Incentive-compatible Experiment Design
(Extended abstract)
Panos Toulis (Harvard Statistics), David C. Parkes (Harvard SEAS), Elery Pfeffer (Pursway, Inc.), James Zou (Microsoft Research), Guy Gildor (Pursway, Inc.)

1. Introduction

Experiments are often the gold-standard for answering a research question. Typically, we think of experimental units as being inanimate objects and the goal of the experiment design is to avoid systematic biases and minimize random errors (Cox & Reid, 2000). Here we propose a setting where the experimental units are strategic agents whose behavior interferes with the experiment. The goal of incentive-compatible experiment design is to design an experiment in a way to promote behavior that is aligned with the behavior of an agent as if it was the only actor in the system.

In this work, we pursue this research agenda by considering the problem of comparing the quality of two viral marketing firms (agents). In particular, we assume that there is a company that wishes to promote a product to a population of individuals (experimental units) by targeting a subset of the population, called the seed set, through a marketing campaign. Two agents, say A and B, claim to possess knowledge about the influence network among individuals, which is hidden to the marketing company, and claim to be able to utilize this knowledge to optimally pick a seed set that will maximize product adoption. Thus, the question we ask is the following: “How to design an experiment that will decide which agent is better able to pick a seed set that will maximize product adoption?” As part of this, we may also be interested to understand which agent has the best knowledge of the influence network, and quantify our uncertainty.

To answer this question, ideally an experimenter would test agents A and B on two separate, disjoint populations that are identical in every other aspect. They should be disjoint in the sense that marketing to individuals in one part does not affect the purchasing decisions of individuals in the other part. Beyond achieving disjointedness, a practical problem is that the experimenter may have limited knowledge about the factors that matter in order to choose balanced populations. Furthermore, there could be geography-specific conditions such as weather or sports events that are hard to predict or know about and affect the outcome. Even if the experimenter has this kind of knowledge, the agents may not believe that the partition is fair, making it challenging to gain acceptance of the results.

Recognizing these challenges, we focus on conducting an experiment without requiring two disjoint (i.e., non-interacting), yet otherwise identical populations. Rather, we simply assume access to individuals in a single population. This requires us to address new challenges that arise because of the strategic interference between agents A and B, as they compete in order to win in the experiment, recognizing interactions between individuals over the influence network underlying the population. In this setting, the kind of concern that we must address is that an agent (say agent A) may free-ride on the influence provided by the seed set of agent B, which in turn can lead agent B to be strategic in choosing the seed set. Note that we assume that there is no way to know that an adoption by an individual is because of targeting by agent A or agent B i.e., there is no tracking. This is because adoption can happen by
navigating to a web site, or going into a store, and quite apart from clicking on a particular message. This strategic behavior may cause an experiment design to be ineffective in assessing the true quality of the agents i.e., the quality that would be observed without strategic interactions. More generally, we’re interested to mitigate the effects of such strategic interactions by promoting equilibrium behavior that approximates an ideal experiment without interactions.

2. Experiment Designs

In addition to asking agents to select a seed set, we may also ask agents to select a test set. A test set contains the individuals in the population that will be used to assess the performance of the agent in selecting influential seed sets. In order to reduce the space of possible experiment designs, we focus our attention on designs that are defined under the following operational constraints. First, either the experimenter predefines a seed set that is used by both agents, or each agent picks at most one seed set. Second, each agent picks at least one test set with knowledge of the seed set(s). Test sets are required to be disjoint from the seed set(s), and both sets have a predefined size. Finally, the experimenter targets individuals in the seed set(s) through a promotional campaign. After a predefined amount of time, product adoption outcomes are observed on the individuals (or units) in the test sets. A unit i in a test set has outcome Yi = 1 if she adopts the product, and Yi = 0 otherwise. Each agent is evaluated according to the total adoption in its own reported test set(s).

In our initial study, we analyze the following experiment designs.

1. “Fixed Seed, One Test” (M1): The experimenter selects a seed set S, and each agent is asked to pick a test set that is disjoint from S.
2. “Split, Variable Seed, One Test” (M2): The experimenter randomly places each unit into set VA or VB. Agent A picks a seed set from VA and agent B picks a seed set from VB. With knowledge of these, agent A picks a test set (disjoint from its seed) from VA and agent B picks a test set (disjoint from its seed) from VB.
3. “Split, Variable Seed, Two Test” (M3): Seed sets are as in M2. With knowledge of these, agent A picks a test set (disjoint from its seed) from VA and a test set (disjoint from B’s seed) from VB, and Agent B picks a test set (disjoint from its seed) from VB and a test set (disjoint from A’s seed) from VA.

3. Influence Model

To get further traction on the problem, we make modeling assumptions in regard to the effect of influence on product adoptions Y. For any given seed set S, we assume that there is an associated intensity λs that captures the amount of influence that units in the set S have on the rest of the population or a subset of it. In particular, the parameter λs affects the characteristics of the influence network G between units in the seed and a test set(s). The product adoption outcomes Y in the test sets are defined according to a distribution conditional on G. In this work, we assume the following data-generation process that is an instance of the aforementioned model:

- Given a seed set S, each unit i has ni incoming edges originating from units in S, such that ni ~ Poisson(λs).
The outcome \( Y_i \) of unit \( i \) is a Bernoulli trial with probability \( 1 - \exp(\varphi_0 + \varphi_i + N_i) \), where \((\varphi_0, \varphi_i)\) are fixed model parameters.

- An agent \( j \) has an associated quality \( p_j \), with \( 0 < p_j < 1 \), that allows it to observe a noisy version \( N_i \) of the #incoming edges in a unit \( i \), such that \( N_i \sim Binom(n, p) \).
- An agent that selects a test set \( T \) using observations \( \{N_i\} \), achieves a score \( u \) that is given by \( u = \sum_{i \in T} Y_i \).

In this model, agents differ qualitatively in two aspects. First, better agents can pick seed sets with stochastically higher intensities \( \lambda \). Second, better agents have better knowledge of the underlying influence network \( G \), and can thus observe more edges on average. In the split designs \( M_1 \) and \( M_2 \) we further assume:

- An agent \( j \) can pick a seed set that realizes a pair of intensities \((\lambda_{j\varphi}, \lambda_{j\mu})\) that lie on a domain that is restricted by the topology of the influence network \( G \); \( \lambda_{j\varphi} \) is the intensity of the agent’s selected seed set on the set \( V_j \) on one half, and \( \lambda_{j\mu} \) is the intensity of the same seed set onto the other half.

Our modeling assumptions have the following implications on the estimation problem: (i) higher intensities mean higher product adoptions on average, (ii) better knowledge of the influence network \( G \) means better outcomes \( Y \) on average, and (iii) very low or very high product adoption rates, i.e., averages of outcomes \( Y \), hurt the statistical efficiency in estimating agents’ abilities (e.g. knowledge of the network \( G \)).

Incentive-compatible experiment design presents unique challenges to the experimenter because the game-theoretic aspects of the experiment are conflicted with its role as a statistical estimation or hypothesis testing device.

4. Game-theoretic challenges

Let’s first consider \( M_0 \). Each agent wants to win the contest, which requires achieving the highest score. Fixing the seed set, we define the straightforward strategy as the action where an agent ranks the test units based on their observed incoming edges \( N_i \), and reports the highest-ranked units as its test set. Other strategies are possible, for example trying to pick a test set that is more risky, but with more potential for upside and thus winning the contest.

**Theorem 1.** In the fixed-seed experiment design \( M_0 \), straightforward behavior in regard to test-set selection is a dominant-strategy equilibrium.

**Proof (sketch).** For any two units \( i, i' \) in the test set such that the observed #incoming edges satisfy \( N_i > N_{i'} \), the outcome \( Y_i \) stochastically dominates \( Y_{i'} \). Therefore, an agent prefers to include unit \( i \) in the test set, rather than unit \( i' \), regardless of the choice of the other agent. Thus, selecting units with the highest observed incoming edges is a dominant-strategy.∎

But the \( M_0 \) design does not allow the estimation of agents’ ability to pick seeds, and recall that the main goal is to design an experiment that will decide which agent is better able to pick a seed set that will maximize product adoption. Let’s consider the \( M_1 \) and \( M_2 \) designs. Agents are also asked to pick seed
sets, and so these designs can in principle be used to estimate agents’ seed-selection ability. However, the designs have different incentive properties. For example, an agent in the $M_1$ design may prefer to pick a seed set that both influences individuals in its own half but does not tend to influence individuals in the other half of the population. This recognizes that the other agent may try to free-ride on its seeding decision. By contrast, an agent in the $M_2$ design may prefer to pick a seed set that is effective in influencing individuals irrespective of whether they are in set $V_A$ or set $V_B$ because it might have the same ability to benefit from this as the other agent.

We present initial empirical results that support this intuition. In a simulation study, we assume a population of 20,000 units, and agents A and B that pick 100 test units in each half. We set agent A to have higher quality ($p_A = 0.3$) than agent B ($p_B = 0.2$). Both agents can pick a pair of intensities $(\lambda_{A1}, \lambda_{A2})$ that lies in an ellipse as in Figure 1, to reflect the symmetry of the two halves induced by the randomization; Agent B is further fixed to a strategy $(\lambda_{B1}, \lambda_{B2})$ such that $\lambda_{B1} > \lambda_{B2}$, and both are slightly smaller than the values considered for agent A in order to reflect agent B’s lower quality. For every strategy profile $(\lambda_{A1}, \lambda_{A2}, \lambda_{B1}, \lambda_{B2})$, we generate the data through the process of Section 3, and calculate how often agents A and B win; these probabilities are shown in Figure 1.

![Figure 1](image.png)  

**Figure 1.** Probabilities of A winning/losing in mechanisms $M_1$ and $M_2$. Green color indicates higher probability and red color indicates lower probability. The asterisk *** marks the point of maximum winning probability. In $M_1$, a better agent prefers to maximize the influence on its own half, and minimize the influence in the other half, thus preferring points in lower-right of the action space $(\lambda_{A1}, \lambda_{A2})$. In $M_2$, a better agent prefers actions that maximize influence in both halves, thus preferring points in the upper-left of the action space.

We observe in design $M_1$ that the better agent prefers reports in the bottom-right of the action space; this corresponds to high $\lambda_{A1}$ and low $\lambda_{A2}$, consistent with the intuition that a better agent would try to select a seed set with high influence on its own half, and low influence on the other half (roughly maximizing $\lambda_{A1} - \lambda_{A2}$). In contrast, in $M_2$ the better agent is happy with intensities $\lambda_{A1}$ and $\lambda_{A2}$ that are high in both halves, roughly maximizing $\lambda_{A1} + \lambda_{A2}$. Thus, design $M_2$ seems to promote a more straightforward behavior than design $M_1$.

5. Statistical challenges
A basic design parameter is the size of test sets. One striking and counterintuitive finding is that the discriminative power of the experiments can be nonmonotonic in test set size. To illustrate, we study the $M_b$ design with 2000 units and follow a simple data-generation process similar to Section 3 (we exclude sampling $Y_j$ for simplicity). In particular, each unit $i$, $i=1, \ldots, 2000$, samples a latent value (i.e., incoming edges) $n_i = \{0, 1 \text{ or } 2\}$ with equal probability $1/3$. Agents have quality $0 < p_j < 1$, $j = A$ or $B$, such that they observe noisy unit values as $N_j \sim \text{Binomial}(n, p)$; thus, higher values for $p_j$ means more knowledgeable agents. Agents are asked to report their highest-ranked units (straightforward behavior), and their score is the sum of the actual values $n_i$ of their reported units.

See Figure 2, showing results for 1000 trials, where both agents adopt the straightforward behavior. We observe what we refer to as the goldfish paradox, noting that a larger sample size (i.e. test set size) can reduce the power of the experiment, which is atypical of classical experiment design in statistics. The intuition is that in the range of 100-550 test units, $A$ and $B$ select from pools of observed moderate-valued unit; i.e., they pick randomly from $\Delta_A = \{i : N_{j_i} = 1\}$ and $\Delta_B = \{i : N_{j_i} = 1\}$ respectively. Since $B$ is a lower-quality agent, $\Delta_B$ contains a higher proportion of actual high-valued units (i.e., units with $n_i = 2$) than $\Delta_A$. Until the pool $\Delta_B$ is exhausted, the chances of agent $B$ winning the game keep increasing relative to the chances of the higher-quality agent $A$.

Figure 2. The goldfish paradox. The green line is the probability that higher-quality agent $A$ ($p_a=0.3$) beats lower-quality agent $B$ ($p_b=0.2$); the red line is the probability that $B$ beats $A$. We observe a nonmonotonic effect of test-set size (for 100-550 test units, the chances of $B$ winning increase relative to $A$.)

Theorem 2. The goldfish paradox exists for any pair $(p_a, p_b)$ of positive agent qualities and any number of possible discrete latent values.

6. Summary

We introduce the problem of incentive-compatible experiment design as the process of designing mechanisms that promote equilibrium behavior that approximate a classical experiment without strategic interactions. In current work we are developing a game-theoretic analysis of our proposed designs $M_1$ and $M_2$, and variants. Our long-term goal is to analyze classical experiment designs, such as blocking or factorial designs, in this incentive-compatible design framework.

7. References