $A_2 = (\lambda_2, \lambda_2^{'})$, and the observed outcomes on the units in the test sets have the following distributions:

$$\begin{aligned} Y_{u1}^{\text{obs}} &\sim \text{Pois}(\lambda_1 + \gamma\lambda_2^{'}), \\ Y_{u2}^{\text{obs}} &\sim \text{Pois}(\lambda_2 + \gamma\lambda_1^{'}). \end{aligned} \tag{33}$$

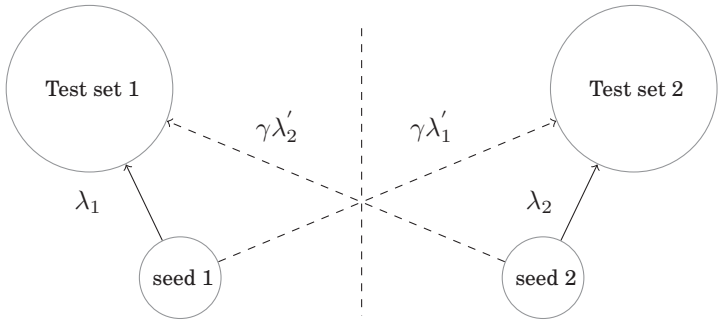

Fig. 1. Test set i has units assigned to agent i, i.e., $\{u \in \mathcal{U} : Z_u = i\}$. Seed set i corresponds to the treatment version A_i. The seed sets influence the purchase rate of units in the test sets, for example, through word-of-mouth effects between units. In particular, $A_i = (\lambda_i, \lambda_i^{'})$, where λ_i is the induced rate from seed set i to test set i, and $\gamma\lambda_i^{'}$ is the induced rate from seed set i to the other test set, where $0 \leq \gamma \leq 1$ is a parameter that models interference. Outcomes, i.e., product purchases, are measured on units in the test sets; the score of agent i will be calculated based on observed purchases in test set i. Arrows indicate induced purchase rates from the seed sets; dashed arrows indicate that the rate is discounted by γ. The presence of interference, where an agent can affect the purchase rate on a test set of another agent, changes how agent select their seed sets, i.e., their treatment versions.

To derive the performance of an agent, say agent 1, we need to replace agent 2 with a replicate of agent 1, playing action $A_2 = (\lambda_1, \lambda_1^{'})$. In this case, the induced rate on the units assigned to agent 1 is actually equal to $\lambda_1 + \lambda_1^{'}$ since, by definition of our interference model in Example 3(c), a rate is discounted only from a seed set of one agent to the test set of another agent. Thus, the performance of agent i for action $\alpha_i = (\lambda_i, \lambda_i^{'})$ is equal to

$$\chi(\alpha_i) = \mathbb{E}\left(Y_u(\mathbf{Z}, \mathbf{A}) | \mathbf{A} = \alpha_i \mathbf{1}, Z_u = i\right) = \lambda_i + \lambda_i^{'}. \tag{34}$$

It can be seen, by inspection of Eq. (33), that the outcomes of one unit depend on the action of the other agent. For example, the outcomes $Y_{\cdot 1}^{\text{obs}}$ on units assigned to agent 1 depend on action λ_1 of agent 1 as well as action $\lambda_2^{'}$ of agent 2. Hence, the observed outcomes for one agent carries statistical information for the action of the other agent. This information should be used in order to correctly estimate the agent qualities, and then the agent of highest quality.

However, the estimation of qualities is not possible through outcomes (33), because there exist multiple action profiles for which the observed outcomes are equally likely. It follows that there is no identifying statistic, and our theory (e.g., Theorem 3.2) cannot be applied. Furthermore, the variance-stabilization transformations that were shown to give more powerful designs in Example 3(b) do not work. This is illustrated in the following example.

Example 3(e). – Poisson outcomes with interference. Consider the setup of Example 3(c) and an experiment $\mathcal{D}$ with the usual score function $\phi_i(Y_{\cdot\cdot}^{\text{obs}}) = \overline{Y_{\cdot i}^{\text{obs}}}$. As the number of experimental units grows, Eq. (33) result in the following asymptotics.

$$\sqrt{k}\left(\overline{Y_{\cdot 1}^{\text{obs}}} - (\lambda_1 + \gamma\lambda_2^{'})\right) \xrightarrow{D} \mathcal{N}(0, \lambda_1 + \gamma\lambda_2^{'}),$$
$$\sqrt{k}\left(\overline{Y_{\cdot 2}^{\text{obs}}} - (\lambda_2 + \gamma\lambda_1^{'})\right) \xrightarrow{D} \mathcal{N}(0, \lambda_2 + \gamma\lambda_1^{'}).$$

Therefore, the probability that agent 1 wins is

$$P_1(\mathbf{A}|\mathcal{D}) = \Pr(\overline{Y_{.1}^{\text{obs}}} > \overline{Y_{.2}^{\text{obs}}}) = \Phi\left(\sqrt{k}\frac{(\lambda_1 - \gamma\lambda_1^{'}) - (\lambda_2 - \gamma\lambda_2^{'})}{\sqrt{\lambda_1 + \gamma\lambda_1^{'} + \lambda_2 + \gamma\lambda_2^{'}}}\right). \tag{35}$$

This design is not incentive-compatible because agent 1 prefers a large $\lambda_1 - \gamma\lambda_1^{'}$ and a small $\lambda_1 + \gamma\lambda_1^{'}$. As can been seen from Figure 1, a purchase rate of $\gamma\lambda_1^{'}$ from the seed set of agent 1 only benefits agent 2. Thus, agent 1 wants to benefit its assigned units (test set 1) while minimizing the spillovers to test set 2 that benefit only agent 2. However, the experimenter wants to know something very different. In particular, given the definition of performance in Example 3(d), the experimenter wants to know the maximum $\lambda_1 + \lambda_1^{'}$ that agent 1 can achieve (and maximum $\lambda_2 + \lambda_2^{'}$, for agent 2). This quantity is of interest because it is the quantity that agent 1 would maximize if a copy of agent 1 substituted agent 2, and also played $(\lambda_1, \lambda_1^{'})$.

Using the variance-stabilizing transformation of Example 3(b), does not solve the problem. In particular, if we use $\phi_i(Y_{\cdot\cdot}^{\text{obs}}) = 2\sqrt{\overline{Y_{\cdot i}^{\text{obs}}}}$ as the score function, then the winning probability of agent 1 becomes

$$P_1(\mathbf{A}|\mathcal{D}) = \Phi\left(\sqrt{k/2}(\sqrt{\lambda_1 + \gamma\lambda_2^{'}} - \sqrt{\lambda_2 + \gamma\lambda_1^{'}})\right).$$

The incentive problem remains because agent 1 still wants achieve a high purchase rate λ_1 on units in test set 1, and a low rate $\lambda_1^{'}$ in units of test set 2.

5.1. Dealing with strategic interference through better designs

We now describe a method to construct an incentive-compatible design in the viral marketing problem with interference. The idea is to introduce a new design that will provide an identifying statistic, and then define appropriate score functions to fulfill the conditions of Theorem 3.2 that guarantee incentive-compatibility.

Example 3(f). – Poisson outcomes with interference – New design. We consider the following new design. The units are split in two groups, say G_1 and G_2. Within each group, units are randomly assigned to the two agents, resulting in 2 test sets per agent. For example, group G_1 has two test sets, namely G_{11} with units assigned to agent 1, and G_{12} with units assigned to agent 2. Similarly, group G_2 has test sets G_{21} with units assigned to agent 1, and G_{22} with units assigned to agent 2. Test sets in the same group may be overlapping. In addition, each agent is free to pick one *seed set*; each seed set is in a separate population that is disjoint from the test sets. The seed set i corresponds to treatment version –agent action– A_i. The outcomes Y, say number of purchases for each unit, for each agent i, will be measured on units only in their two test sets, namely G_{1i} and G_{2i}. This design is depicted in Figure 2.

The outcomes model is similar to the design of Example 3(c) (see also Figure 1). A seed set i –action A_i– induces a rate λ_i on units of group G_i, and a rate $\lambda_i^{'}$ on units of the other group. The rate is assumed to be discounted when the seed set is targeting units in a test set of another agent. For example, units in test set G_{12} will have purchase rate $\lambda_2^{'} + \gamma\lambda_1$; the rate $\lambda_2^{'}$ originates from seed set 2 affecting units in group G_1, and rate λ_1 is from seed set 1 affecting units in G_1, discounted by γ because G_{12} is a test set of agent 2. Thus, action A_i is associated with a pair of rates, $A_i = (\lambda_i, \lambda_i^{'})$.

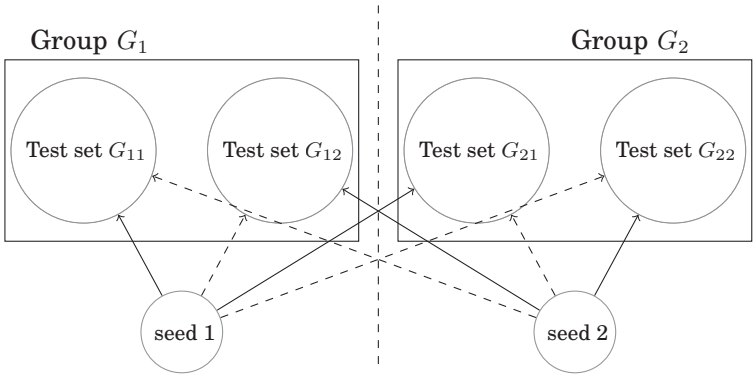

Fig. 2. Test sets G_{1j} and G_{2j} have the units assigned to agent j, i.e., $\{u \in \mathcal{U} : Z_u = i\}$; there are two test sets per agent. Agent i selects an influential seed set i, that corresponds to the treatment version A_i. The seed sets influence the purchase rate of units in the test sets. In particular, $A_i = (\lambda_i, \lambda_i^{'})$, where λ_i is the induced rate from seed set i to a test set with units assigned to i, and $\gamma\lambda_i^{'}$ is the induced rate from seed set i to a test set with units assigned to the other agent. Outcomes are measured on units in the test sets; the score of agent i will be calculated based on observed purchases of units assigned to agent i; for example, agent 1 will be scored based on outcomes of units in G_{11} and G_{21}. Arrows indicate induced purchase rates from the seed sets; dashed arrows indicate that the rate is discounted by γ. Agent scores are calculated based on outcomes in their respective test sets. The presence of interference, where an agent can affect the purchase rate on a test set of another agent, changes how agent select their seed sets, i.e., their treatment versions.

Agent 1's action is $A_1 = (\lambda_1, \lambda_1^{'})$, and agent 2's action is $A_2 = (\lambda_2, \lambda_2^{'})$. Therefore, the observed outcomes of units are distributed as follows:

$$Y_{ui}^{\text{obs}} \sim \begin{cases} \text{Pois}(\lambda_1 + \gamma\lambda_2^{'}), & \text{if } u \in G_{11}, \\ \text{Pois}(\lambda_2^{'} + \gamma\lambda_1), & \text{if } u \in G_{12}, \\ \text{Pois}(\lambda_1^{'} + \gamma\lambda_2), & \text{if } u \in G_{21}, \\ \text{Pois}(\lambda_2 + \gamma\lambda_1^{'}), & \text{if } u \in G_{22}. \end{cases} \tag{36}$$

Using the same interference model (parameter γ of discounted influence) introduced in Example 3(c), the new design of Figure 2 now provides more information about the agent actions, and thus their performance, through outcomes (36). This additional information provides an identifying statistic that can be used to define score functions that make the design of Figure 2 incentive-compatible.

Example 3(g). – Poisson outcomes. By symmetry of the new design, the experimenter is interested to estimate $\chi(A_i) = \lambda_i + \lambda_i^{'}$. Let $\bar{Y}_{ij}$ be the sample mean of outcomes of units in test set G_{ij}, and let $Y = (\bar{Y}_{11}, \bar{Y}_{12}, \bar{Y}_{21}, \bar{Y}_{22})^{\intercal}$. Define the matrices

$$B = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix}, \text{ and } C = \begin{pmatrix} 1 & 0 & \gamma & 0 \\ \gamma & 0 & 1 & 0 \\ 0 & 1 & 0 & \gamma \\ 0 & \gamma & 0 & 1 \end{pmatrix}.$$

Denote the action profile as $\mathbf{A} = (\lambda_1, \lambda_1^{'}, \lambda_2^{'}, \lambda_2)^{\intercal}$. Further, let $D_{\mathbf{A}} = \text{diag}(C\mathbf{A})$ be the diagonal matrix with diagonal elements from the vector $C\mathbf{A}$. By Eq. (36), as the

number of units grows, we have

$$\sqrt{m/4}(Y - C\mathbf{A}) \xrightarrow{D} \mathcal{N}(0, D_{\mathbf{A}}). \tag{37}$$

The term $m/4$ is because there are $m/4$ units per test set. Now define the statistic $T = BC^{-1}Y$. Since $\boldsymbol{\chi}(\mathbf{A}) = (\lambda_1 + \lambda_1^{'}, \lambda_2 + \lambda_2^{'})^{\intercal} = B\mathbf{A}$, it holds, asymptotically,[8]

$$\sqrt{m/4}(T - \boldsymbol{\chi}(\mathbf{A})) \xrightarrow{D} \mathcal{N}(0, BC^{-1}D_{\mathbf{A}}(C^{-1})^{\intercal}B^{\intercal}). \tag{38}$$

Therefore, the new design has identifiable performance, and T is an identifying statistic, with covariance matrix $\Sigma(\mathbf{A}) = BC^{-1}D_{\mathbf{A}}(C^{-1})^{\intercal}B^{\intercal}$.

Now, using notation of Theorem 3.2, define the score function simply as

$$\phi_i(Y_{\cdot\cdot}^{\text{obs}}) = f(T_i) = T_i. \tag{39}$$

Thus, the Jacobian of ϕ is $\mathcal{J}_\phi = \mathbb{I}$, the identity matrix. The matrix $V(\mathbf{A})$ of Theorem 3.2 is calculated as

$$V(\mathbf{A}) = \mathcal{J}_\phi \Sigma(\mathbf{A}) \mathcal{J}_\phi^{\intercal} = BC^{-1}D_{\mathbf{A}}(C^{-1})^{\intercal}B^{\intercal}. \tag{40}$$

Through simple but tedious matrix algebra we obtain,

$$V(\mathbf{A}) = \frac{1}{(1-\gamma^2)^2} \begin{pmatrix} d_1 + \gamma^2 d_2 + d_3 + \gamma^2 d_4 & -\gamma \sum_{i=1}^4 d_i \\ -\gamma \sum_{i=1}^4 d_i & \gamma^2 d_1 + d_2 + \gamma^2 d_3 + d_4 \end{pmatrix}, \tag{41}$$

where (d_i) are the diagonal elements of $D_{\mathbf{A}}$; thus, $d_1 = \lambda_1 + \gamma\lambda_2^{'}$, $d_2 = \gamma\lambda_1 + \lambda_2^{'}$, $d_3 = \lambda_1^{'} + \gamma\lambda_2$, and $d_4 = \gamma\lambda_1^{'} + \lambda_2$. In particular,

$$\sum_{i=1}^4 d_i = (1+\gamma)\left[(\lambda_1 + \lambda_1^{'}) + (\lambda_2 + \lambda_2^{'})\right]. \tag{42}$$

It follows from Theorem Eq. (17) of Theorem 3.2,

$$\begin{aligned} v_f^{ij}(\alpha|\mathbf{A}_{-i}) =& (d_1 + \gamma^2 d_2 + d_3 + \gamma^2 d_4) + (\gamma^2 d_1 + d_2 + \gamma^2 d_3 + d_4) - (-2\gamma \sum_{i=1}^4 d_i) \\ =& (1+\gamma)^2 \sum_{i=1}^4 d_i = (1+\gamma)^3 \left[(\lambda_1 + \lambda_1^{'}) + (\lambda_2 + \lambda_2^{'})\right], \end{aligned}$$

if $i \neq j$, and 0 otherwise. It follows that,

$$\arg\max_{\alpha_i \in \mathcal{A}_i} \left\{ \frac{f(\chi(\alpha_i))}{v_f^{ij}(\alpha_i|\mathbf{A}_{-i})^{1/2}} \right\} \propto \arg\max_{\alpha_i \in \mathcal{A}_i} \left\{ \frac{\lambda_i + \lambda_i^{'}}{\sqrt{(\lambda_1 + \lambda_1^{'}) + (\lambda_2 + \lambda_2^{'})}} \right\}. \tag{43}$$

The expression on the right of Eq. (43) is increasing with respect to $\chi(\alpha_i) = \lambda_i + \lambda_i^{'}$. Therefore, each agent prefers to play actions $(\lambda_i, \lambda_i^{'})$ so as to maximize their sum, $\lambda_i + \lambda_i^{'}$, which is the quantity of interest to the experimenter. Condition (18) of Theorem 3.2 is fulfilled. Thus, incentives are aligned under the new design. Intuitively, the new design allows all agents to benefit from spillovers. For example, in the previous design,

[8]The normality of T follows from normality of Y. The expected value of T is $\mathbb{E}(T) = \mathbb{E}(BC^{-1}Y) = \mathbb{E}(BC^{-1}C\mathbf{A}) = B\mathbf{A}$, and its variance is $\mathbb{V}\text{ar}(T) = \mathbb{V}\text{ar}(BC^{-1}Y) = BC^{-1}\mathbb{V}\text{ar}(Y)(C^{-1})^{\intercal}B^{\intercal} = BC^{-1}(D_{\mathbf{A}}/m)(C^{-1})^{\intercal}B^{\intercal}$.

agent 1 could not benefit from the spillover of seed set 1 to test set 2, because agent 1's score was calculated only on test set 1. However, in the new design, the score of agent 1 includes outcomes from units in the test set G_{21}, which receives spillovers from seed set 1.

6. CONCLUSION

We introduced game theory into experiments where the treatments are determined by actions of strategic agents, and where treatments can interfere with each other. The goal of the experiment is to estimate the agent that is best with respect to a quantity of interest, defined in a context *without* competition; e.g., average number of conversions from the agent's algorithm for viral marketing. However, statistical estimation of the best agent is based on experiment data, generated *with* competition among agents. Thus, the game-theoretic setting poses new challenges to the statistical analysis of experiment data, and may often invalidate well-established experimental design methods. The goal of incentive-compatible experimental design is to promote behaviors by agents that accord to the natural actions the agents would take in the experiment if there was no competition.

When agent actions do not interfere with each other, we showed that incentive-compatible designs are possible through variance-stabilizing transformations of statistics that estimate how agent would perform without competition. Furthermore, we proved a result suggesting that variance stabilization might, more generally, lead to more powerful incentive-compatible experiment designs, in which better agents have higher chances of winning. In the presence of interference, we showed that more elaborate designs are generally necessary to obtain statistics that estimate agent performances. In the context of a viral marketing application, we showed how a better design can be constructed that can account for interference among agents, e.g., when agents are able to free-ride on the advertising campaign of other agents.

REFERENCES

ATHEY, S., LEVIN, J., AND SEIRA, E. 2008. Comparing open and sealed bid auctions: Evidence from timber auctions. Tech. rep., National Bureau of Economic Research.

BESAG, J. AND KEMPTON, R. 1986. Statistical analysis of field experiments using neighbouring plots. *Biometrics*, 231–251.

BICKEL, P. AND DOKSUM, K. 2001. *Mathematical Statistics: Basic Ideas and Selected Topics*. Number v. 1 in Holden-Day series in probability and statistics. Prentice Hall.

BOX, G. E., HUNTER, W. G., HUNTER, J. S., ET AL. 1978. Statistics for experimenters.

COX, C. 1998. Delta method. *Encyclopedia of biostatistics*.

COX, D. R. AND REID, N. 2000. *The theory of the design of experiments*. CRC Press.

DASH, D. AND DRUZDZEL, M. 2001. Caveats for causal reasoning with equilibrium models. In *Symbolic and Quantitative Approaches to Reasoning with Uncertainty*. Springer, 192–203.

DAVID, O. AND KEMPTON, R. A. 1996. Designs for interference. *Biometrics*, 597–606.

PEARL, J. 2000. *Causality: models, reasoning and inference*. Vol. 29. Cambridge Univ Press.

RUBIN, D. B. 1980. Comment. *Journal of the American Statistical Association 75,* 371, 591–593.

TOULIS, P. AND KAO, E. 2013. Estimation of causal peer influence effects. In *Proceedings of The 30th International Conference on Machine Learning*. 1489–1497.

TOULIS, P. AND PARKES, D. C. 2015. Long-term causal effects of interventions in multiagent economic mechanisms. *arXiv preprint arXiv:1501.02315*.

TOULIS, P., PARKES, D. C., PFEFFER, E., ZOU, J., AND GILDOR, G. 2014. Incentive-compatible experiment design (extended abstract). In *Conference on Digital Experimentation (CODE@MIT, 2014)*.

Online Appendix to: Incentive-Compatible Experimental Design

PANOS TOULIS, Harvard University, Department of Statistics
DAVID C. PARKES, Harvard University, SEAS
ELERY PFEFFER, Harvard University, SEAS
JAMES ZOU, Microsoft Research

A. EXTENSION TO MULTIPLE BLOCKS

In this paper, our theory is developed and applied assuming only one block. However, it is straightforward to extend it to multiple blocks in a typical blocking experiment design. In this section, we give an outline of this extension.

The treatment assignment rule ψ now groups units into B blocks based on their covariates, and then randomizes treatment (i.e., the assignment of units to agents) within the blocks; blocking is performed in a deterministic way based on the publicly known covariates $\{X_u\}$, for each unit u. Formally, rule ψ is a probability distribution over the space of pairs of binary matrices $\Psi \stackrel{\text{def}}{=} (\{0,1\}^{m\times B}, \{0,1\}^{m\times n})$.

A pair $(W, Z) \in \Psi$ is called a *treatment assignment*, and has the following interpretation. The element $W_{ub} = 1$ if unit u is assigned to block b, and it is 0 otherwise. Similarly, $Z_{ui} = 1$ if unit u is assigned to agent i, and it is 0 otherwise. Using dot-notation $W_{\cdot b}$ is the bth column of matrix W, $W_{u\cdot}$ is the uth row of W as a $B \times 1$ vector, and $W_{\cdot\cdot} \equiv W$. Similarly for Z and other matrices. Finally the notation $(W, Z) \sim \psi$ will denote a treatment assignment $(W, Z) \in \Psi$, that is sampled according to rule ψ.

Example A1. Consider four experimental units (consumers) and two treatments (marketing agents) that an experimenter wishes to evaluate. In particular, the experimenter is interested to estimate which agent can achieve the highest number of sales. Suppose that, for each unit u, the experimenter and the agents know the marriage status (only covariate). We assume that units $\{1, 2\}$ are not married and $\{3, 4\}$ are, and these correspond to the two blocks $b \in \{1, 2\}$. The experimenter suspects that the outcomes differ systematically based on marriage status, and randomizes treatment within blocks. This design corresponds to treatment assignment rule ψ which samples with equal probability 1/4 from the treatment assignments $\{W, Z\}$ where $Z \in \left\{ \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{pmatrix} \right\}$ and $W = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}$ is the matrix that indicates the blocking. Some examples of dot-notation follow: $W_{1\cdot} = (1\,0)^\intercal$ is the assignment of unit u over blocks, $W_{\cdot 2} = (0\,0\,1\,1)^\intercal$ is the assignment over units in block 2, etc.

With multiple blocks, agents are allowed to play different actions across blocks. We would thus write A_{ib} for the action of agent i in block b, and $\mathcal{A}_{ib}$ for the action space of this action.

With multiple blocks, there is also an additional block index for the potential and observed outcomes. For example, Y^{obs}_{ubi} is now the observed outcome of unit u assigned to block b and agent i; with dot-notation, $Y^{\text{obs}}_{\cdot b\cdot}$ denotes the observed outcomes of units in block b. The experiment design $\mathcal{D}$ has now multiple score functions, ϕ_b, one per block. For example, $\phi_{ib}(Y^{\text{obs}}_{\cdot b\cdot})$ is the score of agent i in block b with data $Y^{\text{obs}}_{\cdot b\cdot}$. Similar extensions are straightforward for the concepts of performance, natural action, and quality.

Given block-specific score functions, the winner of the experiment is the agent who won the majority of blocks, ignoring ties. When there is no interference across and within-blocks, then the experimenter can design an incentive-compatible design within each block using Theorem

DOI 10.1145/2764468.2764525 http://doi.acm.org/10.1145/2764468.2764525

3.2. In this case, each block would have a separate identifying statistic. When the action space of an agent is the product space of the block action spaces, the agent will prefer to maximize its winning probability within each block. Therefore, the incentive-compatibility results of Theorems 3.2 and 4.2 can be readily applied. The same results can be applied in the problem with interference, assuming that there is no between-block interference, i.e., an action of agent i in block b does not affect the outcomes for agent j in some other block b'.

B. PROOFS

THEOREM 3.2. *Fix agent actions* $\mathbf{A}$*, and consider design* $\mathcal{D} = (\psi, \phi)$ *that has an identifying statistic* T *with covariance matrix* $\Sigma(\mathbf{A})$*. Let* $\phi_i(Y_{\cdot\cdot}^{\text{obs}}) = f(T_i)$ *for some function* $f : \mathbb{R} \to \mathbb{R}$*, and let* $v_{ij}(\mathbf{A})$ *be the* ij*th element of* $V(\mathbf{A})$ *defined in Eq.* (16)*. Also define,*

$$v_f^{ij}(\alpha|\mathbf{A}_{-i}) = v_{ii}(\alpha, \mathbf{A}_{-i}) + v_{jj}(\alpha, \mathbf{A}_{-i}) - v_{ij}(\alpha, \mathbf{A}_{-i}) - v_{ji}(\alpha, \mathbf{A}_{-i}).$$

The design $\mathcal{D}$ *is incentive-compatible, if, for every agent* i*,*

$$\arg\max_{\alpha_i \in \mathcal{A}_i} \left\{ \frac{f(\chi(\alpha_i))}{v_f^{ij}(\alpha_i|\mathbf{A}_{-i})^{1/2}} \right\} = \arg\max_{\alpha_i \in \mathcal{A}_i} \{\chi(\alpha_i)\} \stackrel{\text{def}}{=} A_i^\star,$$

for every agent j*, and all actions* $\mathbf{A}_{-i}$*. In such case, we say that* T *is aligned with performance* χ *through score* ϕ*.*

PROOF. For a vector $x \in \mathbb{R}^n$, let $f(x) = (f(x_1), f(x_2), \ldots, f(x_n))^\intercal$. From the Delta theorem [Bickel and Doksum 2001; Cox 1998], and the asymptotic property (15) of the identifying statistic T, we obtain

$$\sqrt{k}\,(f(T) - f(\boldsymbol{\chi}(\mathbf{A}))) \xrightarrow{D} \mathcal{N}(0, \mathcal{J}_\phi \Sigma(\mathbf{A}) \mathcal{J}_\phi^\intercal), \tag{44}$$

where $\mathcal{J}_\phi$ is the Jacobian of f at $\boldsymbol{\chi}(\mathbf{A})$ (by definition, this is a diagonal matrix). The probability that agent i wins over j is equal to

$$\Pr\left(\phi_i(Y_{\cdot\cdot}^{\text{obs}}) > \phi_j(Y_{\cdot\cdot}^{\text{obs}})\right) = \Pr\left(c^\intercal f(T) > 0\right), \tag{45}$$

where $c = (0, \ldots, 1, 0, \ldots, -1, 0, \ldots)^\intercal$, is a $n \times 1$ vector, with zero elements, except for $c_i = 1$ and $c_j = -1$. Using Eq. (44), we have

$$\sqrt{k}\,(c^\intercal f(T) - c^\intercal f(\boldsymbol{\chi}(\mathbf{A}))) \xrightarrow{D} \mathcal{N}(0, c^\intercal \mathcal{J}_\phi \Sigma(\mathbf{A}) \mathcal{J}_\phi^\intercal c). \tag{46}$$

From (46), probability (45) becomes

$$\Pr\left(\phi_i(Y_{\cdot\cdot}^{\text{obs}}) > \phi_j(Y_{\cdot\cdot}^{\text{obs}})\right) = \Phi\left(\frac{f_i(\boldsymbol{\chi}(\mathbf{A})) - f_j(\boldsymbol{\chi}(\mathbf{A}))}{v_f^{ij}(\mathbf{A})^{1/2}}\right) = \Phi\left(\frac{\chi(A_i) - \chi(A_j)}{v_f^{ij}(\mathbf{A})^{1/2}}\right),$$

where $v_f^{ij}(\mathbf{A})$ is given in Eq. (17). Therefore, agent i maximizes its winning probability by playing the natural action, by property (18). □

THEOREM 4.2. *Consider design* $\mathcal{D} = (\psi, \phi)$ *with an identifying statistic* T *with covariance matrix* $\Sigma(\mathbf{A})$*. Suppose Assumption 1 holds. If, for every agent* i*,*

$$\begin{aligned}
&\phi_i(Y_{\cdot\cdot}^{\text{obs}}) \equiv f(T_i), \text{ where } f : \mathbb{R} \to \mathbb{R}, \\
&\mathbb{V}\text{ar}(\phi_i(Y_{\cdot\cdot}^{\text{obs}})) = \text{const.}, \\
&\arg\max_{\alpha_i \in \mathcal{A}_i} f(\chi(\alpha_i)) = \arg\max_{\alpha_i \in \mathcal{A}_i} \{\chi(\alpha_i)\} \stackrel{\text{def}}{=} A_i^\star,
\end{aligned}$$

then design $\mathcal{D}$ *is incentive-compatible.*

PROOF. By Assumption 1 (no interference), $\Sigma(\mathbf{A})$ is diagonal; let $\Sigma(\mathbf{A}) = \text{diag}(\sigma_{ii}^2(\mathbf{A}))$. Then, from Theorem (4.2) and Condition (23),

$$\mathbb{V}\text{ar}(\phi_i(Y_{\cdot\cdot}^{\text{obs}})) = f'(\chi(A_i))^2 \sigma_{ii}^2(\mathbf{A}) = c,$$

for some constant $c > 0$. Also by Condition (23), the Jacobian of ϕ at $\mathbf{A}$, is given by $\mathcal{J}_\phi = \mathrm{diag}(f'(\chi(A_i)))$. Using the notation of Theorem 3.2,

$$V(\mathbf{A}) = \mathcal{J}_\phi \Sigma(\mathbf{A}) \mathcal{J}_\phi^\intercal = \mathrm{diag}(f'(\chi(A_i))^2 \sigma_{ii}^2(\mathbf{A})) = c\mathbb{I}.$$

It follows, $v_f^{ij}(\alpha|\mathbf{A}_{-i}) = 2c$ for any i, j, where v_f^{ij} is defined in Eq. (17), Theorem 3.2. Using Condition (25),

$$\arg\max_{\alpha_i \in \mathcal{A}_i} \left\{ \frac{f(\chi(\alpha_i))}{v_f^{ij}(\alpha_i|\mathbf{A}_{-i})^{1/2}} \right\} = (1/2c) \arg\max_{\alpha_i \in \mathcal{A}_i} \{\chi(\alpha_i)\} = A_i^\star.$$

Thus, all conditions of Theorem 3.2 are fulfilled, and the design $\mathcal{D}$ is incentive-compatible. □

THEOREM 4.4. *Consider an incentive-compatible design $\mathcal{D} = (\psi, \phi)$, where action sets $\mathcal{A}_i \subseteq \mathbb{R}$ are compact, and performance χ is one-to-one and continuous. Let,*

$$\sqrt{k}\left(\phi_i(Y_{\cdot\cdot}^{\mathrm{obs}}) - \chi(A_i)\right) \xrightarrow{D} \mathcal{N}(0, \sigma^2(A_i)),$$

where function $\sigma^2 : \mathcal{A} \to \mathbb{R}^+$ satisfies

$$\chi(\alpha_i') \geq \chi(\alpha_i) \Rightarrow \sigma^2(\alpha_i') \geq \sigma^2(\alpha_i),$$

for every agent i, and all actions $\alpha_i', \alpha_i \in \mathcal{A}_i$. Consider a design $\mathcal{D}' = (\psi, \phi')$, where $\phi_i'(Y_{\cdot\cdot}^{\mathrm{obs}}) = \nu(\phi_i(Y_{\cdot\cdot}^{\mathrm{obs}}))$, for each agent i, with $\nu(\cdot)$ defined by

$$\nu(y) = \int^y \frac{1}{\sqrt{\sigma^2(\chi^{-1}(z))}} dz.$$

Then, design $\mathcal{D}'$ is incentive-compatible and more powerful than $\mathcal{D}$, if $\nu(\cdot)$ is convex, or $1/\sqrt{\sigma^2(\chi^{-1}(\cdot))}$ and $\sigma^2(\chi^{-1}(\cdot))$ are both convex.

PROOF. From the univariate Delta theorem,

$$\sqrt{k}\left(\nu(\phi_i(Y_{\cdot\cdot}^{\mathrm{obs}})) - \nu(\chi(A_i))\right) \xrightarrow{D} \mathcal{N}(0, 1),$$

since $\nu'(\chi(A_i))^2 \sigma^2(A_i) = 1$, by Eq. (31). For brevity, set $\chi(A_i) \stackrel{\text{def}}{=} \chi_i$ and $\sigma^2(A_i) \stackrel{\text{def}}{=} \sigma_i^2$. Without loss of generality, assume $\chi_i \geq \chi_j$. The probability that agent i wins over agent j in design $\mathcal{D}'$ is equal to,

$$P_1(\mathbf{A}|\mathcal{D}') = \Phi\left(\sqrt{k/2}(\nu(\chi_i) - \nu(\chi_j))\right).$$

In the old design, $\mathcal{D}$, this probability is equal to

$$P_1(\mathbf{A}|\mathcal{D}) = \Phi\left(\sqrt{k}\frac{\chi_i - \chi_j}{\sqrt{\sigma_i^2 + \sigma_j^2}}\right).$$

Case 1 – Convex $\nu(\cdot)$. By convexity of ν we have

$$\frac{\nu(\chi_i) - \nu(\chi_j)}{\chi_i - \chi_j} \geq \nu'(\chi_j). \tag{47}$$

By definition (29), $\nu'(\chi_j)^2 \sigma_j^2 = 1$. By property (30), $\sigma_i^2 \geq \sigma_j^2$ since $\chi_i \geq \chi_j$. Hence, $\nu'(\chi_i)^2 \sigma_i^2 = 1 \Rightarrow \nu'(\chi_i)^2 \leq \nu'(\chi_j)^2$. It follows,

$$\nu'(\chi_j)^2 \sigma_j^2 + \nu'(\chi_j)^2 \sigma_i^2 \geq 2 \Rightarrow$$
$$\nu'(\chi_j) \geq \sqrt{\frac{2}{\sigma_i^2 + \sigma_j^2}}. \tag{48}$$

Combining (47) and (48), we obtain

$$\frac{\nu(\chi_i)-\nu(\chi_j)}{\sqrt{2}} \geq \frac{\chi_i-\chi_j}{\sqrt{\sigma_i^2+\sigma_j^2}} \Rightarrow \Phi\left(\sqrt{k/2}(\nu(\chi_i)-\nu(\chi_j))\right) \geq \Phi\left(\sqrt{k}\frac{\chi_i-\chi_j}{\sqrt{\sigma_i^2+\sigma_j^2}}\right),$$

which implies that design $\mathcal{D}'$ is more powerful than $\mathcal{D}$.

Case 2 – $1/\sqrt{\sigma^2(\chi^{-1}(\cdot))}$ *and* $\sigma^2(\chi^{-1}(\cdot))$ *are both convex.* It holds,

$$\begin{aligned}\frac{\nu(\chi_i)-\nu(\chi_j)}{\chi_i-\chi_j} &= \frac{1}{\chi_i-\chi_j}\int_{\chi_j}^{\chi_i}\frac{1}{\sqrt{\sigma^2(\chi^{-1}(z))}}dz \geq \frac{1}{\sqrt{\sigma^2(\chi^{-1}((\chi_j+\chi_i)/2))}} \\ &\geq \frac{1}{\sqrt{\sigma^2(\chi^{-1}(\chi_j))/2+\sigma^2(\chi^{-1}(\chi_i))/2}} \stackrel{\text{def}}{=} \sqrt{\frac{2}{\sigma_i^2+\sigma_j^2}}.\end{aligned}$$

The first inequalty is obtained by convexity of $1/\sqrt{\sigma^2(\chi^{-1}(\cdot))}$, and the second by convexity of $\sigma^2(\chi^{-1}(\cdot))$. To finish the proof we follow the same arguments as in Case 1. □

C. REMARKS ON VARIANCE STABILIZATION

In Theorem 4.2, the variance of the score functions ϕ_i is stabilized (made constant) through a transformation f. Such transformations that stabilize the variance of a statistic, are known as *variance-stabilizing* transformations in statistics, and they are of fundamental importance in various tasks, such as hypothesis testing and estimation. For example, consider a sample average of n independent Poisson random variables with mean λ. The asymptotic distribution of the sample average is $\bar{Y} \sim \text{Poisson}(\lambda/n)$. In the limit, $\sqrt{n}(\bar{Y}-\lambda) \xrightarrow{D} \mathcal{N}(0,\lambda)$. This asymptotic result is not useful to construct a confidence interval for the unknown parameter λ because the variance of $\bar{Y}$ depends on that unknown parameter. However, through the Delta theorem, $2\sqrt{n}(\sqrt{\bar{Y}}-\sqrt{\lambda}) \xrightarrow{D} \mathcal{N}(0,1)$ i.e., the variance of $\sqrt{\bar{Y}}$ is constant; the statistic $\sqrt{\bar{Y}}$ can be used to obtain *exact* confidence intervals for λ.

In our setting, the variance stabilization helps to mitigate the risk-return trade-off that strategic agents can undertake in an experiment. Loosely speaking, when the variance is stabilized a worse agent cannot benefit by being more risky, and a better agent cannot benefit by being more conservative. Rather, incentives are aligned such that every agent will do its best, assuming the remaining conditions of Theorem 4.2 are fulfilled.

D. DISCUSSION

Our approach to design incentive-compatible experiments has been through the use of an identifying statistic, i.e., a statistic that can estimate the agent performances without competition. In many situations, such a statistic exists, e.g., by using sample summaries (means, variances, etc), and then appealing to the central limit theorem. In most realistic cases, a key assumption will be that the outcomes have a known parametric form. In this paper, we made such parametric assumptions in our viral marketing example.

However, an experimenter might be unwilling to make such parametric modeling assumptions. An alternative would then be either to use a nonparametric test for the quantities of interest (i.e., agent performances), or a randomization-based analysis. The former includes a wide-class of nonparametric methods, and we plan to investigate it in future work. It should be noted, however, that even nonparametric tests have crucial underlying assumptions, e.g., exchangeability of observed data, that are not easy to validate. In many situations, such assumptions are more critical than, for example, normality assumptions that can be quite robust under many scenarios [Box et al. 1978, Appendix 3A]. The latter method of randomization-based analysis usually starts from a null hypothesis which aims to provide evidence for the likelihood of certain observed quantities, e.g., through p-values. However, it is hard to test such hypotheses in our setting because agents can freely choose the versions of the treatment to apply. Therefore, one cannot use the null hypothesis to *impute* counterfactuals, i.e., outcomes that would have been observed under a different randomization because agents act in a strategic, non-random way.

In the case with interference, the assumption that an identifying statistic exists has two components. First, it is required that the experimenter has a good idea about the *model* of interference, e.g., that an agent action affects the outcomes for another agent linearly, as in Example 3(c). Assumptions on the model of interference are frequent in practice because they help to deal with interference after the experiment has been performed [Besag and Kempton 1986]. Second, it is required that the experimenter knows exactly the hyperparameters of the assumed interference model. In the viral marketing problem of Section 5, a scalar parameter γ was used to model interference. In our examples, we assumed that γ was known. One way to avoid this problem is to treat such parameters of interference as *nuisance* parameters, and then use a suitable statistical method; e.g., use profile likelihood instead of the true, but unknown, likelihood to obtain proxies for the maximum-likelihood estimates. A Bayesian approach would be to set priors for such parameters and then obtain a posterior predictive distribution for the unknown agent performances. Agents would then be scored according to this posterior distribution, but this would not alter the core of our methodology.