# Balancing External Representativity and Networked Interference in Large-Scale Rural Experiments[1]

Alejandro Noriega and Alex Pentland*†

*Corresponding: {noriega, pentland} @mit.edu

**This paper presents a methodology for a) modeling the geo-spatial and social interaction factors that drive interference (SUTVA violations) in randomized field experiments; and b) eliciting a set of non-dominated sample options that approximate the Pareto-optimal tradeoff between interference and external representativity as functions of sample choice. We develop and test the methodology in the context of a large-scale health experiment in rural Mexico, involving more than 5,000 pregnant women and 600 health clinics across five states. Relevant for the practitioner, we show the methodology is computationally tractable and can be implemented leveraging novel open sourced geo-spatial data and software tools.**

[1] This document is a draft paper accepted for presentation at the Annual Bank Conference on Development Economics (ABCDE), World Bank 2016. Improvements in synthesis and presentation are yet to be implemented.

†Alejandro Noriega is a PhD candidate at the MIT Human Dynamics Laboratory, and MS. in Technology and Policy at the MIT Institute for Data, Systems and Society (IDSS). Alex Pentland is director of the MIT Human Dynamics Laboratory, Academic Director of the Harvard-MIT-ODI DataPop Alliance, and Faculty Director of the MIT Connection Science Research Initiative.

# 1 Introduction

Interference in experimental studies refers to situations where subjects' outcomes may depend on their own treatment status as well as on the treatment status of other subjects. It is not uncommon to encounter this phenomena in the context of social and health sciences. Community health effects of vaccination, effects of police patrolling policies on displaced criminal activity across neighborhoods, social mobilization in political and marketing campaigns (see (*1*), for a famous example), and effects of auditing policies on corruption across bureaucratic units, are examples of social phenomena where interference or indirect effects are likely to occur (*2*).[1]

Neglecting the presence of interference in field experiments may cause bias and increase variance of causal estimators (*3*). Approaches for dealing with interference include: multilevel designs, where indirect effects are explicitly modeled and estimated (*4*); designs where researchers attempt to isolate experimental units from indirect effects; and cluster randomization designs, where units are grouped in fairly self-contained clusters (*5*) and randomization is implemented at the cluster level (*6*).[2]

On the one hand, today's 'big data' fosters our ability to extrapolate by means of rich covariate sets (see (*7*) and (*8*) for cutting-edge formalizations and use cases). On the other, a natural strategy for coping with interference is to choose an experimental sample where units, or clusters of units, are most apart from each other interference-wise. However, interference-minimizing criteria for sample selection may compromise representativity of the experimental sample against external populations of interest. In particular, common sampling approaches amenable to extrapolation such as sampling for heterogeneity (*9*) and proportional sampling (*10*) are likely to misalign to some degree against interference-minimizing sample selection.

This paper presents a methodology for a) modeling the geo-spatial and social interaction factors that drive interference in field experiments; and b) eliciting a set of non-dominated sample options that approximate the Pareto-optimal tradeoff between interference and external representativity. We develop and test the methodology in the context of a large-scale health experiment in rural Mexico, involving more than 600 health clinics and 5,000 pregnant women. Relevant for the practitioner, we show the methodology is computationally tractable and can be implemented leveraging novel open sourced geo-spatial data and software tools.

The structure of the paper is as follows. Section 2 describes the large-scale health intervention in rural Mexico that serves as context and test platform for this research. Section 3 formally

[1] The no-interference assumption is implicit in the potential outcomes framework for causal inference when we enumerate all potential outcomes of $i$ as $Y_i(d_i) \quad \forall d_i \in D_i$, where $D_i$ is the set of possible treatment statuses of $i$.

[2] It is often the case in development research contexts that clusters are naturally defined by fundamental geographic, social, infrastructure or administrative structures, such as villages, schools, health clinics, neighborhoods, or administrative units. For example, rural health interventions are rarely implemented at the individual rather than the health clinic or physician level. This is the case of the Prospera experiment, on which this paper is focused.

describes the challenge of extrapolation in field experiments and its relation with interference in sample choice. Section 4 builds the inter-clinic interference network based on an interference gravity model, implemented through novel open sourced geo-spatial data and software tools. Section 5 introduces greedy search algorithms for eliciting non-dominated sample options that trade off among representativity and interference, and presents results on their application to the Mexico experiment. Section 6 discusses sample choice and Section 7 concludes.

# 2 A Large-Scale Health Experiment in Rural Mexico

Mexico's social assistance program Prospera is one of the largest conditional cash transfer programs in the world, and second largest in Latin America. Among other services, it provides health care to nearly 6M families and 30M beneficiaries (*11*). Most of Prospera operations occur in rural localities where physical remoteness drives key challenges in provisioning Prospera's health and financial services. Moreover, information in Prospera flows through traditional means such as fliers, radio announcements, and door-to-door communication.

Mexican national authorities[3] have endeavored to introduce digital means of communication with and among Prospera beneficiaries, based on UNESCO's RapidPro mobile platform successfully implemented in countries like Rwanda (*12*), Uganda and Sierra Leona (*13*). In this context, we have participated in designing a large-scale randomized controlled experiment to assess the effect of a potential intervention on health outcomes of Prospera beneficiaries. The experiment focuses on maternal and child health, and involves more than $600$ health clinics and $5,000$ pregnant women across five states. Three treatment arms are tested, consisting of top-down, peer-to-peer, and down-to-top feedback communications.

Design of the Prospera experiment had to address external validity. In particular, the study is meant to help assess the potential value of the program innovations national-wide. Moreover, the study had to address potential interference issues, given the large-scale and field nature of the experiment.

# 3 External Representativity and Interference in Sample Choice

Let PATE denote the *population average treatment effect*. Following Imai, King et al. 2008 (*14*), the *estimation error* of PATE can be decomposed as:

$$\Delta = \text{PATE} - D = \Delta_S + \Delta_T = \left(\Delta_{S_X} + \Delta_{S_U}\right) + \left(\Delta_{T_X} + \Delta_{T_U}\right) \tag{1}$$

[3] In partnership with a set of academic institutions and NGOs, such as the MIT Media Laboratories.

where $D$ is the baseline *difference in means* estimator, $\Delta_S$ and $\Delta_T$ are *sample selection error* and *treatment assignment error* components, and $\Delta_{T_X}$, $\Delta_{T_U}$, $\Delta_{S_X}$, and $\Delta_{S_U}$ are their subcomponents associated to observed and unobserved covariate sets $X$ and $U$.

Techniques such as weighting and stratified sampling ensure that $\Delta_{S_X} = 0$. Random sampling, in its different forms (e.g., simple or stratified random sampling), ensures that $E[\Delta_{S_U}] = 0$, that is, in expectation. However, as pointed out by Muller 2013 (*9*), sampling at random is commonly "not practically feasible, or researchers have the more ambitious aim of generalizing beyond a single, prespecified population".[4]

Extrapolation is possible from non-random samples, and is most often conducted by leveraging observed covariates $X$. The weighted or conditional ATE estimator can eliminate observed component $\Delta_{S_X}$, and unobserved component $\Delta_{S_U}$ in as much as $X$ relates to $U$ (*14*). See Bareinboim and Pearl 2015 (*7*) and Hartman et. al 2015 (*8*) for a formal framework and cutting-edge use cases of extrapolation leveraging covariates.

The richer the set of available covariates $X$, the smaller $E[\Delta_{S_U}]$ is.[5] Relevantly, in the current age of 'big data', we often have rich sets of covariates for populations of millions of individuals. In the case of Prospera, rich census and institutional data is available, providing more than 200 covariates at the clinic, village, household, and individual levels.[6] However, extrapolation based on non-randomized samples depends on common support over $X$, i.e., the sample covariate distribution must be positive in every strata $x$ where the population covariate distribution is positive (*support condition* in (*16*)). Moreover, the weighted estimator variance depends both on the within-strata variance and the within-strata sample size $n_{s_i}$.

Hence, common approaches to sampling are sampling for heterogeneity, and proportional and optimal stratified sampling (*10*). Each sampling strategy induces an ideal or *target covariate distribution*, for which we can search for an experimental sample whose covariate distribution is a close match.[7] However, samples selected on this representativity criteria alone may incur in high interference among sample units. Similarly, interference-minimizing samples may present covariate distributions that significantly differ from the population or target distribution.

Section 5 introduces a methodology for eliciting the potential tradeoff among these two objectives, and present the researcher with a set of Pareto non-dominated sample options to decide on.

[4] In the Prospera experiment for example, operational considerations constrain sample $S$ to clinics within five states. Additionally, there is interest in extrapolating results to villages that currently lack mobile connectivity, hence assessing potential health benefits of digital inclusion policies.

[5] Analogous to conditional ignorability and the *back-door-path* criteria in controlling $\Delta_{T_U}$ (see sections 4.1 and 4.2 in (*15*)).

[6] Out of which, some of the most relevant include education, indigenism, newborn weight and measures, birth defect rates, health clinics' equipment inventory, etc.

[7] For example, proportional stratified sampling induces the population covariate distribution as its target distribution.

# 4 An Interference Gravity Model using Open Data and GIS Tools

### The Gravity Model

Reasoning about the effects of sample choice on interference presupposes a model – implicit or explicit – of how indirect treatment effects are transmitted. In contexts of rural development interventions, physical proximity often characterizes the quantity of exposure to indirect effects. Spatial proximity interference models range from more agnostic approaches where proximity is defined according to whether a subject is located within a certain radius of a treated subject, to approaches that model the rate at which indirect effects decay over distance (*2*).[8]

Density of experimental subjects with respect to total population is as well an expected driver of social interaction and interference. For example, at a similar distance, a couple of subjects in two nearby rural villages composed of a few households are more likely to interact than a couple of subjects in two semi-urban areas composed of thousands of households.

We propose a simple class of gravity models where potential interference between two social clusters – e.g., villages, schools, or health clinics – is driven positively by the density of experimental subjects relative to the general population, and negatively by the distance between them. Gravity models of this type have a long standing history in economics and other social sciences in capturing spatially mediated social interactions.[9] Let $d_{ij}$ denote the distance between clusters $i$ and $j$, then expected interference among them can be modeled by

$$I_{ij} = \frac{f(m_i, m_j)}{g(d_{ij})} = \frac{a\, m_i m_j}{{d_{ij}}^b} \tag{2}$$

where $m_i$ denotes $i$'s density mass, defined as the ratio of the number of experimental subjects in cluster $i$ to its total population. Functions $f$ and $g$ allow for generalizations of the model, which can be informed by substantive knowledge or prior studies in the field.[10] In the traditional formulation of gravity models $f = a\, m_i m_j$ and $g = {d_{ij}}^b$, with $a, b \in \mathbb{R}^+$. In the case of the Prospera experiment, threshold functions were used in $f$ and $g$ so that interference was meaningless for distances above and densities below thresholds set with the assistance of field experts.

The following section discusses the choice of *distance metric* associated to $d_{ij}$, and possibilities enabled by today's available spatial data and open software tools.

[8] See (*2*), Ch. 8, for an introduction and detailed working examples on spatial interference in field experiments.

[9] from trade (see Bergstrand JH. 1985 or Deardorff 1998) and migration flows (see Ravenstein EG. 1889 or Karemera et. al 2000 ), to transportation flows (see Erlander 1990) and epidemics (Xia et. al 2004).

[10] Extensions are also available for cases where researchers might be interested on gravity analysis where the explicit location of individual units within the clusters is relevant, as opposed to the general location of the cluster. This extensions are analogous to the N-Body problem in Newtonian gravitation (Anderson 2011).

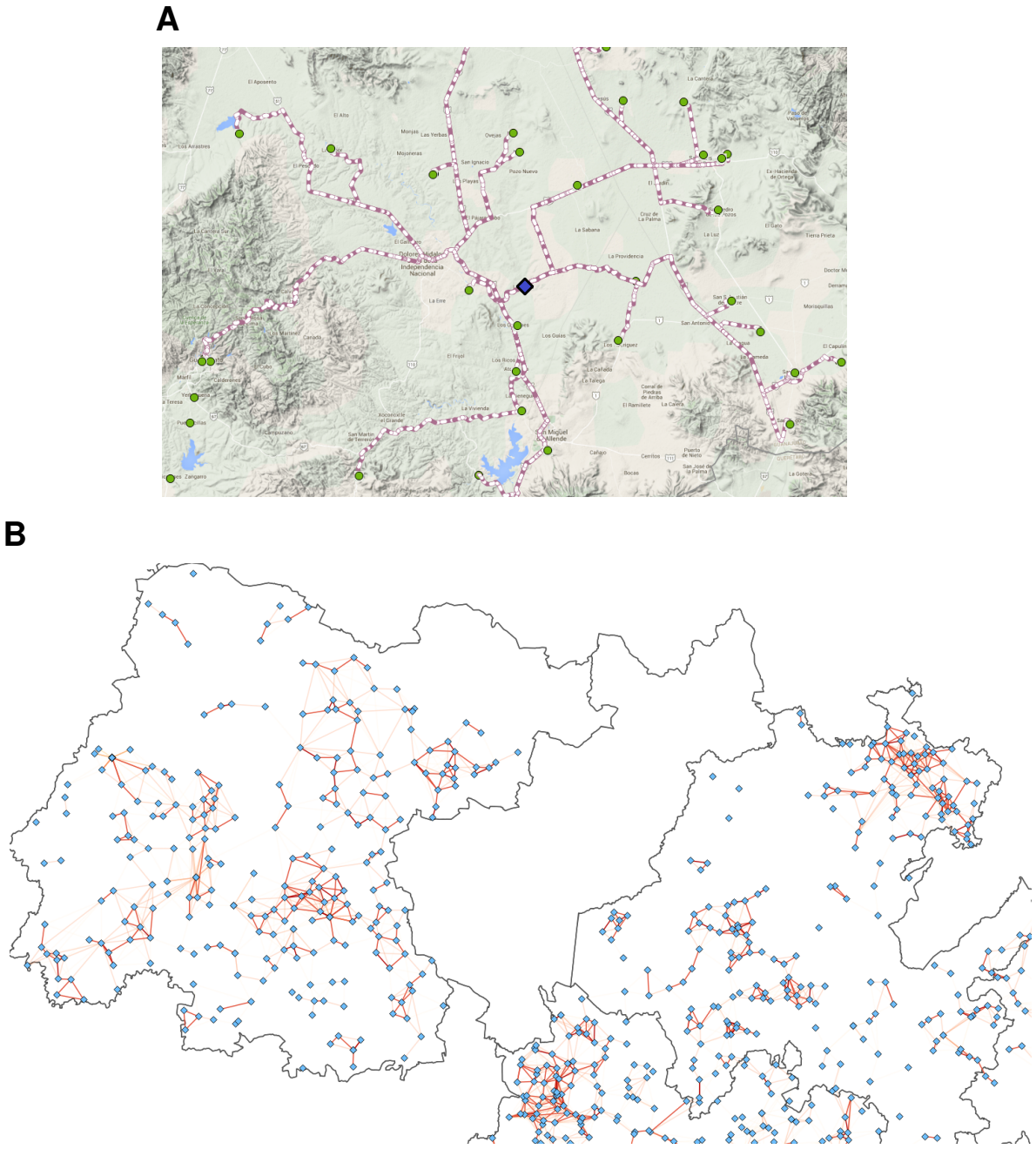


Figure 1: Panel (**A**) visualizes the paths of pregnant women (purple and white lines) from their home villages (green circles) to their assigned Prospera health clinic (blue diamond). Panel (**B**) visualizes the inter-clinic interference network for all Prospera clinics over a region partially covering the states of Guanajuato, Puebla and Mexico. Graded orange lines indicate varying degrees of expected interference according to the interference gravity model.

**Implementation using Open Data and GIS Tools**

Open spatial data and software tools today available to research teams include: spatial data on roads and dirt-roads networks for entire countries, geo-specified census and household survey data at the local polygons level, tools for spatial data querying and visualization, network analytic tools for routing analysis, among others. This section illustrates the use of such tools in the context of modeling interference, yet their application is manifold in the process of field experiment design and analysis.

Proximity analyses require definition of a *distance metric* through which to compute distances $d_{ij}$. In the context of Prospera experiment, we moved from unrealistic straight-line geographic distances to a more meaningful metric of *walk-and-road route* distance.

We obtained spatial data of the Mexican road network through OpenStreetMap repository.[11] This data is also publicly available through the Mexican National Statistics Office (INEGI). We used open sourced tools QGIS and PostGIS for spatial visualization and analysis respectively.[12] Distance between points $A$ and $B$ was defined as the shortest routing distance connecting them, accounting for the possibility of walking route segments outside of the road network. Shortest routes were computed according to differentiated time costs for walking and road commutes, and implemented using PostGIS' Dijkstra shortest path algorithm (Dijkstra 1959). For example, figure 1A visualizes the paths of pregnant women from their villages to an assigned Prospera health clinic, including walking and road segments.

An interference network was computed based on the gravity model and distance metric for all Prospera health clinics across five states. Figure 1B visualizes the network for the states of Guanajuato and Puebla.

# 5 Eliciting the Interference vs. Representativity Tradeoff

This section turns to the joint analysis of interference and representativity. Algorithm 1 defines an interference-minimizing greedy heuristic which progressively removes nodes with highest degree until graph $G(N, E)$ is reduced to the desired sample size $|N| = n_s$.[13] Figure 2 shows results of applying Algorithm 1 to the Prospera experiment sample choice, where mean interference of clinics was reduced in 98.8% as the algorithm pruned the initial pool of 1690 eligible clinics into a 600 clinic sample. The algorithm runs in polynomial time on network size $|N|$, and can easily be run on a desktop computer for network sizes on the order of thousands.

[11] Called 'shape files', or .shp files.

[12] The most common proprietary alternative is Arc Geographic Information System (ArcGIS.), which provides both visualization and analysis capabilities

[13] In graph notation $G(N, E)$, $N$ denotes the set of nodes and $E$ the set of edges.

**Algorithm 1** Interference-minimizing greedy heuristic

*Require:* Interference graph $G(N, E)$ and desired sample size $n_s$

**while** graph size $|N| > n_s$ **do**

Compute degrees (i.e., interference) for each node

$N = N \setminus i'$ , where $i' \in N$ is the node with highest degree.

Sample $S = N$

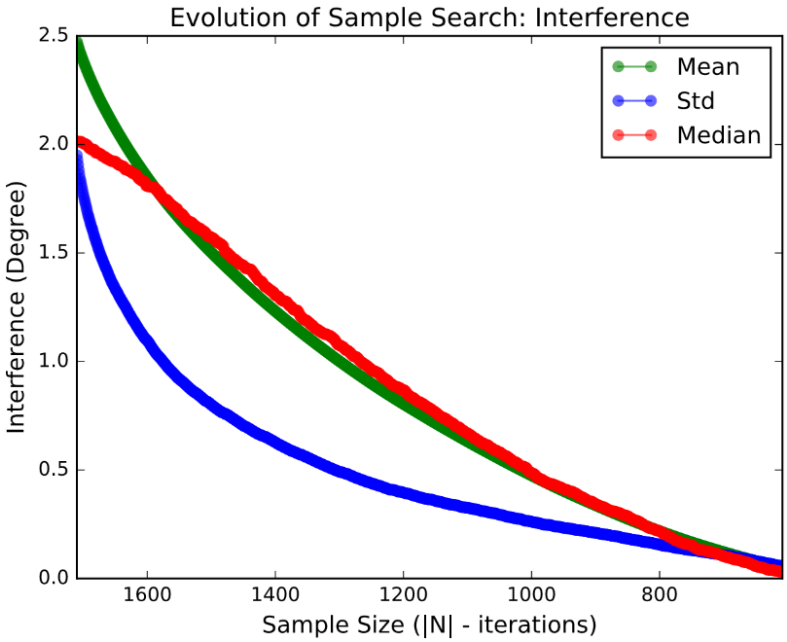


Figure 2: Evolution of the interference-minimizing greedy algorithm. Mean interference among clinics was reduced 98.8% as the algorithm pruned the initial pool of 1710 eligible clinics into a 600 clinic sample (left to right).

However, minimizing interference may compromise external representativity. In the Prospera experiment, we care about a relevant set of covariates including population's social lag, illiteracy, indigenism, type of clinic, average home-to-clinic distance, and pre-treatment outcome variable levels such as of newborn weight and birth diseases. Indeed, Kolmogorov-Smirnov (KS) tests showed that the marginal covariate distributions of the interference-minimizing sample yielded by Algorithm 1 significantly differ from the target population covariate distribution.

Algorithm 2 searches for sample choices that maximize representativity and minimize interference, with the goal of eliciting a set of Pareto non-dominated options, i.e., options that

constitute the tradeoff between objectives.[14]

**Algorithm 2** Bi-objective greedy heuristic minimizing interference and representativity gap

*Require:* Interference graph $G(N, E)$, desired sample size $n_s$, target covariate distribution $J^*(X)$
  $J_N(X) :=$ covariate distribution of nodes in $N$
  **while** graph size $|N| > n_s$ **do**
    Compute degrees (i.e., interference) for each node
    Choose top$_k$ nodes with highest degree
    **for** $i$ in top$_k$ nodes **do**
      Remove $i$ from $N$
      Compute covariate distribution distances between marginals of $J^*(X)$ and $J_{N \setminus i}(X)$
      Add $i$ back to $N$
    $N = N \setminus i'$, where $i' \in$ top$_k$ is the node that most reduced covariate distances
  Sample $S = N$

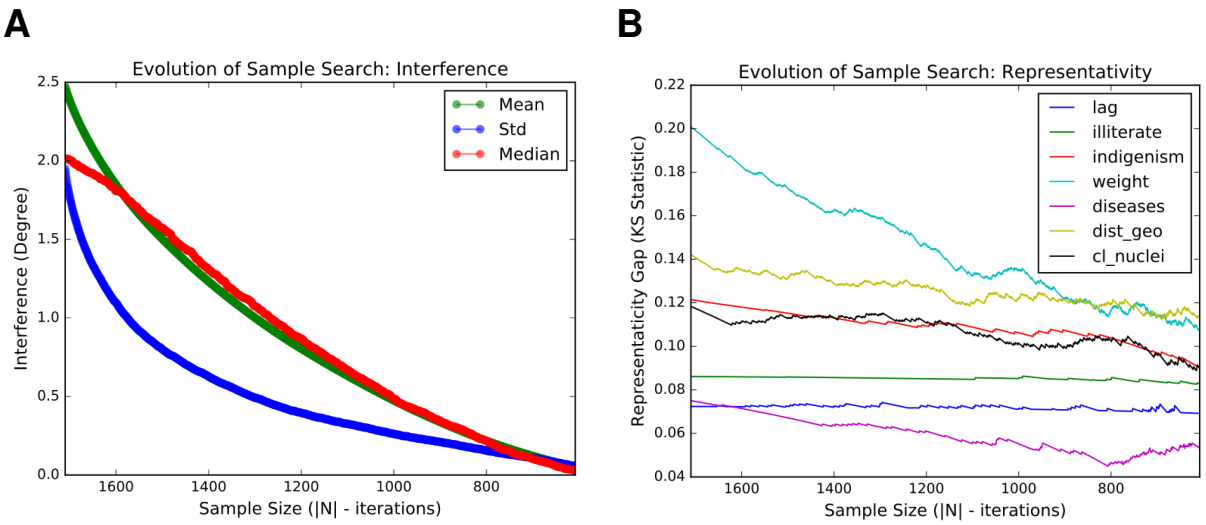


Figure 3: Evolution of Algorithm 2's bi-objective sample search as it prunes the initial pool of 1710 eligible clinics (left to right), using $k = 25$. Panel (**A**): Interference minimization. Panel (**B**): Covariate distance minimization.

Figure 3 shows how the algorithm progressively removes nodes that minimize interference (Figure 3A) and covariate distance between sample and target distributions (Figure 3B), for a

[14]Some practitioners may desire to implement more sophisticated and computationally expensive search heuristics, such as a multi-objective genetic algorithm (MOGA) (see Deb et. al 2002, and the DEAP library for a commonly used open source MOGA implementation).

given $k$. Running Algorithm 2 for different $k$ values provides a set of sample options that trade-off between objectives. Figure 4 shows the set of Pareto non-dominated sample options elicited for the Prospera experiment, where mean interference values range in the [.05, .25] interval, and representativity (average KS statistic) ranges in the $[.05, .12]$ interval. It also compares them against $1000$ random samples. Computational complexity of Algorithm 2 is polynomial in network size $|N|$, and in practice can be run for thousands of nodes on a desktop computer.

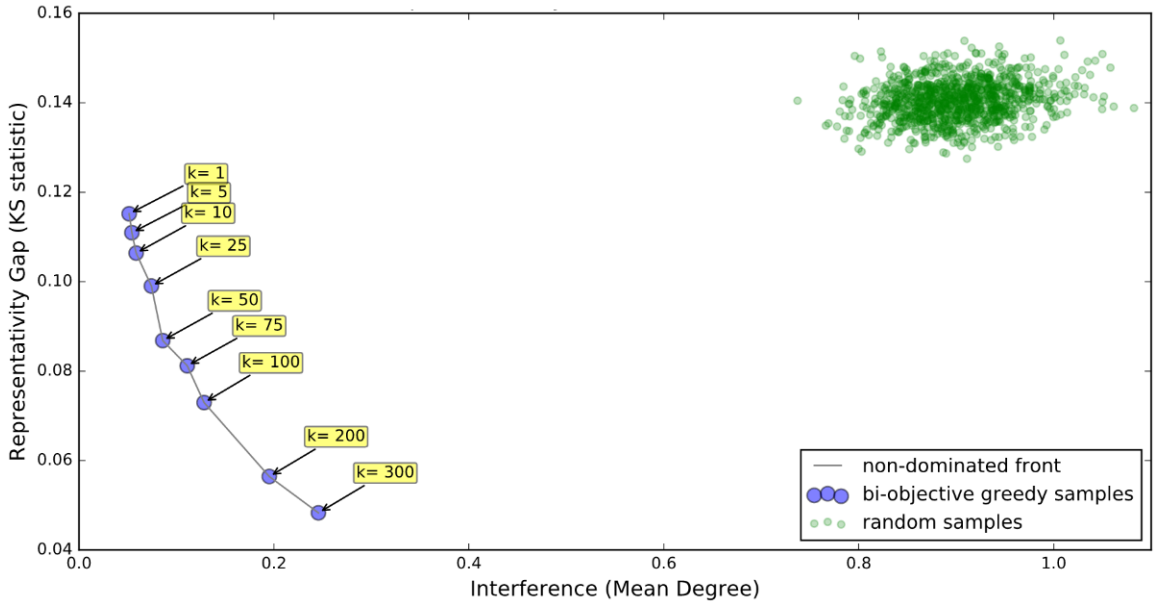


Figure 4: Interference vs. External Representativity tradeoff. The set of Pareto non-dominated sample options for the Prospera experiment, elicited through Algorithm 2 varying parameter $k$, approximate the tradeoff among objectives and dominate performance of $1000$ random samples.

# 6 Sample Choice

Section 5's methodology searches for sample options that optimize both interference and representativity, and approximate the tradeoff among them. Researchers can then perform power calculations and sensitivity analysis over the non-dominated sample options provided, and choose that which best fits the research objectives and context of the study.

For example, in the context of Prospera, and assuming a linear model on direct and indirect effects, we can compute exact p-values for sharp null hypotheses on PATE. This simulations can include the estimation effects that both interference and representativity have on the weighted or conditional estimator. We can then run sensitivity analysis over a grid of values of indirect coefficients, direct coefficients, and within-strata variances, and analyze rejection regions.

# 7 Conclusions and Future Work

In this study we have proposed a methodology for a) modeling the geo-spatial and social interaction factors that drive interference in randomize field experiments; and b) eliciting a set of non-dominated sample options that approximate the Pareto-optimal tradeoff between interference and external representativity as functions of sample choice. We developed and tested the methodology in the context of the Prospera experiment, a large-scale health RCT in rural Mexico, involving more than 5,000 pregnant women and 600 health clinics across five states. Relevant for the practitioner, we showed the methodology is computationally tractable for studies involving networks of thousands of nodes, and can be implemented leveraging novel open sourced geo-spatial data and software tools.

As pointed out by Gerber and Green 2012 (*2*), the study of the existence and nature of indirect or 'spillover' effects is of paramount importance, as it provides relevant insights for conducting subsequent research studies, as well as for the design of policy itself (see (*17*), for a famous example). Moreover, there is an active body of literature advancing methodologies for analyzing indirect effects, such as recent methods for computing exact p-values on non-sharp null hypothesis on network experiments (*18*). The Prospera experiment implementation commenced in January 2016, and will remain active for nearly twelve months. Interesting near-future work will study indirect effects of the Prospera interventions once results on outcome variables are available, and address questions such as: were behavioral spillover effects present in the experiment? What parametrization of the interference gravity model would have best fit observed spillovers? How can Prospera deployment of RapidPro technologies leverage spillover effects?

# References


1. R. M. Bond, C. J. Fariss, J. J. Jones, A. D. Kramer, C. Marlow, J. E. Settle, and J. H. Fowler, "A 61-million-person experiment in social influence and political mobilization," *Nature*, vol. 489, no. 7415, pp. 295–298, 2012.

2. A. Gerber and D. Green, *Field Experiments: Design, Analysis, and Interpretation*. W. W. Norton, 2012.

3. G. Imbens and D. Rubin, *Causal Inference in Statistics, Social, and Biomedical Sciences*. Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction, Cambridge University Press, 2015.

4. B. Sinclair, M. McConnell, and D. P. Green, "Detecting spillover effects: Design and analysis of multilevel experiments," *American Journal of Political Science*, vol. 56, no. 4, pp. 1055–1069, 2012.

5. J. Ugander, B. Karrer, L. Backstrom, and J. Kleinberg, "Graph cluster randomization: Network exposure to multiple universes," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 329–337, ACM, 2013.

6. R. Hayes and S. Bennett, "Simple sample size calculation for cluster-randomized trials.," *International journal of epidemiology*, vol. 28, no. 2, pp. 319–326, 1999.

7. E. Bareinboim and J. Pearl, "Causal inference from big data: Theoretical foundations and the data-fusion problem," tech. rep., DTIC Document, 2015.

8. E. Hartman, R. Grieve, R. Ramsahai, and J. S. Sekhon, "From sate to patt: combining experimental with observational studies to estimate population treatment effects," *JR Stat. Soc. Ser. A Stat. Soc.(forthcoming). doi*, vol. 10, p. 1111, 2015.

9. S. M. Muller, "Causal interaction and external validity: obstacles to the policy relevance of randomized evaluations," *The World Bank Economic Review*, p. lhv027, 2015.

10. T. Y. Chen, T. Tse, and Y.-T. Yu, "Proportional sampling strategy: a compendium and some insights," *Journal of Systems and Software*, vol. 58, no. 1, pp. 65–81, 2001.

11. M. Oportunidades, "Mexicos targeted and conditional transfers: Between oportunidades and rights," *Economic & political weekly*, vol. 46, no. 21, p. 49, 2011.

12. F. Ngabo, J. Nguimfack, F. Nwaigwe, C. Mugeni, D. Muhoza, D. R. Wilson, J. Kalach, R. Gakuba, C. Karema, and A. Binagwaho, "Designing and implementing an innovative sms-based alert system (rapidsms-mch) to monitor pregnancy and reduce maternal and child deaths in rwanda," *Pan African Medical Journal*, vol. 13, no. 31, 2012.

13. A. B. Labrique, L. Vasudevan, E. Kochi, R. Fabricant, and G. Mehl, "mhealth innovations as health system strengthening tools: 12 common applications and a visual framework," *Global Health: Science and Practice*, vol. 1, no. 2, pp. 160–171, 2013.

14. K. Imai, G. King, and E. A. Stuart, "Misunderstandings between experimentalists and observationalists about causal inference," *Journal of the royal statistical society: series A (statistics in society)*, vol. 171, no. 2, pp. 481–502, 2008.

15. S. L. Morgan and C. Winship, *Counterfactuals and causal inference*. Cambridge University Press, 2014.

16. V. J. Hotz, G. W. Imbens, and J. H. Mortimer, "Predicting the efficacy of future training programs using past experiences at other locations," *Journal of Econometrics*, vol. 125, no. 1, pp. 241–270, 2005.

17. A. Banerjee, A. G. Chandrasekhar, E. Duflo, and M. O. Jackson, "The diffusion of microfinance," *Science*, vol. 341, no. 6144, p. 1236498, 2013.

18. S. Athey, D. Eckles, and G. W. Imbens, "Exact p-values for network interference," tech. rep., National Bureau of Economic Research, 2015.