Electronic Journal of Statistics
Vol. 17 (2023) 464–518
ISSN: 1935-7524
https://doi.org/10.1214/23-EJS2111

Design and analysis of bipartite experiments under a linear exposure-response model*

Christopher Harshaw[†1] **Fredrik Sävje**[2]
David Eisenstat[3] **Vahab Mirrokni**[3] **Jean Pouget-Abadie**[3]

[1] *UC Berkeley*

[2] *Yale University*

[3] *Google Research*

Abstract: A bipartite experiment consists of one set of units being assigned treatments and another set of units for which we measure outcomes. The two sets of units are connected by a bipartite graph, governing how the treated units can affect the outcome units. In this paper, we consider estimation of the average total treatment effect in the bipartite experimental framework under a linear exposure-response model. We introduce the Exposure Reweighted Linear (ERL) estimator, and show that the estimator is unbiased, consistent and asymptotically normal, provided that the bipartite graph is sufficiently sparse. To facilitate inference, we introduce an unbiased and consistent estimator of the variance of the ERL point estimator. Finally, we introduce a cluster-based design, Exposure-Design, that uses heuristics to increase the precision of the ERL estimator by realizing a desirable exposure distribution.

MSC2020 subject classifications: Primary 62K99, 62D10; secondary 62G99.
Keywords and phrases: Causal inference, potential outcomes, bipartite experiments, weighting estimators, cluster designs.

Received July 2022.

Contents

arXiv: 2103.06392
*Christopher Harshaw gratefully acknowledges support from an NSF Graduate Research Fellowship (DGE1122492), NSF Grant CCF-1562041, ONR Award N00014-20-1-2335, as well as support from Google as a Summer Intern and a Student Researcher.
[†]Corresponding author e-mail: charshaw@berkeley.edu

1. Introduction

Two-sided marketplaces are rife with interesting but difficult causal questions. What happens to demand if shipping times or fees are reduced? What happens to people's willingness to use ride-hailing apps if more drivers are enrolled in specific cities? What happens to long term user behavior if a hotel booking platform changes its recommendation engine? The causal impact of these changes is hard to quantify, even when using randomized experiments, because marketplace dynamics often violate a central tenet of conventional experimentation: the Stable Unit Treatment Value Assumption, abbreviated SUTVA. This assumption stipulates that the treatment assigned to one unit does not affect any other units. Violations of this assumption is a phenomenon known as interference, which is often present in the case of marketplace experiments and complicates causal analysis.

The bipartite experimental framework offers a useful formalism to study two-sided market experiments and other violations of SUTVA that can happen along the edges of a bipartite graph. This stands in contrast with interference that occurs on graphs where all units are of the same type (e.g. users of a social network). In the bipartite experimental framework, we distinguish two types of units: units that can be subject to an intervention and units whose responses are of interest to the experimenter. We assign treatment to the former and measure the outcomes of the latter. The causal impact of treating one group of units is measured on the other group by tracking the *exposure* to treatment that the latter group receives, informed by the knowledge of the bipartite graph between them. The treatment status of a single unit may affect the measured outcomes of many units and, likewise, a measured outcome may be affected by many treatment units.

As an example, consider a marketplace where buyers compete for limited goods, some of which may be perfectly or partially substitutable. Their demand of these goods form a bipartite graph that potentially can be inferred by the marketplace owner. The owner of the marketplace would like to determine the causal effect of discounting prices on buyers' marketplace behavior through a randomized experiment. Randomly assigning certain buyers to receive a discounted price is often not possible, and might even be prohibited, in which case randomization is only possible at the item-level. At the same time, simply comparing discounted goods with non-discounted goods runs the risk of severe bias. A discounted good may do well against a non-discounted substitutable good, which does not accurately reflect a world where either both or neither are discounted. To address this, the marketplace owner decides to monitor this change at the buyer level. The causal effect can be teased out by tracking both buyers' behavioral changes and exposure to discounted goods.

In addition to the assumptions on the potential outcomes implicitly encoded in the bipartite graph, we consider a setting where the effects of the treatments take a particular form. Similar assumptions are common in the interference literature. One such assumption is the existence of an exposure mapping, which posits that outcomes are some simple function of the treatment assignments of neighboring units in the bipartite graph [59, 5]. In this work, we study estimation of an all-or-nothing treatment effect in the bipartite experimental framework under a linear exposure-response model, where exposures are linear functions of assignments and responses are linear functions of the exposures.

The main contributions of this work are summarized as follows:

- We describe the Exposure-Reweighted Linear (ERL) estimator, an unbiased linear estimator of the average total treatment effect under the linear exposure-response model. We show that the ERL estimator is consistent and asymptotically normal, provided the graph is sufficiently sparse.
- We describe a variance estimator, which can be used to construct confidence intervals via a normal approximation. We show that under mild conditions on the exposure distribution, the variance estimator is unbiased and consistent. Additionally, we prove asymptotic validity of normal-based

confidence intervals using the variance estimator.
- We describe the behavior of the ERL estimator when the linear exposure-response model does not hold, showing that it still estimates an interpretable and policy relevant causal quantity.
- We describe Exposure-Design, a cluster-based design which aims to increase the precision of the ERL estimator. The design achieves this by increasing the variance of individual exposures while decreasing the covariance between exposures of different units. This improves precision in several settings of interest.

2. Related works

Within the wide-ranging causal inference literature, our work falls squarely within the subset relying on the potential outcomes framework [46, 35]. The design and analysis of randomized experiments in the presence of interference has garnered much attention, spanning vaccination trials [57], agricultural studies [37], voter-mobilization field experiments [55], and viral marketing campaigns [2, 20]. It is beyond the scope of the current paper to extensively review the literature on causal inference under interference. Instead, we direct readers to the review article by Halloran and Hudgens [28].

Our work is primarily motivated by marketplace experiments. Evidence of interference in marketplaces has been noted across industries for various experimental designs [27]. Reiley [50], Einav et al. [21] and Holtz et al. [32] study the interference bias that results from supply-side randomization, while Blake and Coey [11] and Fradkin [24] consider this problem in the context of demand/user-side randomization. Basse et al. [10] and Liu et al. [39] compare supply-side randomization to two-sided randomization as well as to budget-split designs, showing bias can be reduced in the context of certain ad auction experiments. More recently, Johari et al. [36] characterize which randomization scheme (supply-side, demand-side, or two-sided) leads to reduced bias as a function of market balance.

We consider a slightly different experimental setting, introduced by Zigler and Papadogeorgou [61], characterized by random assignment of treatment on one side of the bipartite graph (demand- or supply-side), while outcomes are measured on the other side. The advantage of this framework is that the bipartite graph defines an exposure function (similar to Aronow and Samii [5]), which is assumed to solely determine an unit's outcome, making the estimation problem more tractable. Zigler and Papadogeorgou [61] study causal estimands which are more closely related to direct effects rather than the all-or-nothing treatment effect considered here. A drawback of their work is that the analysis of their estimators requires that the bipartite graph be the union of many small connected components.

To make estimation and inference more tractable when the graph is complex, methodologists opt for stronger structural assumptions on the outcomes. An exposure-response assumption similar to the one we use here is adopted by Pouget-Abadie et al. [49], who introduce a cluster-based design for general

bipartite graphs, consider a similar estimand, and are also motivated by marketplace experiments. Later, Doudchenko et al. [18] proposed a class of generalized propensity score estimators for this framework, which are unbiased for both experimental and observational settings under standard assumptions and a similar exposure-response assumption.

To the best of our knowledge, our work is the first to propose methods for provably valid inference (e.g., confidence intervals) in the bipartite settings and to jointly consider estimators and designs which improve overall precision of treatment effect estimators. While the cluster design of Pouget-Abadie et al. [49] is based on the intuition of achieving a large spread of exposures, it disregards the correlation of exposures and is not directly tied to the performance of an estimator. Additionally, while the estimators proposed by Doudchenko et al. [18] are unbiased, they are based on a different approach which requires fitting a generalized propensity score function. Neither of these papers present methods for valid inference.

3. Experimental setting

In the bipartite experimental framework, the units that receive treatment are distinct from the units on which the outcomes are measured. For example, Zigler and Papadogeorgou [61] apply the framework to analyze how interventions on power plants' pollution affect the hospitalization rates among nearby hospitals. We discuss the general bipartite framework in Section 3.1 and the linear exposure response assumption in Section 3.2.

3.1. Bipartite experiments

In the bipartite experiment setting, there are two groups of units: the *diversion units*, to which treatment is applied, and the *outcome units*, on which outcomes are measured. We denote the set of m diversion units by V_d and the set of n outcome units by V_o.

Each of the m diversion units receives a (random) binary treatment $z_i \in \{0,1\}$, and we collect these treatments into a treatment vector, $\boldsymbol{z} = (z_1, z_2, \ldots z_m) \in \{0,1\}^m$. The distribution over the random treatment vectors is called the *design* of the experiment and it is chosen by the experimenter. Each of the outcome units $i \in V_o$ is associated with a potential outcome function $Y_i(\boldsymbol{z})$, which maps the treatment assignments to the observed value, which is a real number. In the bipartite setting, we assume that each potential outcome function depends only on the treatment of a neighborhood set of diversion units. More formally, there exists a *neighborhood mapping* $\mathcal{N} : V_o \to 2^{V_d}$ such that for all outcome units $i \in V_o$,

$$Y_i(\boldsymbol{z}) = Y_i(\boldsymbol{z}') \quad \text{if } z_j = z'_j \text{ for all } j \in \mathcal{N}(i) \ .$$

Throughout the paper, we assume that the neighborhood mapping is known and correctly specified, so that the above condition holds. We recover the standard

Stable Unit Treatment Value Assumption (SUTVA) when the diversion units are identified with the outcome units and the neighborhood mapping is the identity function.

The number of potential outcomes for each outcome unit grows exponentially in the size of its neighborhood. Zigler and Papadogeorgou [61] avoid this issue by assuming that the bipartite structure is the union of many small connected components. While this is a useful assumption in some settings, it typically does not hold in a marketplace setting where we know that more varied interactions occur: buyers may interact with a variety of products. Without further restrictions on the structure of the neighborhoods or the potential outcome functions, inference of any causal estimand is impossible [9, 54]. One example where this is impossible is when the neighborhood of each outcome unit is all diversion units. For this reason, we restrict our attention to settings where a stronger assumption on the potential outcomes is reasonable.

3.2. Linear exposure-response model

In order to admit tractable inference of causal estimands, we consider a linear exposure-response model, which consists of two underlying assumptions: a linear exposure assumption and a linear response assumption, which we state formally below.

In the linear exposure-response model, we suppose that there is a weighted bipartite graph between diversion units and outcomes units, where the edges have non-negative weights $w_{i,j} \geq 0$, which we arrange into an n-by-m incidence matrix $\boldsymbol{W}$. An edge $w_{i,j}$ represents the influence of diversion unit j on the outcome units i. We say that outcome unit i and diversion unit j are *incident* if the weight $w_{i,j}$ is positive. The degree of a diversion unit is the number of outcome units it is incident to, and the largest degree among all diversion units is denoted d_d. The degree of an outcome unit is defined similarly and the largest degree among all outcome units is denoted d_o. For simplicity, we assume that each outcome unit has at least one incident diversion unit, $d_o \geq 1$, and the weights incident to an outcome unit are normalized to sum to one. That is, the rows of the incidence matrix $\boldsymbol{W}$ sum to one. We also assume that this weighted bipartite graph is known to the experimenter. In many market experiments, the experimenter may construct an approximation of this graph from historical data.

The *linear exposure assumption* is that the treatment assignments influences the potential outcomes only through a linear combination. This imposes more structure on the relationship between the diversion and outcome units than under arbitrary interference. More formally, for each outcome unit $i \in V_o$, the *exposure* of outcome unit i is

$$x_i(\boldsymbol{z}) = \sum_{j \in V_d} w_{i,j} z_j \ ,$$

and for all pairs of assignment vectors $\boldsymbol{z}$ and $\boldsymbol{z}'$ with $x_i(\boldsymbol{z}) = x_i(\boldsymbol{z}')$, we have that $Y_i(\boldsymbol{z}) = Y_i(\boldsymbol{z}')$. This implies that the neighborhood mapping is such that

$\mathcal{N}(i) = \{j : w_{i,j} > 0\}$.

We define the exposure vector $\boldsymbol{x}(\boldsymbol{z}) = (x_1(\boldsymbol{z}), x_2(\boldsymbol{z}) \dots x_n(\boldsymbol{z}))$ to be the arrangement of the n individual exposures into a vector. Because the exposure is a function of treatment, the experimental design determines the exposure distribution. This linear exposure assumption is a generalization of the partial and stratified interference assumptions discussed by Hudgens and Halloran [34]. When the treatment assignment vector $\boldsymbol{z}$ is clear from context, we write simply x_i and $\boldsymbol{x}$ for the ith exposure and the exposure vector, respectively. Similarly, we write Y_i for the outcomes. Using matrix-vector notation, we may write the exposure vector as $\boldsymbol{x}(\boldsymbol{z}) = \boldsymbol{W}\boldsymbol{z}$. Due to the normalization of the weights and the binary values of the treatment assignment, each exposure takes values in the range $[0, 1]$.

The *linear response assumption* is that for each outcome unit, the potential outcome is a linear function of its exposure. That is, for each outcome unit $i \in V_o$, there exists parameters α_i and β_i such that

$$Y_i(\boldsymbol{z}) = \alpha_i + \beta_i x_i(\boldsymbol{z}) \ .$$

We refer to α_i as the unit-specific intercept and β_i as the unit-specific slope. Note that the linear function does not need to be the same between units. These coefficients are unknown to the experimenter, and the experimenter only observes the outcome $Y_i(\boldsymbol{z})$, along with the sampled assignment vector $\boldsymbol{z}$ and the resulting exposure vector $\boldsymbol{x}$.

We refer to the *linear exposure-response model* as the combination of the linear exposure assumption and the linear response assumption. The linear exposure-response model imposes restrictions on the potential outcomes, but allows for more complex structure in the bipartite graph than previous work. This trade-off is preferable in settings such as marketplace experiments, where we know that a complex bipartite structure exists and we are more comfortable with making simplifying assumptions about potential outcomes. For further discussion on empirical and theoretical evidence for complex structure in marketplace experiments, we refer the reader to Blake and Coey [11], Fradkin [25], and Johari et al. [36].

Structural assumptions on the outcomes similar to the linear exposure-response assumption presented here are commonly made throughout the interference literature. The *linear-in-means* (LIM) model posits that a unit's response is a linear function of their own treatment, and the mean of the treatments of their group [42]. The LIM model has been extended in various ways in the context of partial interference [7, 47] and social network experiments [12, 59]. Chin [16] investigates the use of machine learning estimators for the total average treatment effect under a variation of the LIM when the terms in the linear model of arbitrary functions of treatment. Basse et al. [10] study model-assisted estimators and designs under the "normal sum-model" which is similar to the linear exposure-response considered here, but with a normal noise term. We remark that the bipartite setting with the linear exposure-response assumption recovers the standard SUTVA setting when diversion units are identified with the outcome units and the weight matrix is the identity.

The linear exposure-response assumption in this work bears some resemblance to a semiparametric partially linear model with heteroskedastic errors [51]. A key assumption in those superpopulation models is that the outcomes of all units have a shared linear term (which is typically the estimand) together with an additive heteroskedastic noise term whose distribution can depend on unit-level information such as covariates. We differ from this approach in at several ways. First, we use a design-based perspective, meaning that randomization of treatment assignments forms the sole basis for inference, i.e. units are not assumed to be drawn i.i.d. from a larger distribution. Second, the linear exposure-response assumption considered in this paper allows for units to have distinct linear terms β_i as well as additive terms α_i, neither of which is assumed to be related to any covariates.

From one perspective, the linear exposure-response model is a strong assumption. It requires that the response for each unit is exactly a linear function in the exposure. This rules out, for example, that different diversion units have different impacts on a single outcome unit, other than what can be captured by the weights in the linear combination in the exposures. But from another perspective, the model is completely unrestrictive: it does not limit the heterogeneity between units at all. That is, knowing the response function for one unit tells us nothing about the response function of other units. While there are few settings in which the linear exposure-response model will hold exactly, it will sometimes be a useful approximation given its unrestrictiveness with respect to heterogeneity. In Section 6, we analyze the behavior of the ERL estimator under a general non-linear response assumption, finding that it estimates a best linear approximation to the average response. However, we leave it to future work to more precisely characterize the behavior of estimator under general responses and we assume the linear exposure-response model holds exactly throughout the paper, unless otherwise stated.

3.3. Causal estimand

We are interested in understanding the contrast between two possible worlds: one where all diversion units receive treatment and one where they all receive control. For an individual outcome unit, this contrast is captured by the individual treatment effect, $\tau_i = Y_i(\mathbf{1}) - Y_i(\mathbf{0})$ for $i \in V_o$. Just as in the typical SUTVA setting, we cannot hope to estimate the individual treatment effects well because only one potential outcome is observed for any one unit. In light of this, we opt to estimate an aggregated causal quantity. In this paper, we are interested in the Average Total Treatment Effect (ATTE), which is the average contrast between the scenario when all diversion units receive treatment and when all diversion units receive control. More precisely, ATTE is defined as

$$\tau = \frac{1}{n}\sum_{i=1}^{n}\tau_i = \frac{1}{n}\sum_{i=1}^{n}\Big[Y_i(\mathbf{1}) - Y_i(\mathbf{0})\Big]$$

Under the linear exposure-response assumption, the ATTE is proportional to

the average of the slope terms, as shown in the following proposition.

Proposition 3.1. *Under the linear exposure-response assumption, the ATTE is* $\tau = \frac{1}{n}\sum_{i=1}^{n}\beta_i$.

Proof. The individual treatment effect of outcome unit i is equal to its slope, as

$$\tau_i = Y_i(\mathbf{1}) - Y_i(\mathbf{0}) = \left[\beta_i x_i(\mathbf{1}) + \alpha_i\right] - \left[\beta_i x_i(\mathbf{0}) + \alpha_i\right] = \beta_i \ ,$$

where we have used that $x_i(\mathbf{1}) = 1$ and $x_i(\mathbf{0}) = 0$. The result follows by taking the average of the individual treatment effects. □

There are two main challenges in estimating the ATTE in this setting. First, we want to estimate the average of the slopes of many different linear response functions, but we observe only one point from each of the distinct linear response functions. Although stated in somewhat unfamiliar terms, this is the fundamental problem of causal inference [31]. The second challenge is that of constructing a treatment design which realizes a desirable exposure distribution. This is a difficult task when the bipartite weight matrix has non-trivial overlapping structures. In the remainder of the paper, we focus on addressing these two challenges by developing an estimator and a class of designs which together accurately estimate the ATTE.

3.4. Cluster designs

Some of the analysis in this paper assumes that the treatment is assigned according to a *independent cluster design*, where the diversion units are grouped into clusters and treatment is assigned to an entire cluster. More formally, we say that a partition $C_1, C_2, \ldots, C_\ell$ of the diversion units is a *clustering*, which we denote as $\mathcal{C} = \{C_1, C_2, \ldots, C_\ell\}$. That is, the clusters $C_1, C_2, \ldots, C_\ell$ are disjoint and their union is set of diversion units V_d. Given a clustering $\mathcal{C}$, a treatment assignment from the corresponding *independent cluster design* is drawn in the following way: independently for each cluster, we assign all diversion within a cluster to have either treatment $z_i = 1$ with probability p and treatment $z_i = 0$ with probability $1 - p$. For notational simplicity, we consider the treatment probability p to be fixed for all clusters, but our results extend to settings where each cluster has its own treatment probability. Note that the class of independent cluster designs is completely specified by $\mathcal{C}$ and p.

4. The exposure reweighted linear estimator

The Exposure Reweighted Linear (ERL) estimator, which is an estimator of the ATTE under the linear exposure-response assumption, is defined as

$$\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n} Y_i\left(\frac{x_i - \mathbb{E}[x_i]}{\mathrm{Var}(x_i)}\right) \ . \tag{1}$$

The ERL estimator requires knowledge of the mean and variance of each of the marginal exposure distributions under the treatment design. For several commonly used designs such as Bernoulli and independent cluster designs, these exposure characteristics can be computed directly. For arbitrary designs, the expectation and variance of the exposures may need to be estimated using samples drawn from the treatment design. We assume here that these exposure characteristics are known exactly. The ERL estimator can be used under any treatment design and not just the cluster-based treatment design we propose in Section 7.

The ERL estimator belongs to the class of linear estimators, as it is a (random) linear function of the observed outcomes. It shares similarities with the style of Horvitz–Thompson estimators, which weight an outcome by the probability of observing that outcome [45, 33]. On the other hand, the ERL estimator weights an outcome by the normalized distance of the exposure from its mean. When there are many possible values of exposures, such as under the linear exposure-response model, the type of weighting done by the Horvitz–Thompson estimator would lead to excessively large variance.

4.1. Unbiasedness and consistency of the ERL *estimator*

In this section, we analyze the behavior of the ERL estimator as a point estimator of the average total treatment effect (ATTE). First, we show that the ERL estimator is unbiased. Then we show consistency and asymptotic normality of the ERL estimator, provided that the bipartite graph is not too dense. Theorem 4.1 ensures that under mild conditions on the treatment design, there is no systematic bias in the ERL estimator.

Theorem 4.1 (Unbiasedness)**.** *Suppose the design is such that each exposure has positive variance,* $\mathrm{Var}(x_i) > 0$*. Under the linear response assumption, the* ERL *estimator is unbiased for the ATTE:* $\mathbb{E}[\hat{\tau}] = \tau$*.*

Next, we analyze the asymptotic behavior of the ERL estimator. In the asymptotic analysis, we suppose that there is a sequence of bipartite experiments, in which the number of outcome and diversion units are growing to infinity. Strictly speaking, all quantities of the experiment such as the bipartite graph, the outcomes, the treatment design, and so on, should be indexed by an integer indicating which experiment in the sequence the variable refers to, but we drop these subscripts for notational clarity.

We make two additional assumptions about the bipartite experiments in this asymptotic sequence. The first is that the potential outcomes are bounded. The second is that the design has limited dependence between treatment assignments.

Assumption 1 (Bounded Potential Outcomes)**.** *The potential outcomes are bounded in absolute value* $|Y_i(\boldsymbol{z})| \leq M$*, where* M *is a constant.*

Assumption 2 (Design Conditions)**.** *The treatments assignments are distributed according to an independent cluster design, where the probability of treatment*

p is bounded away from 0 *and* 1 *by a constant in the asymptotic sequence. Additionally, the sizes of clusters are bounded by k, which is a constant in the asymptotic sequence.*

Assumption 2 rules out certain classes of treatment designs, such as complete randomization (i.e. group balanced designs). While it may be possible to obtain similar asymptotic results under such designs, we here limit our consideration to those satisfying Assumption 2. Under these assumptions, we prove that ERL is consistent when the bipartite graph is not too dense.

Theorem 4.2 (Consistency)**.** *Under Assumptions 1 and 2, the mean squared error of the* ERL *estimator is bounded as* $\mathbb{E}[(\hat{\tau}-\tau)^2] = \mathcal{O}\big(d_d d_o^3/n\big)$*. Thus, the estimator is consistent if* $d_d d_o^3 = o(n)$*.*

Theorem 4.2 shows that the convergence rate of the ERL estimator is at least $\sqrt{d_d d_o^3/n}$, where d_d and d_o are the maximum degrees of the diversion and outcome units, respectively. The main technical assumption that we require for consistency is that $d_d d_o^3 = o(n)$ in the asymptotic sequence. Informally, this condition ensures that the bipartite graph is not too dense as it grows. Indeed, we expect the convergence of any estimator to worsen as the graph becomes more complex and dense. While consistency may hold under weaker conditions for particular designs, an assumption on the graph density must be made: in the complete bipartite graph where all outcome units receive the same exposure, consistent estimation is impossible.

We now discuss a specific class of instances where this condition $d_d d_o^3 = o(n)$ holds. Suppose that each diversion unit has fixed degree d_d, which is a constant with respect to m and n. The average degree of an outcome unit is then $\bar{d}_o = d_d m/n$. Assuming that the maximum outcome degree d_o is within a constant factor of the average, this yields that the term $d_o = \mathcal{O}(m/n)$. Using that the diversion degrees are constant, we get that $d_d d_o^3 = o(n)$ if $m = o(n^{4/3})$. Thus, in graphs with constant diversion degrees where the edges are roughly evenly distributed between outcome units, the premise of Theorem 4.2 holds when m grows at a rate slower than $n^{4/3}$. Observe that the growth condition $m = o(n^{4/3})$ arises from the fact that all diversion units have the same degree. Generally speaking, the underlying bipartite graph need not have constant degree across all diversion units and so the condition $d_d d_o^3 = o(n)$ may still hold for faster growth rates on m, in certain instances.

4.2. Asymptotic normality of the ERL *estimator*

We next characterize the limiting distribution of the estimator. In particular, we show that the sampling distribution of the ERL estimator converges to a normal distribution as the size of the bipartite experiment grows, provided that the graph remains sparse. This result is derived under the same asymptotic regime as above. In order to prove the central limit theorem, we require an additional assumption on the asymptotic sequence of bipartite experiments. Namely, we

require that the variance of the ERL estimator decreases no faster than the parametric rate.

Assumption 3. *The normalized variance of the* ERL *estimator* $n \cdot \mathrm{Var}(\hat{\tau})$ *is bounded away from zero asymptotically.*

Assumption 3 rules out settings in which we can estimate the ATTE at an unusually fast rate. It is theoretically possible to estimate ATTE at a faster than parametric rate, but these settings are not practically relevant. Assumption 3 rules out primarily three scenarios. The first is when the magnitude of the potential outcomes approaches zero in the asymptotic sequence. This requires that almost all potential outcomes approach zero; the magnitude of the potential outcomes are generally non-zero even when their average is zero. The second scenario is when the design almost perfectly pinpoints the potential outcomes. This can be formalized as the variance of each individual term of the estimator diminishes asymptotically, i.e. $\mathrm{Var}(\hat{\tau}_i) \to 0$, where $\hat{\tau}_i = Y_i(\boldsymbol{z})(x_i - \mathbb{E}[x_i])/\mathrm{Var}(x_i)$. The third scenario is when the ERL estimator converges at a parametric rate, but the asymptotic variance happens to be exactly zero. All of these scenarios are knife-edge cases that we have good reason to believe would not materialize in practice. Even if they do, the estimator would still be unbiased and consistent, but its asymptotic distribution might not be normal.

We are now ready to present a central limit theorem.

Theorem 4.3 (Asymptotic Normality)**.** *Under Assumptions 1, 2, and 3, and supposing that* $d_d^4 d_o^{10} = o(n)$*, the* ERL *estimator is asymptotically normal:*

$$\frac{\hat{\tau} - \tau}{\sqrt{\mathrm{Var}(\hat{\tau})}} \xrightarrow{d} \mathcal{N}(0,1) \ .$$

The proof relies on Stein's method for bounding distances between distributions (see, e.g. Ross [52]). We use Stein's method because standard techniques for establishing central limit theorems which rely heavily on independence are not applicable in the bipartite experimental framework where exposures are necessarily correlated. Stein's method has been used to investigate limiting behavior of other estimators in the recent interference literature [5, 15, 48].

The assumptions on the asymptotic growth of the bipartite graph may be interpreted similarly as those appearing in Theorem 4.2. Namely, they prevent the bipartite graph from becoming too dense. The growth assumptions required for asymptotic normality (Theorem 4.3) are stronger than those required required for consistency (Theorem 4.2). The conditions in Theorem 4.3 are only sufficient, and we conjecture that they are not necessary for asymptotic normality. One aspect the growth conditions ensure is that the variance of the exposures does not converge to zero at a too fast rate. If we ensure positivity of the exposures through some other means, then the growth conditions can be weakened. For example, assuming that $\mathrm{Var}(x_i) \geq c > 0$, the growth conditions are relaxed to $d_d^4 d_o^4 = o(n)$. However, weakening the growth conditions beyond this would require a different analysis, either by a more careful application of Stein's method or by different means all together.

Assumption 2 allows for a broad class of designs. For example, unit-level Bernoulli randomization satisfies this condition, but this design does not consider the structure of the bipartite graph and will generally perform poorly. To derive analytical results for this broad class of designs, the growth conditions on the bipartite graph are quite restrictive, and they may be too restrictive in certain settings where more dense interaction patterns occur. If one restricts focus to a smaller class of designs, these growth conditions could potentially be weakened. The key implication of Assumption 2 together with the growth conditions is that the variance of the exposures is large and the correlation between most pairs of exposures is small. Heuristically, these conditions on the exposure distribution are the main aspects required for consistency and normality. We describe a design in Section 7 that directly targets the exposure distribution to satisfy these conditions, and it will therefore be better behaved than many of the designs allowed by Assumption 2.

5. Variance estimation and confidence intervals

In this section, we present methods for constructing confidence intervals for the ATTE in the bipartite setting under the linear exposure-response assumption. If we knew the variance of the ERL estimator, we could use Theorem 4.3 directly to construct asymptotically valid confidence intervals. However, because the variance of the ERL estimator depends on the unobserved potential outcomes, we must construct an estimator of the variance.

In the finite population experimental settings with binary treatments, unbiased variance estimation is not possible without strong assumptions on the heterogeneity between units [35]. In light of this negative result, experimenters tend to favor conservative variance estimators that lead to valid but overly wide confidence intervals. In contrast to the typical experimental settings, we show that unbiased variance estimation is possible in the bipartite setting under the linear response assumption when the exposures take many (i.e. more than two) values.

5.1. Derivation of the variance estimator

To the best of our knowledge, this approach to constructing a variance estimator is new. We will therefore describe its derivation somewhat carefully here. Our approach begins by first decomposing the ERL estimator into a weighted average of individual effect estimators, and then further decomposing the variance of the ERL estimator as the average of covariances of these individual effect estimators.

Define $\hat{\tau}_i \triangleq Y_i(x_i - \mathbb{E}[x_i])/\operatorname{Var}(x_i)$ to be the individual terms in the ERL estimator. We may interpret $\hat{\tau}_i$ as an unbiased, but very imprecise, estimator of the individual treatment effect τ_i. The ERL estimator can now be written as the average of these quantities: $\hat{\tau} = (1/n)\sum_{i=1}^{n} \hat{\tau}_i$. The variance of the ERL

estimator can therefore be written as

$$\mathrm{Var}(\hat{\tau}) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\tau}_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{Cov}(\hat{\tau}_i, \hat{\tau}_j) \ .$$

Our approach to constructing a variance estimator will be to construct simple weighting estimators for each of the $\mathrm{Cov}(\hat{\tau}_i, \hat{\tau}_j)$ terms. These weighting estimators will be expressed as $\widehat{C}_{i,j} = Y_i Y_j R_{i,j}(x_i, x_j)$, where $Y_i Y_j$ is the product of observed outcomes and $R_{i,j}(x_i, x_j)$ is a weighting function which takes the observed exposures as inputs, such that our variance estimator will be the simple average of these weighted products of outcomes:

$$\widehat{\mathrm{Var}}(\hat{\tau}) \triangleq \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\widehat{C}_{i,j} \triangleq \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}Y_i Y_j R_{i,j}(x_i, x_j) \ .$$

The goal here is to make each individual estimator $\widehat{C}_{i,j}$ unbiased for each individual covariance $\mathrm{Cov}(\hat{\tau}_i, \hat{\tau}_j)$, so that the entire variance estimator $\widehat{\mathrm{Var}}(\hat{\tau})$ will be unbiased. Moreover, if the individual estimators $\widehat{C}_{i,j}$ are sufficiently uncorrelated, then the overall variance estimator will achieve high precision.

We consider three different types of terms in the double sum. The first type is when the covariance between unit-level estimators $\mathrm{Cov}(\hat{\tau}_i, \hat{\tau}_j)$ is known to be zero. In particular, the linear response assumption implies that $\mathrm{Cov}(\hat{\tau}_i, \hat{\tau}_j) = 0$ when $\mathrm{Cov}(x_i, x_j) = 0$. Therefore, we set the weighting function $R_{i,j}(x_i, x_j)$ to be exactly zero whenever $\mathrm{Cov}(x_i, x_j) = 0$, which means that $\widehat{C}_{i,j} = 0$.

The second type is the non-zero, off-diagonal terms: $\mathrm{Cov}(\hat{\tau}_i, \hat{\tau}_j) \neq 0$ and $i \neq j$. For these pairs, we define the overall weighting function $R_{i,j}(x_i, x_j)$ to be

$$R_{i,j}(x_i, x_j) = Q_{i,j}(x_i, x_j) - S_{i,j}(x_i, x_j) \ ,$$

where $Q_{i,j}(x_i, x_j)$ and $S_{i,j}(x_i, x_j)$ are defined as

$$\begin{aligned} Q_{i,j}(x_i, x_j) &= \frac{x_i - \mathbb{E}[x_i]}{\mathrm{Var}(x_i)} \cdot \frac{x_j - \mathbb{E}[x_j]}{\mathrm{Var}(x_j)}, \\ S_{i,j}(x_i, x_j) &= a_{i,j}(x_i x_j - \mathbb{E}[x_i x_j]) + b_{i,j}(x_i - \mathbb{E}[x_i]) + c_{i,j}(x_j - \mathbb{E}[x_j]) \ , \end{aligned}$$

and the coefficients $a_{i,j}$, $b_{i,j}$, $c_{i,j}$ are obtained as solutions to the following system of linear equations:

$$\Sigma_{i,j}\begin{bmatrix} a_{i,j} \\ b_{i,j} \\ c_{i,j} \end{bmatrix} = \begin{bmatrix} \mathrm{Var}(x_i x_j) & \mathrm{Cov}(x_i, x_i x_j) & \mathrm{Cov}(x_j, x_i x_j) \\ \mathrm{Cov}(x_i x_j, x_i) & \mathrm{Var}(x_i) & \mathrm{Cov}(x_j, x_i) \\ \mathrm{Cov}(x_i x_j, x_j) & \mathrm{Cov}(x_i, x_j) & \mathrm{Var}(x_j) \end{bmatrix}\begin{bmatrix} a_{i,j} \\ b_{i,j} \\ c_{i,j} \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \ .$$

The two weighting functions have been constructed such that $Y_i Y_j Q_{i,j}(x_i, x_j)$ and $Y_i Y_j S_{i,j}(x_i, x_j)$ are unbiased estimators of $\mathbb{E}[\hat{\tau}_i \hat{\tau}_j]$ and $\mathbb{E}[\hat{\tau}_i]\mathbb{E}[\hat{\tau}_j]$, respectively. Because $\mathrm{Cov}(\hat{\tau}_i, \hat{\tau}_j) = \mathbb{E}[\hat{\tau}_i \hat{\tau}_j] - \mathbb{E}[\hat{\tau}_i]\mathbb{E}[\hat{\tau}_j]$, this means that the overall

estimator $\widehat{C}_{i,j} \triangleq Y_i Y_j R_{i,j}(x_i, x_j) = Y_i Y_j Q_{i,j}(x_i, x_j) - Y_i Y_j S_{i,j}(x_i, x_j)$ is an unbiased estimator of the covariance $\mathrm{Cov}(\hat{\tau}_i, \hat{\tau}_j)$. We refer the reader to Appendix B for more details.

The third case is the diagonal terms: $i = j$. In principle, we could use the same functions $Q_{i,j}(x_i, x_j)$ and $S_{i,j}(x_i, x_j)$ as for the off-diagonal terms, but the system of equations is underdetermined, so the coefficients $a_{i,j}$, $b_{i,j}$, and $c_{i,j}$ are not uniquely determined. To address this, we set $c_{i,i} = 0$ when $i = j$, and obtain the coefficients $a_{i,i}$ and $b_{i,i}$ as solutions to the following system of linear equations:

$$\Sigma_{i,i} \begin{bmatrix} a_{i,i} \\ b_{i,i} \end{bmatrix} = \begin{bmatrix} \mathrm{Var}(x_i^2) & \mathrm{Cov}(x_i^2, x_i) \\ \mathrm{Cov}(x_i^2, x_i) & \mathrm{Var}(x_i) \end{bmatrix} \begin{bmatrix} a_{i,i} \\ b_{i,i} \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} .$$

The existence of unique solutions to these systems of linear equations requires that the exposures and their products are not perfectly correlated. The matrices $\Sigma_{i,j}$ allow us to capture this. Note that when $i \neq j$, $\Sigma_{i,j}$ is the 3-by-3 covariance matrix of the two exposures x_i, x_j, and their product $x_i x_j$. When $i = j$, $\Sigma_{i,j}$ is the 2-by-2 covariance matrix of the exposure x_i and its square x_i^2. For a given pair $i, j \in V_o$ (whether distinct or not), a unique solution to the corresponding system of linear equations above exists if $\det(\Sigma_{i,j}) > 0$. To understand this, note that the determinant of a covariance matrix is a quantitative measure of the linear dependence of a set of random variables, which prompted Wilks [60] to refer to $\det(\Sigma)$ as the "generalized variance".

In Appendix B.1, we derive the coefficients in closed form. The coefficients can be expressed as simple functions of various statistics of the joint distribution of exposure pairs. The exposure distribution depends on the underlying bipartite graph and the experimental design, both of which are known to the experimenter. Thus, the coefficients of the variance estimator can be computed before the experiment begins. In the case of an independent cluster design, the coefficients can be computed exactly. For more complicated designs, the statistics in the coefficients can be estimated to high precision via a Monte Carlo procedure [23]. Throughout the paper, we assume that the exact coefficients are used.

5.2. Unbiased and consistent variance estimation

The following theorem demonstrates that, under certain condition on the exposure distribution described above, the variance estimator is unbiased.

Assumption 4 (Non-degenerate Exposures)**.** *For each pair of outcome units $i, j \in [n]$, the joint distribution of their exposures x_i, x_j satisfies the non-degeneracy condition* $\det(\Sigma_{i,j}) > 0$.

Theorem 5.1 (Unbiased Variance Estimator)**.** *Under Assumption 4 and the linear response assumption, the variance estimator of the* ERL *point estimator is unbiased, i.e.* $\mathbb{E}[\widehat{\mathrm{Var}}(\hat{\tau})] = \mathrm{Var}(\hat{\tau})$.

As discussed above, the exposure distribution is known to the experimenter, so they can check whether the non-degeneracy conditions in Assumption 4 hold before the experiment is run. In particular, this information may inform the experimenter's choice of design. When $i = j$, the condition states that x_i and x_i^2 are not perfectly correlated random variables, which occurs exactly when x_i is supported on at least three distinct values. It is often possible to ensure this condition holds through careful selection of the experimental design. However, such a condition cannot be satisfied when an outcome unit is incident to only a single diversion unit, as only two exposures are ever observed in that case. This impossibility is in line with the fact that unbiased variance estimation is generally not possible in experimental settings with binary treatments. In the case of a pair of exposures (i.e. $i \neq j$), the condition states that exposures x_i, x_j, and their product $x_i x_j$ are not perfectly correlated random variables. Again, this may be achieved by careful selection of the experimental design; however, this condition cannot be satisfied when two outcome units have identically weighted edges (i.e. $w_{i,k} = w_{j,k}$ for all $k \in V_d$), because this would imply that the exposures are identical, $x_i = x_j$.

When Assumption 4 does not hold, our proposed variance estimator will be ill-defined or biased. Indeed, it is possible that no unbiased variance estimator exists in such settings. In this case, one can replace the problematic $\mathrm{Cov}(\hat{\tau}_i, \hat{\tau}_j)$ terms, which cannot be estimated directly, with upper bounds that can be estimated. For example, if the exposure x_i takes only two values, so that $\det(\Sigma_{i,i}) = 0$, then one can replace the problematic term in the variance with the upper bound: $\mathrm{Cov}(\hat{\tau}_i, \hat{\tau}_i) = \mathrm{Var}(\hat{\tau}_i) = \mathbb{E}[\hat{\tau}_i^2] - \mathbb{E}[\hat{\tau}_i]^2 \leq \mathbb{E}[\hat{\tau}_i^2]$, as Aronow and Samii [3] do when they invoke Young's inequality. To unbiasedly estimate this term, we can modify our weighting estimator as $R_{i,i}(x_i) = Q_{i,i}(x_i)$. Similarly, if $\det(\Sigma_{i,j}) = 0$ for some distinct outcome units $i \neq j$, then one can replace the corresponding covariance term with an upper bound obtained from the Cauchy-Schwarz and AM–GM inequalities:

$$\mathrm{Cov}(\hat{\tau}_i, \hat{\tau}_j) \leq \sqrt{\mathrm{Var}(\hat{\tau}_i)\,\mathrm{Var}(\hat{\tau}_j)} \leq \frac{1}{2}\Big(\mathrm{Var}(\hat{\tau}_i) + \mathrm{Var}(\hat{\tau}_j)\Big) \ .$$

An unbiased estimator for the above term may be obtained by modifying the weighting function so that $R_{i,j}(x_i, x_j) = 1/2 \cdot (R_{i,i}(x_i, x_i) + R_{j,j}(x_j, x_j))$. Under Assumptions 1 and 2, replacing one of these individual terms in this way leads to a positive bias of the normalized variance estimator $n \cdot \widehat{\mathrm{Var}}(\hat{\tau})$ which is on the order $\mathcal{O}(1/n)$. Thus, the variance estimator remains asymptotically unbiased as long as we apply these upper bounds to $o(n)$ terms. This is stated formally in the proposition below:

Proposition 5.2. *Let $\mathcal{E} = \{(i,j) \in V_o \times V_o : \det(\Sigma_{i,i}) = 0\}$ be the pairs of outcome units for which Assumption 4 is not satisfied. Consider the alternative variance estimator $\widehat{\mathrm{Var}}_{con}(\hat{\tau})$ defined by the new weighting function:*

$$R'_{i,j}(x_i, x_j) = \begin{cases} R_{i,j}(x_i, x_j) & \text{if } (i,j) \notin \mathcal{E} \\ Q_{i,j}(x_i, x_i) & \text{if } (i,j) \in \mathcal{E} \text{ and } i = j \\ \frac{1}{2}(R'_{i,i}(x_i, x_i) + R'_{j,j}(x_i, x_j)) & \text{if } (i,j) \in \mathcal{E} \text{ and } i \neq j \end{cases} \ .$$

Then, under Assumptions 1 and 2, the normalized alternative variance estimator is conservative in expectation with bounded bias: $0 \leq \mathbb{E}[n \cdot \widehat{\mathrm{Var}}_{con}(\hat{\tau})] - n \cdot \mathrm{Var}(\hat{\tau}) = \mathcal{O}(|\mathcal{E}|/n)$. *Thus, the normalized alternative variance estimator is asymptotically unbiased if* $|\mathcal{E}| = o(n)$.

Even when the terms can be estimated without bias, it could still be preferable to apply the bound if $\det(\Sigma_{i,j}) \approx 0$, as the individual covariance estimators would have high variance in this case, causing the overall variance estimator to be imprecise. For the rest of the section, we analyze the variance estimator assuming that Assumption 4 holds exactly, but we conjecture that many of our results will go through for the conservative variance estimator described above.

We now present conditions under which our proposed variance estimator is consistent in mean squared error. To this end, we define $\Delta = \min_{i,j \in [n]} \det(\Sigma_{i,j})$ to be the smallest non-degeneracy measure.

Theorem 5.3. *Under Assumptions 1, 2, and 4, the mean squared error of the normalized variance estimator is bounded as*

$$\mathbb{E}\Big[\big(n \cdot \mathrm{Var}(\hat{\tau}) - n \cdot \widehat{\mathrm{Var}}(\hat{\tau})\big)^2\Big] = \mathcal{O}\Big(\frac{1}{n} \cdot \Big(d_d^3 d_o^7 + \frac{1}{\Delta^2}\Big)\Big) \ .$$

Thus, the normalized variance estimator is consistent if $d_d^3 d_o^7 = o(n)$ *and* $\Delta = \omega(n^{-1/2})$.

In Theorem 5.3, we analyze convergence of the normalized variance estimator to the normalized variance, which is bounded below by a positive constant, according to Assumption 3. This normalization ensures that both the variance and its estimator are on appropriate scales so that the mean squared error does not trivially approach zero. In light of Assumption 3, this says that the variance estimator converges to the variance at a faster rate than the variance of the ERL estimator converges to zero. A stronger requirement for the rate of convergence yields a stronger restriction on the growth conditions of the bipartite graph. We also require that the degeneracy measure Δ does not approach zero too quickly, which is a quantitative strengthening of Assumption 4.

Theorem 5.3 holds for a broad class of designs under Assumption 2, including Bernoulli randomization at the level of individual diversion units. An experimental design that takes the structure of the bipartite graph into account would generally yield consistency under weaker conditions than those needed for Theorem 5.3. In particular, a design which decorrelates the individual covariance estimators $\widehat{C}_{i,j}$ and make their variances small will yield improved precision of the variance estimator. Interestingly, improving the precision of the ERL estimator and improving the precision of its variance estimator are generally different goals; a design which improves one may not necessarily improve the other. A detailed investigation into this trade-off is beyond the scope of the current paper.

5.3. Asymptotically valid confidence intervals

We may now use our variance estimator together with the asymptotic normality to construct well-motivated confidence intervals. We construct a confidence interval at the $1-\alpha$ confidence level as

$$\hat{\tau} \pm \Phi^{-1}(1-\alpha/2)\sqrt{\widehat{\mathrm{Var}}(\hat{\tau})} \ ,$$

where $\Phi^{-1} : [0,1] \to \mathbb{R}$ is the quantile function of the standard normal deviate. The following result follows from the consistency of the variance estimator (Theorem 5.3) together with asymptotic normality of the ERL estimator (Theorem 4.3).

Corollary 5.4. *Under Assumptions 1-4 and further supposing that* $d_o^4 d_d^{10} = o(n)$ *and* $\Delta = \omega(n^{-1/2})$*, the Wald-type confidence interval using the proposed variance estimator is asymptotically valid:*

$$\lim_{n\to\infty} \Pr\Big(\tau \in \Big[\hat{\tau} \pm \Phi^{-1}(1-\alpha/2)\sqrt{\widehat{\mathrm{Var}}(\hat{\tau})}\Big]\Big) = 1-\alpha \ .$$

It is possible that the proposed variance estimator may take negative values for some treatment assignments. This may happen when the variance is near zero and the variance estimator is imprecise relative to the variance. When the variance estimator takes a negative value, this construction of confidence intervals is not well-defined. Experimenters may here opt for a more conservative estimator of the variance in order to ensure that it never is negative. For example, in the simulation study, we use the absolute value of the estimator described here to ensure non-negativity.

6. Analyzing ERL without the linear response assumption

Our previous analysis of the ERL estimator relied on the linear response assumption. In this section, we show that without the linear response assumption, the ERL estimator can be interpreted as capturing as an average of linear approximations of each unit's dose response function to treatment intensities among the relevant diversion units.

The following theorem derives the expectation of the ERL estimator without the linear response assumption.

Theorem 6.1. *Let the potential outcome functions be arbitrary functions of the exposures:* $Y_i(\boldsymbol{z}) = Y_i(x_i)$*. Then, the expectation of the* ERL *estimator is*

$$\mathbb{E}[\hat{\tau}] = \frac{1}{n}\sum_{i=1}^{n} \tilde{\beta}_i \ ,$$

where $\tilde{\beta}_i$ *is the coefficient of the exposure* x_i *in a unit-specific OLS regression of* Y_i *on* x_i*:* $\tilde{\beta}_i = \frac{\mathrm{Cov}(x_i, Y_i)}{\mathrm{Var}(x_i)}$.

Theorem 6.1 shows that under a general (non-linear) response assumption, the ERL estimator may be interpreted as estimating the average of the slopes of the best linear fit of the outcome to the exposure. We emphasize that this regression cannot be run by the experimenter because the outcomes are not known.

This result is related to several previous results within and outside causal inference. Realizing that most conditional expectation functions are not linear, some statisticians and econometricians have advocated for an interpretation of linear regression as capturing an interpretable approximation of the underlying relationship between the outcome and the regressors [13, 41, 26]. Specifically for causal inference, Angrist [1] highlights that when linear regression is used to estimate treatment effects in an observational setting, the estimator captures a variance-weighted average of unit-level causal effects (see also [4] and [56]). In a vein similar to these results, Theorem 6.1 shows that the ERL estimator captures a policy-relevant causal quantity even if the linear response assumption does not hold. The difference is that the effect it captures is an unweighted average over the units, and the approximation is with respect to each unit's response function.

Under the linear response assumption, this regression-based estimand is equal to the average total treatment effect (ATTE) defined in Section 3.3. However, these two estimands will not coincide for arbitrary response functions and designs. Aside from the linear response assumption, there are several scenarios where we would expect the ATTE and the regression-based estimand to be similar. One such scenario is when the design very closely approximates the Bernoulli design with respect to the exposures, so that exposures have mean 1/2 and concentrate around 0 and 1. When the design is exactly Bernoulli the regression-based estimand is exactly equal to the ATTE, which matches the intuition from the no-interference setting. Another scenario is when the response function is well-approximated by a linear function. An extensive investigation into formal conditions under which the regression-based estimand and the causal estimand (ATTE) are equivalent or similar is beyond the scope of this paper.

7. A cluster design for targeting exposure distribution

In this section, we describe Exposure-Design, an independent cluster design which aims to improve precision of the ERL estimator by constructing a desirable exposure distribution. To this end, we first show in Section 7.1 that increasing the variance of exposures and decreasing the covariance between exposures can lead to improved precision of the ERL estimator in settings of interest. We use the formal results in this setting as inspiration for a heuristic for more general settings. In Section 7.2, we present a clustering objective that aims to achieve such exposure distributions, thereby improving the precision of ERL estimator. Finally, we present a heuristic algorithm for optimizing this clustering objective in Section 7.3.

7.1. An ideal exposure distribution

Like all linear estimators, the ERL estimator will incur a large mean squared error when the coefficients for the observed outcomes are large. In particular, if the variance of an exposure $\mathrm{Var}(x_i)$ is close to zero, the corresponding term of the estimator in equation (1) on page 472 will become large, yielding a high mean squared error even though the estimator is unbiased. In general, experimenters should use designs for which the corresponding exposure variances are large.

However, large exposure variances should not be the only property of the exposure distribution that experimenters focus on. Consider a naive design that places equal probability on two treatment vectors: either all diversion units receive treatment ($\boldsymbol{z} = \mathbf{1}$) or all diversion units receive control ($\boldsymbol{z} = \mathbf{0}$). Under this design, all of the exposures have variance ones, which is the largest possible variance in this setting. However, we observe either all of the treatment outcomes or all of the control outcomes, but never a mix of the two; in fact, the estimator itself takes only two values. Thus, the ERL estimator will be very imprecise under this design, despite the individual exposure variances being as large as possible. This raises the question: how should we construct a design that improves the precision of the ERL estimator?

This is a challenging task, since the precision of the ERL estimator depends on the unobserved outcomes. Indeed, a universally optimal design does not exist [29]. However, we argue that a good heuristic is to construct the design so that the variance of the exposures are large and the covariances between most pairs of exposures are close to zero. As discussed at the end of Section 4.1, a design which directly targets these aspects of the exposure distribution can be expected to ensure high precision of the ERL estimator under weaker growth conditions on the bipartite graph than those presented in our analysis.

As an illustration to motivate this heuristic, consider the scenario where all of the individual treatment effects are zero, i.e. the response functions are of the form $Y_i(x_i) = \alpha_i$. Studies of these sort are sometimes called uniformity trails or A/A tests. In this scenario, the MSE of the ERL estimator is

$$\mathbb{E}[(\hat{\tau} - \tau)^2] = \frac{1}{n^2}\left[\sum_{i=1}^{m} \alpha_i^2 \frac{1}{\mathrm{Var}(x_i)} + 2\sum_{i<j} \alpha_i \alpha_j \frac{\mathrm{Cov}(x_i, x_j)}{\mathrm{Var}(x_i)\,\mathrm{Var}(x_j)}\right] .$$

As the individual variance terms increase, the first sum decreases. The effect of the second term depends on the sign of the product of intercepts, $\alpha_i\alpha_j$. Generally speaking, these intercepts are unknown to the experimenter. For the sake of this discussion, consider when the outcomes $Y_i(\boldsymbol{z})$ are non-negative, in which case all intercepts α_i and their products $\alpha_i\alpha_j$ are non-negative. In this case, decreasing the correlation between exposures would decrease the second term, leading to an overall decrease in the MSE of the ERL estimator. Note that the same hold if all α_i are negative. If the α_i have mixed signs, it is beneficial to introduce positive correlation when $\alpha_i\alpha_j < 0$, but it will generally not be possible to know at the design stage for which pairs of units this holds.

7.2. Clustering objective for targeting exposure distribution

In the previous section, we noted that a reasonable heuristic is to assign treatments to the diversion units so that the variance of exposures is large and the covariance between most exposures is small. However, as argued in Section 3.2, constructing a treatment distribution which realizes a desired exposure distribution is generally not possible due to overlapping structures in the bipartite graph. In this section, we present an optimization formulation for an independent cluster design that aims to achieve large exposure variance and small correlations between exposures, to the extent that this is possible given the bipartite graph.

We propose choosing a cluster design which maximizes the following objective function:

$$\max_{\text{clustering } \mathcal{C}} \quad \sum_{i=1}^{n}\left[\text{Var}(x_i) - \phi \sum_{i \neq j} \text{Cov}(x_i, x_j)\right] . \qquad \text{(Exposure-Design)}$$

The variance and covariance of the exposures above are with respect to the random treatment assignments of the corresponding independent cluster design. The first term in the objective is the sum of the exposure variances, so maximizing this term will encourage large exposure variances. The second term penalizes positive correlation between exposures, and maximizing it encourages small correlation. It is impossible to induce a negative correlation between exposures in the class of independent cluster designs, so the second term of the objective attains its maximum when the exposures are uncorrelated. The correlation penalizing parameter $\phi \geq 0$ controls the relative emphasis between large exposure variances and small exposure correlations. When $\phi = 0$, then the emphasis is placed entirely on increasing individual exposure variance; this is typically undesirable, as the optimal solution is often a single cluster containing all diversion units, which results in a design where either all diversion units receive treatment or all diversion units receive control. Increasing ϕ places more emphasis on decorrelating exposures.

A key insight to solving the Exposure-Design formulation is that it may be reformulated as a *correlation clustering* problem, which is well-studied in the algorithms literature [8, 58, 14]. The existing computational understanding of these correlation clustering problems is another reason to use the Exposure-Design objective. The following proposition states the re-formulation of the Exposure-Design objective into the correlation clustering variant, denoted Corr-Clust.

Proposition 7.1. *For each pair of diversion units $i, j \in V_d$, define the value $\omega_{i,j} \in \mathbb{R}$ as*

$$\omega_{i,j} = (1+\phi)\sum_{k=1}^{m} w_{k,i}w_{k,j} - \phi\Big(\sum_{k=1}^{m} w_{k,i}\Big)\Big(\sum_{k=1}^{m} w_{k,j}\Big) , \qquad (2)$$

where $w_{k,i}$ is the weight of the edge between the kth outcome unit and the ith diversion unit. Exposure-Design *is equivalent to the following clustering prob-*

lem:

$$\max_{\text{clusterings } \mathcal{C}} \quad \sum_{C_r \in \mathcal{C}} \sum_{i,j \in C_r} \omega_{i,j} \quad . \tag{Corr-Clust}$$

Although Corr-Clust is a variant of the weighted maximization-type correlation clustering problems previously studied in the literature [14, 58], it is not equivalent to previously studied formulations in an approximation-preserving sense, as it takes positive and negative values. Given that weighted maximization correlation clustering is NP-Hard [14], it is reasonable to presume that our formulation Corr-Clust is also computationally hard. However, these computational complexity considerations are beyond the scope of this paper.

Exposure-Design places no explicit constraint on the number of clusters produced by the clustering algorithm. However, our analysis in Section 4.1 suggests that limiting the cluster sizes, and thereby reducing correlation between exposures, helps to achieve consistency and normality of the ERL estimator. This desirable cluster structure is not captured by the optimization problem itself, but we handle it through our local search heuristic described in Section 7.3. The Exposure-Design objective does not directly minimize MSE of the ERL estimator, but should instead be understood as a useful heuristic.

The Exposure-Design is conceptually similar to the correlation-clustering based design presented in [49], but it differs in several key ways. Exposure-Design provides experimenters the flexibility to trade-off larger exposure variances with more de-correlated exposures by setting the parameter ϕ. In contrast, the cluster design of Pouget-Abadie et al. [49] focuses solely on the exposure variance by maximizing what is referred to as "empirical dose variance" in their paper. As we demonstrate in Appendix C, their objective is equal to ours when the trade-off parameter is set to $\phi = 1/(n-1)$. In this sense, the cluster design of Pouget-Abadie et al. [49] can be viewed as a specific instance of the more general Exposure-Design, where a greater emphasis is placed on maximizing the exposure variances. More importantly, the Exposure-Design presented in this paper is is arguably better aligned with the objective of minimizing variance of the ERL estimator, while the correlation-clustering based design of Pouget-Abadie et al. [49] is motivated by the intuition that extreme exposures are helpful in this setting, without any reference to an explicit estimator or its statistical properties.

7.3. Local search heuristic for Exposure-Design

We now describe a local search heuristic for optimizing Exposure-Design. The local search is initialized with the singleton clustering and iteratively seeks to improve the clustering. In each iteration, the algorithm loops through random pairs of diversion units $i, j \in V_d$ and moves diversion unit j to the cluster currently containing diversion unit i if that change improves the objective value, subject to a user-defined constraint on the clusters. The local search algorithm

is presented more formally below as Algorithm 1.

Algorithm 1: Local Search($\boldsymbol{W}, \phi, k, T$, cluster constraints)

```
Initialize singleton clustering C = {{1}, {2}, ... {m}}
for iterations t = 1 ... T do
    Choose permutation π on the diversion units uniformly at random.
    for diversion units i ∈ π do
        Randomly select a diversion unit j with probability proportional
          to (W^⊺ W)_{i,j}
        Let C and C' be the clusters containing diversion units i and j,
          respectively.
        if moving j from cluster C' to C increases objective value and
          satisfies user-defined constraints then
            Move diversion unit j from cluster C' to C.
return clustering C
```

Given a diversion unit i, we use *wedge sampling* to select unit j proportional to $(\boldsymbol{W}^\intercal \boldsymbol{W})_{i,j} = \sum_{k=1}^{n} w_{k,i} w_{k,j}$ [17]. We use wedge sampling because picking pairs of units for which $(\boldsymbol{W}^\intercal \boldsymbol{W})_{i,j}$ is large often results in a large correlation clustering weight $\omega_{i,j}$. Performance improvements are obtained by computing the correlation clustering weights $\omega_{i,j}$ only when they are needed to evaluate changes in the objective. In particular, the first term of equation (2) on page 484 is an inner product whose computation scales with the sparsity of the bipartite graph and the second term is the product of sums that can be pre-computed.

Diversion unit j is moved into the cluster containing diversion unit i if two conditions are met: the objective increase and the user-defined cluster constraints are satisfied. We recommend that experimenters choose constraints which limit the cluster sizes in some way. For example, the experimenter may choose to constraint the number of diversion units within a cluster. In our implementation, we constrain the sum of the (unweighted) degrees of diversion units within a cluster to be a fixed fraction of the total number of edges. In this way, no cluster has too many outgoing edges to outcome units. This implicitly limits the amount of dependence between exposures, which is one of the key aspects underlying the design conditions in Assumption 2 of our analysis.

Our local search algorithm is different from the one presented in [49], which approximates the Gram matrix $\boldsymbol{W}^\intercal \boldsymbol{W}$ offline as the sum of a sparse matrix and a rank-one matrix, so that the algorithm works with an approximation to the objective. In contrast, our algorithm accepts and rejects changes based on the exact value of the objective. Relative to [22], this local search does not consider moving units to new empty clusters, nor does it consider merging clusters. Moves of the first type seem consistently unprofitable in our setting. As for merges, we find that the algorithm is able to essentially perform them by moving one diversion unit at a time.

8. An application to online marketplace experiments

In this section, we apply our proposed methodology to a simulated marketplace experiment based on a product review dataset from the Amazon marketplace [43, 30]. The Amazon product review dataset contains 83 million reviews made by 121 thousand customers on 9.8 million items. In this application, we imagine running an experiment where we change the pricing mechanism of items in the marketplace, and are interested in how customers' satisfaction is affected by this change. The items sold in the marketplace are the diversion units and the customers in the marketplace are the outcome units.

We do not observe customer demand nor purchasing history in this data set. Instead, we use whether a customer reviewed an item as a proxy for item demand. Thus, an edge is present in the bipartite graph if a customer reviewed an item and all edges incident to an outcome unit are uniformly weighted. This means that a customer's exposure is the unweighted average of the treatment status of the items they have previously reviewed. While the use of the review proxy for demand makes the simulation quite stylized, we believe the proxy sufficiently aligned with item demand to make the simulation informative. But we use this dataset only to illustrate our statistical methodology on a bipartite graph arising from real online interactions; we do not claim that our simulation is representative of online marketplaces more generally.

We generate potential outcomes in the simulation via an exposure-response function. The outcomes could be seen as the satisfaction of a customer given their exposure. The responses in this study are simulated, but we can imagine that they are either reported directly by a customer or inferred from the customer's future behavior. In the case of a linear response, a positive slope indicates an increase in customer satisfaction as a result of the new pricing mechanism, while a negative slope indicates a decrease in satisfaction as a result of the new pricing mechanism.

We preprocess the Amazon produce review dataset for computational tractability in the same manner as Pouget-Abadie et al. [49]. We begin by removing customers that have reviewed fewer than 100 items. Next, we execute a balanced partitioning algorithm [6] on the entire bipartite graph to create groups of customers and groups of items. After this preprocessing, we define the diversion units to be the item groups and the outcome units to be the customer groups. The resulting bipartite graph has one thousand outcome units, 2.4 million diversion units, and 7.1 million edges. Given the denseness of edges, this bipartite graph can hardly be seen as satisfying the growth conditions used to prove consistency and asymptotic normality (specified in Section 4.1). This application may therefore be seen as a stricter test of the performance of the proposed design and estimator, as the current setting would require weaker growth conditions.

We investigate the statistical properties of the ERL estimator, the variance estimator, and the resulting confidence intervals under various treatment designs in this application. In particular, we compare our proposed Exposure-Design to several existing designs: the Bernoulli design, the correlation clustering design of Pouget-Abadie et al. [49], and the balanced partitioning cluster design of [19],

as implemented by Aydin et al. [6]. Although the balanced partitioning design was not developed for the bipartite setting, we may expect it to achieve high precision estimates if the clustering produces decorrelated exposures with large variances.

We generate the potential outcomes in three simulations, where we vary the response functions that are used. The first two simulations feature linear response functions and the third simulation features a non-linear response function. We draw the parameters at random, but keep them (and thus the potential outcomes) fixed between simulation rounds. These simulation settings are listed below.

- **S1: (Mostly) Positive Effects.** In this simulation, we set almost all of the individual treatment effects to be positive. More precisely, we sample the slope terms as $\beta_i \sim \mathcal{N}(2,1)$ and the intercept terms as $\alpha_i \sim \mathcal{N}(-1,3/8)$.
- **S2: (Nearly) Zero Effects.** In this simulation, we set all the individual treatment effects close to zero, while varying the baseline outcomes. The outcomes are chosen to be mostly positive. More precisely, we sample the slope terms as $\beta_i \sim \mathcal{N}(0,1/2)$ and the intercept terms as $\alpha_i \sim \mathcal{N}(2,3/8)$.
- **S3: Non-Linear Response.** In this simulation, we use a non-linear response function to specify the potential outcomes. In particular, the response of outcome unit i is $Y_i(x_i) = 4x_i(x_i-1)+\alpha_i$, where $\alpha_i \sim \mathcal{N}(0,1/8)$. Under this response, all individual treatment effects are 0. Because the linear response assumption is not satisfied, we should not expect that our statistical analyses (unbiasedness, consistency, normality, etc) will hold exactly.

We run Exposure-Design with different values of the correlation penalty parameter ϕ, chosen from a grid of ten points between $[0,2]$. The clustering itself is obtained using our local search heuristic presented in Section 7.3. Recall that the correlation clustering objective of Pouget-Abadie et al. [49] may be obtained by setting $\phi = 1/(n-1)$ for the Exposure-Design. For this reason, we compute the corresponding cluster by running our local search heuristic with $\phi = 0.001 \approx 1/(n-1)$.

Table 1
Simulation results

		Exposure Design ($\phi = 0.223$)	Exposure Design ($\phi = 1.0$)	Correlation Clustering	Balanced Partitioning	Bernoulli
S1	RMSE	0.049	0.057	0.087	0.0718	0.659
	CI Width	0.220	0.239	0.329	0.284	2.576
	CI Coverage	91.5%	91.5%	94.0%	94.1%	95.1%
S2	RMSE	1.81	2.24	2.05	1.86	43.83
	CI Width	7.04	8.72	8.00	7.21	190.8
	CI Coverage	94.8%	94.8%	95.0%	94.5%	95.1%
S3	RMSE	0.86	1.09	0.90	0.78	24.37
	CI Width	3.47	4.35	3.82	3.39	95.15
	CI Coverage	95.7%	95.5%	96.9%	96.9%	95.1%

A summary of the main results from these simulations appears in Table 1. For each treatment design and simulation, we sample $15,000$ exposure vectors, compute the observed outcomes, and construct the corresponding ERL and variance estimators. Given the ERL and variance estimators, we construct the confidence intervals as described in Section 5, with absolute value corrections when the variance estimator takes a negative value. For each simulation setting and treatment design, we report the root mean square error (RMSE) of the ERL estimator, the average width of the 95% confidence intervals, and the coverage of the 95% confidence intervals.

In Table 1, we show results for EXPOSURE-DESIGN with parameters $\phi = 0.223$ and $\phi = 1.0$. The parameter value $\phi = 1.0$ is chosen arbitrarily, while the value $\phi = 0.223$ is the value for which EXPOSURE-DESIGN typically achieves the smallest mean squared error across each of the simulation settings. We emphasize that selecting ϕ in this way (i.e. a course grid search to find the smallest MSE) cannot be performed by an experimenter and is presented here only to inform our discussion. Still, it is interesting that a single penalty parameter minimized MSE for all simulations considered here. See Figure 3 for comparison of the mean squared error of EXPOSURE-DESIGN when the correlation penalty parameter ϕ is varied.

We draw particular attention to a few features in these results. EXPOSURE-DESIGN achieves the smallest RMSE in the simulations which satisfy the linear response assumption. All cluster-based designs achieve significantly smaller RMSE than the Bernoulli design, which emphasizes the importance of carefully considering the exposure distribution when the growth conditions (specified in Section 4.1) are not satisfied. The confidence intervals in Simulation 1 under EXPOSURE-DESIGN cover notably below the nominal 95% level, indicating that either the sampling distribution of the ERL estimator is not sufficiently approximated by a normal or the variance estimator is not sufficiently concentrated at this sample size. This does not contradict the theory developed earlier in this paper, because the growth conditions required for validity of the confidence intervals can hardly be seen as holding with such a dense graph. The confidence intervals in Simulation 3 cover slightly above the nominal 95% level, which is a result of conservative bias in the variance estimate due to non-linearity of the response.

Figure 1 contains histograms of the ERL estimator for each simulation and design, where the rows correspond to the designs and the columns correspond to the simulations. The dotted vertical line in the plot is the true ATTE. In all simulations, the distribution of the ERL estimator appears unimodal, centered at or close to the ATTE, and (roughly) normal. This is to be expected for Simulations 1 and 2 where the linear response assumption holds.

Perhaps surprisingly, the statistical analyses seem to hold in Simulation 3, which features a highly non-linear response. The apparent unbiasedness of the ERL estimator here can be explained by Theorem 6.1 in the following way: the quadratic responses in Simulation 3 yield zero treatment effect for all units. Although the best linear approximation to each quadratic response does not well-approximate the quadratic response itself, the linear approximation has

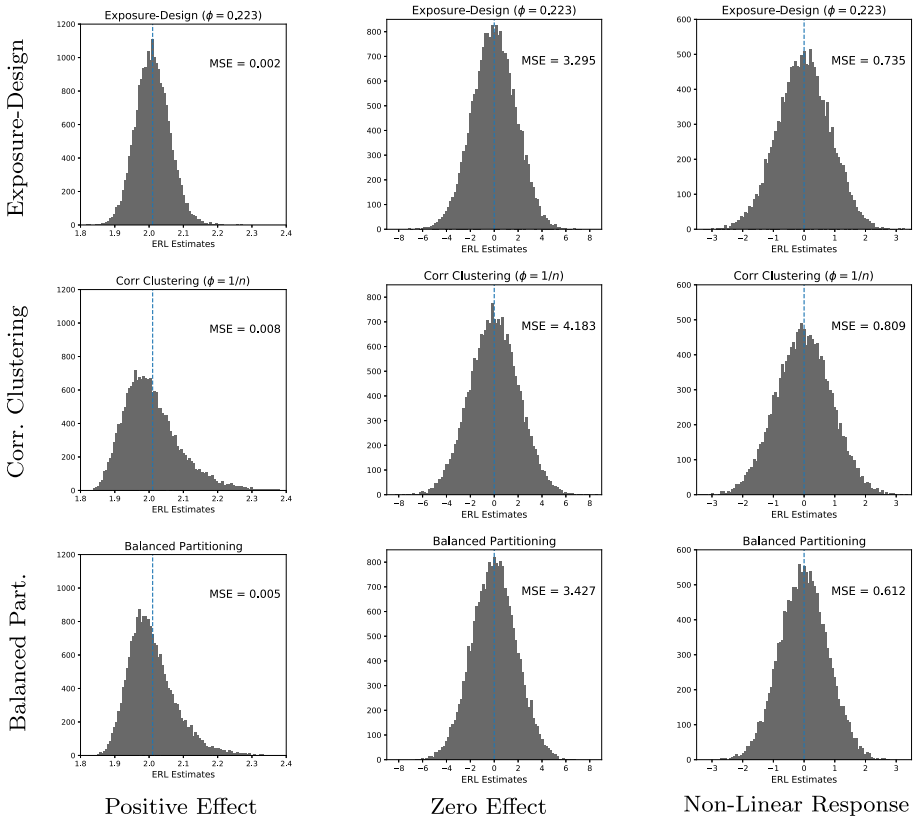

FIG 1. *Histograms of the* ERL *estimator in simulations.*

zero slope and so, in this sense, captures each individual ITE—and thus the ATTE—exactly.

Figure 2 contains histograms of the variance estimator for each simulation and treatment design, where the rows correspond to the designs and the columns correspond to the simulations. The dotted vertical line in the plot is the is the true variance, as estimated by the empirical distribution of the estimator from the simulations. The variance estimator is unbiased in Simulations 1 and 2, which aligns with Theorem 5.1. However, because the empirical variance estimate is used, the blue line is close to (but not exactly) the true variance. Increasing the number of sampled exposure vectors decreases this error, but drawing more than 20 thousand samples is prohibitively expensive given the size of the data. In Simulation 1, the mean squared error of all cluster-based designs is so small that the variance estimator takes negative values with non-negligible probability. In Simulation 3, the response is highly non-linear so that the variance estimator incurs a positive bias, which results in a coverage slightly above the nominal level. Interestingly, the variance estimator under the balanced

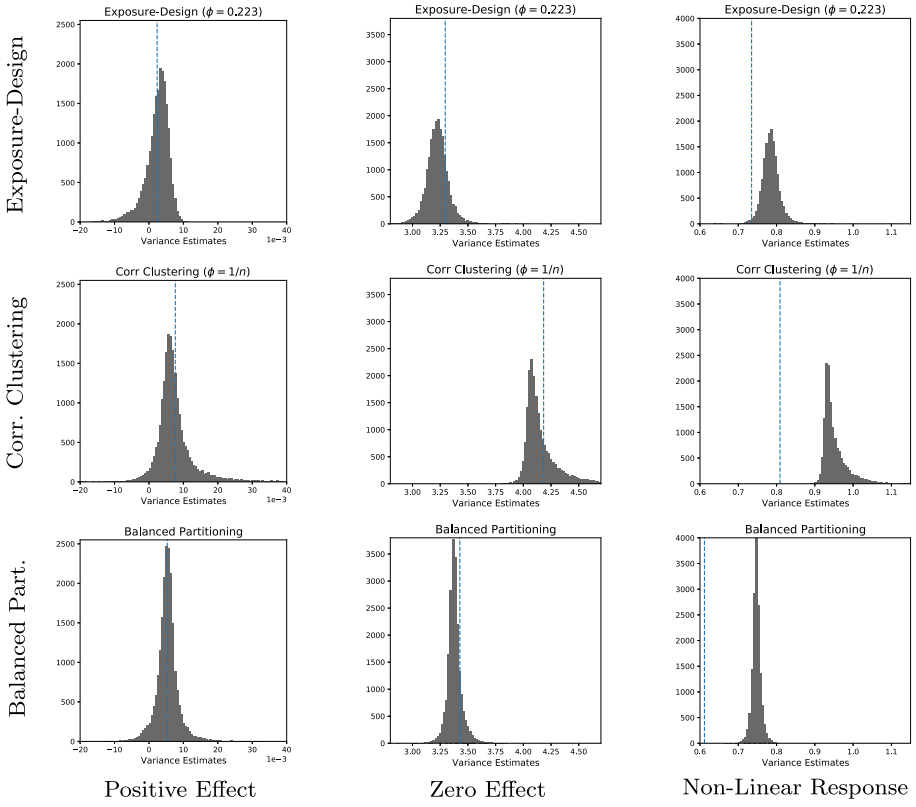

FIG 2. *Histograms of the variance estimator in simulations.*

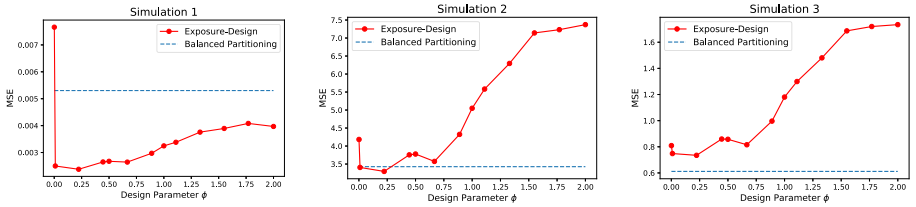

FIG 3. *MSE of the* ERL *estimator as trade-off parameter ϕ is varied. First values of ϕ are* .001 *and* .01.

partitioning design is more concentrated around its mean, which is worth further investigation but is beyond the scope of this paper.

Figure 3 contains a plot for each simulation, where the mean squared error is plotted against the correlation penalizing parameter ϕ. The mean squared error of the balanced partitioning design appears as a dotted blue line. In Simulations 1 and 2 where the linear response assumption holds, there is a range of

values of ϕ where EXPOSURE-DESIGN achieves lower mean squared error than the balanced partitioning design. In our simulations, the choice of $\phi \approx 1/4$ typically achieves lowest mean squared error. While EXPOSURE-DESIGN performed better than balanced partitioning in some settings we investigated, the simulation exercise does not support the conclusion that EXPOSURE-DESIGN is generally preferable over balanced partitioning. Indeed, no design is optimal across all types of potential outcomes [29]. We encourage experimenters to select ϕ (and more generally, select designs) by running tests on simulated data.

9. Concluding remarks

The Exposure Reweighted Linear (ERL) estimator provides a way to estimate and obtain confidence intervals for the average total treatment effect in bipartite experiments under a linear exposure-response assumption.

When employing the EXPOSURE-DESIGN design in practice, we recommend that the experimenter choose the value of the trade-off parameter ϕ by running simulations of the experiment using available models of the outcomes when possible. When this is not possible, we find in our simulations that $\phi \approx 1/4$ typically yields improvements in the precision of the ERL estimator over the previously proposed correlation clustering design of Pouget-Abadie et al. [49] where $\phi = 1/(n-1)$. We suspect that in most settings of interest, the ERL estimator will enjoy increased precision under any treatment design that ensures that exposures have large variance and are decorrelated, either explicitly or implicitly.

The performance of the ERL estimator is sensitive to the choice of experimental design. For example, the precision of the ERL estimator under Bernoulli randomization will generally be poor. As we saw in our simulation study, the root mean square error of the ERL estimator under Bernoulli randomization was about ten times as large as under the clustered designs. For this reason, experimenters should avoid designs, such as Bernoulli and complete randomization, that do not take the structure of the bipartite graph into account. If such a design is unavoidable, or if precision is inadequate despite a deliberate choice of design, experimenter can consider implementing covariate adjustments. It is beyond the scope of this paper to formally extend the ERL estimator to accommodate complex adjustments. A simple approach we expect will work well when the covariates are discrete (or discretized) is post-stratification [44]. With this approach, the ERL estimator is applied separately in subsets of the sample based on covariate information. These estimators are then aggregated to an overall estimator in a second step. If the covariates used to form the bins are informative of the potential outcomes and the number of bins is small relative to the number of units, this approach is expected to improve precision – sometimes drastically so. A related approach is to incorporate covariate information already in the design stage, with the aim of ensuring that the exposures are unrelated to covariates for the realized assignment. This could be implemented in a cluster-type design by introducing dependencies between the treatment

assignments in different clusters. One way of introducing such dependencies is rejection sampling, or re-randomization, where assignments are drawn until one with acceptable balance properties are found [40, 38]. However, the theory described in this paper does not apply when there are dependencies between clusters.

There are several open questions suggested by this work. Given that the design we describe in this paper is heuristically motivated, there is likely room for large improvements. One technical challenge is to construct a design for which consistency and asymptotic normality of the ERL estimator may be established under weaker growth conditions on the bipartite graph. Another important open question is whether it is possible to develop methods for valid inference in bipartite experiments that go beyond the linear exposure-response assumption. Finally, it would be of practical and methodological interest to develop estimation techniques that are robust to misspecification in the bipartite graph as well as estimation techniques that perform well in the presence of greater structural interference (i.e., when outcomes are influenced by the exposures of other units). The results in Sävje [53] regarding estimation of treatment effects under a misspecified exposure mapping might extend to this setting, but that remains to be shown. Answering these methodological questions around the bipartite experimental framework will increase its relevance and applicability in practice.

Appendix A: Analysis of the ERL estimator

In this section, we present proofs of unbiasedness, consistency, and asymptotic normality of ERL estimator appearing in Section 4 of the main paper. Before continuing, we introduce some notation used in the proofs. We begin by defining for each outcome unit $i \in V_o$, an estimate of the individual treatment effect τ_i, which is

$$\hat{\tau}_i \triangleq Y_i(\boldsymbol{z})\Big(\frac{x_i(\boldsymbol{z}) - \mathbb{E}[x_i(\boldsymbol{z})]}{\mathrm{Var}(x_i(\boldsymbol{z}))}\Big) \ .$$

Observe that the ERL estimator is the average of these estimates of the individual treatment effects, i.e. $\hat{\tau} = (1/n)\sum_{i=1}^{n} \hat{\tau}_i$. Throughout the proofs, we will often reason about the behavior of the ERL estimator through the properties of the individual treatment effect estimates.

Next, we introduce the concept of *dependency neighborhoods* [52]. Let a_1, $a_2, \dots a_n$ be random variables indexed by the integers $[n]$ and collect these random variables into the set $\mathcal{A} = \{a_i : i \in [n]\}$. For each variable a_i, we define the *dependency neighborhood* as

$$\mathcal{I}(i) \subset \mathcal{A} \text{ such that } a_i \text{ is jointly independent of the variables } \mathcal{A} \setminus \mathcal{I}(i) \ .$$

In other words, a random variable a_i is jointly independent of all variables not contained in its dependency neighborhood, but is dependent on variables contained in its dependency neighborhood. We take the convention that $i \in \mathcal{I}(i)$ and so that each dependency neighborhood as cardinality at least 1. A measure of dependence between the random variables is the *maximum dependency degree*,

which is $D = \max_{i\in[n]}|\mathcal{I}(i)|$. Note that independent random variables satisfy $D = 1$ and that completely dependent random variables have $D = n$.

For the remainder of the proof, we focus our discussion of dependency neighborhoods and degrees to the collection of errors of the individual treatment effects,

$$a_1 = \tau_1 - \hat{\tau}_1, \quad a_2 = \tau_2 - \hat{\tau}_2, \quad \dots \quad a_n = \tau_n - \hat{\tau}_n \ .$$

We begin by showing that in this case, the maximum dependency degree may be bounded in terms of the degrees of the bipartite graph and the dependence in the treatment assignments.

Lemma A.1. *The dependency degree of the individual treatment effect errors is bounded by* $D \leq kd_d d_o$.

Proof. The first part of this proof is to establish a necessary condition for an individual treatment effect error a_j to be in the dependency neighborhood of a_i, i.e. $a_j \in \mathcal{I}(i)$. We begin by re-writing the exposures under a cluster design. Recall that the exposures are defined as $x_i = \sum_{j=1}^{m} w_{i,j} z_j$. For each cluster $C \in \mathcal{C}$, define $w_{i,C} = \sum_{j\in C} w_{i,j}$ and define z_C to be the ± 1 cluster treatment assignment variable which is 1 if diversion units in C are treated and -1 otherwise. If $w_{i,C} \neq 0$, then we say that cluster C is *incident* to outcome unit i. Define $S(i) = \{z_C : w_{i,C} \neq 0\}$ to be the cluster treatment assignments which influence the exposure x_i. Under the cluster design, the exposure for outcome unit i may be written as

$$x_i = \sum_{C\in\mathcal{C}} w_{i,C} z_C = \sum_{C\in S(i)} w_{i,C} z_C \ .$$

By the linear-response assumption, the individual treatment effect error a_i is a function of the exposure x_i. Moreover, a_i is a function of the cluster treatment assignment variables in $S(i)$. Let us denote this relationship by writing $a_i = g_i(S(i))$, where g_i is a function of the cluster treatment variables $z_C \in S(i)$. Let $B \subset V_o$ be a collection of outcome units. We remark that joint independence of cluster treatment assignments implies joint independence of individual treatment effect errors:

$$S(i) \perp\!\!\!\perp \{S(j) : j \in B\} \Rightarrow a_i \perp\!\!\!\perp \{a_j : j \in B\} \ .$$

Under an independent cluster design, the cluster treatment assignments $S(i)$ are jointly independent of the cluster treatment assignments $\{S(j) : j \in B\}$ when the corresponding sets of clusters are disjoint, i.e. $S(i) \cap (\cup_{j\in B} S(j)) = \emptyset$. Thus, the individual treatment effect estimate a_i is jointly independent of the collection of individual treatment effect estimates $\{a_j : j \in B\}$ when outcome unit i is not incident to any cluster that is incident to an outcome unit in B. In other words, $a_j \in \mathcal{I}(i)$ only if outcome units i and j are incident to a common cluster.

Fix an outcome unit $i \in V_o$. The remainder of the proof is a simple counting argument which uses this necessary condition to establish that $|\mathcal{I}(i)| \leq kd_d d_o$. In particular, we will count the number of outcome units that are incident to

one of the clusters that are incident to i. Because the degree of outcome unit i is at most d_o, it is incident to at most d_o clusters. Each of these clusters has at most k diversion units, by Assumption 2. Because the degree of all diversion units j is at most d_d, the number of outcome units which are incident to at least one of these clusters is at most kd_dd_o. Thus, we have established that

$$D = \max_{i \in V_o} |\mathcal{I}(i)| \leq kd_dd_o \ . \qquad \square$$

The following lemma derives a lower bound the exposure variances in terms of the treatment assignment probability and the maximum degree of the outcome units.

Lemma A.2. *If each pair of treatment assignments is non-negatively correlated, then each exposure variance is lower bounded as* $\mathrm{Var}(x_i) \geq \frac{p(1-p)}{d_o}$.

Proof. We begin by expanding the variance of the exposure x_i by

$$\begin{aligned}
\mathrm{Var}(x_i) &= \mathrm{Var}(\sum_{j=1}^{m} w_{i,j} z_j) \\
&= \sum_{i=1}^{m} \Big[\mathrm{Var}(w_{i,j} z_j) + \sum_{\ell \neq j} \mathrm{Cov}(w_{i,j} z_j, w_{i,\ell} z_\ell) \Big] \\
&= \sum_{i=1}^{m} \Big[w_{i,j}^2 \, \mathrm{Var}(z_j) + \sum_{\ell \neq j} w_{i,j} w_{i,\ell} \, \mathrm{Cov}(z_j, z_\ell) \Big] \\
&\geq \sum_{i=1}^{m} w_{i,j}^2 \, \mathrm{Var}(z_j) \\
&= p(1-p) \sum_{i=1}^{m} w_{i,j}^2 \ ,
\end{aligned}$$

where the inequality follows because the weights $w_{i,j}$ are non-negative and the assignments are non-negatively correlated and the last equality follows because z_i are $0, 1$ random variables with $\Pr(z_i = 1) = p$.

We complete the proof by lower bounding the sum of the squares of the weights. Recall that the sum of the weights is 1 and there are at most d_o non-negative terms in the sum. Using this together with the inequality that relates ℓ_2 to ℓ_1 norms in d-dimensions, $\|\cdot\|_2^2 \geq \frac{1}{d}\|\cdot\|_1^2$, we have that

$$\sum_{i=1}^{m} w_{i,j}^2 \geq \frac{1}{d_o} \sum_{i=1}^{m} w_{i,j} = \frac{1}{d_o} \ . \qquad \square$$

The following lemma is a bound on the moments of the errors of the individual treatment effects.

Lemma A.3. *The sth moment of the error of the individual treatment effect estimates is bounded by*

$$\mathbb{E}[|\tau_i - \hat{\tau}_i|^s] \leq \left[M\Big(2 + \frac{d_o}{p(1-p)}\Big) \right]^s .$$

Proof. We begin by remarking that $|Y_i(\boldsymbol{z})| \leq M$ implies that each of the individual slopes are also bounded in absolute value as $|\beta_i| \leq 2M$. Recall that by the linear response assumption, $Y_i(\boldsymbol{z}) = \beta_i x_i + \alpha_i$ and by the linear exposure assumption (along with the normalization of the edge weights), setting $\boldsymbol{z} = \mathbf{0}, \mathbf{1}$ results in an exposure of $\xi = 0, 1$. Thus, when considering $\boldsymbol{z} = \mathbf{0}, \mathbf{1}$, the bound $|Y_i(\boldsymbol{z})| \leq M$ implies that $|\beta_i + \alpha_i| \leq M$ and $|\alpha_i| \leq M$, which is enough to establish that $|\beta_i| \leq 2M$.

We now proceed by proving a bound on $|\tau_i - \hat{\tau}_i|$, which holds for any realization of the random variables:

$$\begin{aligned}
|\tau_i - \hat{\tau}_i| &= \Big|\beta_i - Y_i(\boldsymbol{z})\Big(\frac{x_i - \mathbb{E}[x_i]}{\mathrm{Var}(x_i)}\Big)\Big| \\
&\leq |\beta_i| + \frac{|Y_i(\boldsymbol{z})| \cdot |x_i - \mathbb{E}[x_i]|}{\mathrm{Var}(x_i)} && \text{(triangle inequality)} \\
&\leq 2M + \frac{M}{\mathrm{Var}(x_i)} && \text{(definition of } M \text{ and above)} \\
&\leq M\Big(2 + \frac{1}{\mathrm{Var}(x_i)}\Big) && \text{(collecting terms)} \\
&\leq M\Big(2 + \frac{d_o}{p(1-p)}\Big) && \text{(Lemma A.2)}
\end{aligned}$$

The moment bound follows by applying the bound above. □

A.1. Expectation of the ERL *estimator (Theorems 4.1 and 6.1)*

In this section, we derive the expectation of the ERL estimator, both with and without the linear exposure-response assumption. First, we derive the expectation under the linear exposure-response assumption.

Theorem 4.1. *Suppose the design is such that each exposure has positive variance,* $\mathrm{Var}(x_i) > 0$. *Under the linear response assumption, the* ERL *estimator is unbiased for the ATTE:* $\mathbb{E}[\hat{\tau}] = \tau$.

Proof. By linearity, the expectation of the estimator is

$$\mathbb{E}[\hat{\tau}] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[Y_i(\boldsymbol{z}) \left(\frac{x_i(\boldsymbol{z}) - \mathbb{E}[x_i(\boldsymbol{z})]}{\mathrm{Var}(x_i(\boldsymbol{z}))} \right) \right] .$$

By Proposition 3.1, the ATTE is the average of the slope terms β_i. Thus, to complete the proof we show that each expectation terms inside the sum is equal

to the corresponding slope β_i. Using the linear response assumption,

$$\begin{aligned}
&\mathbb{E}\left[Y_i(\boldsymbol{z})\left(\frac{x_i(\boldsymbol{z})-\mathbb{E}[x_i(\boldsymbol{z})]}{\mathrm{Var}(x_i(\boldsymbol{z}))}\right)\right] \\
&= \mathbb{E}\left[(\beta_i x_i(\boldsymbol{z})+\alpha_i)\left(\frac{x_i(\boldsymbol{z})-\mathbb{E}[x_i(\boldsymbol{z})]}{\mathrm{Var}(x_i(\boldsymbol{z}))}\right)\right] \\
&= \beta_i \mathbb{E}\left[x_i(\boldsymbol{z})\left(\frac{x_i(\boldsymbol{z})-\mathbb{E}[x_i(\boldsymbol{z})]}{\mathrm{Var}(x_i(\boldsymbol{z}))}\right)\right] + \alpha_i \mathbb{E}\left[\left(\frac{x_i(\boldsymbol{z})-\mathbb{E}[x_i(\boldsymbol{z})]}{\mathrm{Var}(x_i(\boldsymbol{z}))}\right)\right] \\
&= \beta_i\left(\frac{\mathbb{E}[x_i(\boldsymbol{z})^2]-\mathbb{E}[x_i(\boldsymbol{z})]^2}{\mathrm{Var}(x_i(\boldsymbol{z}))}\right) + \alpha_i\left(\frac{\mathbb{E}[x_i(\boldsymbol{z})]-\mathbb{E}[x_i(\boldsymbol{z})]}{\mathrm{Var}(x_i(\boldsymbol{z}))}\right) \\
&= \beta_i
\end{aligned}$$

□

Next, we derive the expectation of the ERL estimator under a general (non-linear) response assumption.

Theorem 6.1. *Let the potential outcome functions be arbitrary functions of the exposures:* $Y_i(\boldsymbol{z}) = Y_i(x_i)$. *Then, the expectation of the* ERL *estimator is*

$$\mathbb{E}[\hat{\tau}] = \frac{1}{n}\sum_{i=1}^{n} \tilde{\beta}_i \ ,$$

where $\tilde{\beta}_i$ *is the coefficient of the exposure* x_i *in a unit-specific OLS regression of* Y_i *on* x_i*:* $\tilde{\beta}_i = \frac{\mathrm{Cov}(x_i,Y_i)}{\mathrm{Var}(x_i)}$.

Proof. We begin by deriving the expectation of an individual term in the ERL estimator. To this end, observe that

$$\mathbb{E}[\hat{\tau}_i] = s\mathbb{E}\Big[Y_i\Big(\frac{x_i-\mathbb{E}[x_i]}{\mathrm{Var}(x_i)}\Big)\Big] = \frac{\mathbb{E}[Y_i x_i]-\mathbb{E}[Y_i]\mathbb{E}[x_i]}{\mathrm{Var}(x_i)} = \frac{\mathrm{Cov}(x_i,Y_i)}{\mathrm{Var}(x_i)} \ .$$

The proof is completed by linearity of expectation. □

We remark that Theorem 4.1 follows from Theorem 6.1 by observing that under a linear response assumption that $Y_i = \beta_i x_i + \alpha_i$, we have that $\mathrm{Cov}(x_i, Y_i) = \mathrm{Cov}(x_i, \beta_i x_i + \alpha_i) = \beta_i \,\mathrm{Var}(x_i)$.

A.2. Consistency of ERL *estimator (Theorem 4.2)*

We are now ready to establish the consistency of the ERL estimator. Before doing so, we restate the theorem here.

Theorem 4.2. *Under Assumptions 1 and 2, the mean squared error of the* ERL *estimator is bounded as* $\mathbb{E}[(\hat{\tau}-\tau)^2] = \mathcal{O}\big(d_d d_o^3/n\big)$. *Thus, the estimator is consistent if* $d_d d_o^3 = o(n)$.

Proof. We begin by proving a finite sample bound on the mean squared error of the ERL estimator, and then we finish the proof by taking the limit in the asymptotic sequence. Note that the mean squared error may be broken down into the errors of the individual treatment effect estimates via

$$\mathbb{E}[(\tau - \hat{\tau})^2] = \mathbb{E}\Big[\Big(\frac{1}{n}\sum_{i=1}^{n}(\tau_i - \hat{\tau}_i)\Big)^2\Big] = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}[(\tau_i - \hat{\tau}_i)(\tau_j - \hat{\tau}_j)] \ .$$

Note that the term in the inner sum is the covariance of the errors in the individual treatment effect estimators. By definition of the dependency neighborhoods, only terms $j \in \mathcal{I}(i)$ are dependent and so only these terms will have non-zero covariance. Using this and the second moment bound in Lemma A.3, we have that

$$\begin{aligned}
\mathbb{E}[(\tau - \hat{\tau})^2] &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\in\mathcal{I}(i)}\mathbb{E}[(\tau_i - \hat{\tau}_i)(\tau_j - \hat{\tau}_j)] && \text{(dep. neighborhoods)}\\
&\le \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\in\mathcal{I}(i)}\sqrt{\mathbb{E}[(\tau_i - \hat{\tau}_i)^2]\mathbb{E}[(\tau_j - \hat{\tau}_j)^2]} && \text{(Cauchy-Schwarz)}\\
&\le \frac{1}{n^2}\sum_{i=1}^{n}|\mathcal{I}(i)|\cdot\Big[M\Big(2 + \frac{d_o}{p(1-p)}\Big)\Big]^2 && \text{(Lemma A.3)}\\
&\le \frac{D}{n}\Big[M\Big(2 + \frac{d_o}{p(1-p)}\Big)\Big]^2 \ . && \text{(max dep. degree)}
\end{aligned}$$

By using the bound $D \le kd_dd_o$ given in Lemma A.1, we have the finite-sample bound on the mean squared error:

$$\mathbb{E}[(\tau - \hat{\tau})^2] \le \frac{kd_dd_o}{n}\Big[M\Big(2 + \frac{d_o}{p(1-p)}\Big)\Big]^2 = \mathcal{O}\big(d_dd_o^3/n\big) \ ,$$

where the final equality comes from interpreting the finite-sample bound in the context of the asymptotic sequence. In particular, by Assumptions 1 and 2, we have that M is a constant, p is bounded away from 0 and 1 by a constant, and k is a constant. □

A.3. Asymptotic normality of the ERL *estimator (Theorem 4.3)*

We establish asymptotic normality of the ERL estimator by using Stein's method. In particular, we use the following result from Ross [52]:

Lemma A.4 (Lemma 3.6 of Ross [52])**.** *Let $a_1, a_2, \dots a_n$ be random variables such that $\mathbb{E}[a_i^4] < \infty$, $\mathbb{E}[a_i] = 0$, $\sigma^2 = \mathrm{Var}(\frac{1}{n}\sum_{i=1}^n a_i)$, and define $X = (\frac{1}{n}\sum_{i=1}^n a_i)/\sigma$. Then for a standard normal $Z \sim \mathcal{N}(0,1)$, we have*

$$d_W(X, Z) \le \frac{D^2}{\sigma^3 n^3}\sum_{i=1}^{n}\mathbb{E}[|a_i|^3] + \sqrt{\frac{28}{\pi}}\cdot\frac{D^{3/2}}{n^2\sigma^2}\sqrt{\sum_{i=1}^{n}\mathbb{E}[a_i^4]} \ ,$$

where D is the maximum dependency degree of the random variables and $d_W(\cdot,\cdot)$ is the Wasserstein distance.

We will use Lemma A.4 to prove asymptotic normality of the ERL estimator. Before continuing, let us restate the theorem.

Theorem 4.3. *Under Assumptions 1, 2, and 3, and supposing that $d_d^4 d_o^{10} = o(n)$, the* ERL *estimator is asymptotically normal:*

$$\frac{\hat{\tau} - \tau}{\sqrt{\mathrm{Var}(\hat{\tau})}} \xrightarrow{d} \mathcal{N}(0,1) \ .$$

Proof. Our strategy may be described in two main steps: first, we use Lemma A.4 to derive a finite-sample bound on the Wasserstein distance between the distribution of $(\tau - \hat{\tau})/\sqrt{\mathrm{Var}(\hat{\tau})}$ and a standard normal. Next, we us this bound to argue that this Wasserstein distance approaches 0 in the limit of the asymptotic sequence under the above conditions.

We seek to apply Lemma A.4 where the random variables are the errors of the individual treatment effect estimates; that is,

$$a_1 = \tau_1 - \hat{\tau}_1, a_2 = \tau_2 - \hat{\tau}_2, \ldots a_n = \tau_n - \hat{\tau}_n \ .$$

Note that $\frac{1}{n}\sum_{i=1}^n a_i = \frac{1}{n}\sum_{i=1}^n \tau_i - \hat{\tau}_i = \tau - \hat{\tau}$ and $\mathrm{Var}(\frac{1}{n}\sum_{i=1}^n a_i) = \mathrm{Var}(\hat{\tau})$ so that the random variable X in Lemma A.4 is equal to $(\tau - \hat{\tau})/\sqrt{\mathrm{Var}(\hat{\tau})}$, which is indeed the random variable we wish to characterize. Let us show that the conditions of Lemma A.4 are satisfied: first, recall that $\hat{\tau}_i$ are unbiased estimates of τ_i so that a_i has mean zero. Second, because the potential outcomes are bounded by a constant M, the support of a_i is bounded so the fourth moments are finite. Thus, we may apply Lemma A.4 in this setting.

We will use the Lemma A.3 to bound the sum of the third and fourth moments. In particular, Lemma A.3 implies that

$$\sum_{i=1}^n \mathbb{E}[|a_i|^3] \le n \cdot \left[M\left(2 + \frac{d_o}{p(1-p)}\right)\right]^3 \text{ and } \sum_{i=1}^n \mathbb{E}[|a_i|^4] \le n \cdot \left[M\left(2 + \frac{d_o}{p(1-p)}\right)\right]^4 .$$

Using this moment bound together with the bound on the maximum dependence degree D (Lemma A.1) on the result of Lemma A.4, we obtain that

$$d_W\left(\frac{\tau - \hat{\tau}}{\sqrt{\mathrm{Var}(\hat{\tau})}}, Z\right) \le \frac{(kd_d d_o)^2}{\sigma^3 n^2} \cdot \left[M\left(2 + \frac{d_o}{p(1-p)}\right)\right]^3 + \sqrt{\frac{28}{\pi}} \cdot \frac{(kd_d d_o)^{3/2}}{\sigma^2 n^{3/2}} \cdot \left[M\left(2 + \frac{d_o}{p(1-p)}\right)\right]^2 .$$

We now interpret this finite-sample bound in the context of the asymptotic sequence. By Assumptions 1 and 2, we have that M is a constant, p is bounded

away from 0 and 1 by a constant, and k is a constant. It follows that the Wasserstein distance between $(\tau-\hat{\tau})/\sqrt{\mathrm{Var}(\hat{\tau})}$ and a standard normal is asymptotically bounded as

$$d_W\Big(\frac{\tau-\hat{\tau}}{\sqrt{\mathrm{Var}(\hat{\tau})}}, Z\Big) = \mathcal{O}\Big(\frac{d_d^2 d_o^5}{\sigma^3 n^2} + \frac{d_d^{3/2} d_o^{7/2}}{\sigma^2 n^{3/2}}\Big)$$

By Assumption 3, we have that $\mathrm{Var}(\hat{\tau}) = \Omega(1/n)$, which means that this bound becomes

$$d_W\Big(\frac{\tau-\hat{\tau}}{\sqrt{\mathrm{Var}(\hat{\tau})}}, Z\Big) = \mathcal{O}\Big(\frac{d_d^2 d_o^5}{n^{1/2}} + \frac{d_d^{3/2} d_o^{7/2}}{n^{1/2}}\Big) = \mathcal{O}\Big(\frac{d_d^2 d_o^5}{n^{1/2}}\Big) \ .$$

By assumption, the asymptotic sequence satisfies $d_d^4 d_o^{10} = o(n)$. Thus, the Wasserstein distance between $(\tau-\hat{\tau})/\sqrt{\mathrm{Var}(\hat{\tau})}$ and a standard normal approaches 0 in this asymptotic sequence. □

Appendix B: Variance estimation and confidence intervals

In this section, we present the proofs of unbiasedness and consistency of the variance estimator together with a proof of the asymptotic validity of the normal based confidence intervals.

B.1. Closed form expressions for coefficients

We begin by deriving closed form expressions for the coefficients $a_{i,j}$, $b_{i,j}$, and $c_{i,j}$ which are obtained as solutions to the system of linear equations:

$$\Sigma_{i,j}\begin{bmatrix} a_{i,j} \\ b_{i,j} \\ c_{i,j}\end{bmatrix} = \begin{bmatrix} \mathrm{Var}(x_i x_j) & \mathrm{Cov}(x_i, x_i x_j) & \mathrm{Cov}(x_j, x_i x_j) \\ \mathrm{Cov}(x_i x_j, x_i) & \mathrm{Var}(x_i) & \mathrm{Cov}(x_j, x_i) \\ \mathrm{Cov}(x_i x_j, x_j) & \mathrm{Cov}(x_i, x_j) & \mathrm{Var}(x_j)\end{bmatrix}\begin{bmatrix} a_{i,j} \\ b_{i,j} \\ c_{i,j}\end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0\end{bmatrix} \ .$$

Distinct outcome units $(i \neq j)$ Note that the matrix in the linear system above is the 3-by-3 covariance matrix of the exposures x_i, x_j and their product $x_i x_j$. Recall that we have defined this matrix to be $\Sigma_{i,j}$. A unique solution exists if and only if $\det(\Sigma_{i,j}) > 0$, which is the say that the generalized variance of x_i, x_j, and $x_i x_j$ is nonzero. In this case, the solution coefficients may be explicitly derived via Cramer's rule for solving linear systems as

$$\begin{aligned} a_{i,j} &= \frac{\mathrm{Var}(x_i)\,\mathrm{Var}(x_j) - \mathrm{Cov}(x_i, x_j)^2}{\det(\Sigma_{i,j})}, \\ b_{i,j} &= \frac{\mathrm{Cov}(x_i, x_j)\,\mathrm{Cov}(x_i x_j, x_j) - \mathrm{Var}(x_j)\,\mathrm{Cov}(x_i x_j, x_i)}{\det(\Sigma_{i,j})}, \\ c_{i,j} &= \frac{\mathrm{Cov}(x_i, x_j)\,\mathrm{Cov}(x_i x_j, x_i) - \mathrm{Var}(x_i)\,\mathrm{Cov}(x_i x_j, x_j)}{\det(\Sigma_{i,j})} \ . \end{aligned}$$

The determinant $\det(\Sigma_{i,j})$ is a polynomial in the entries of the matrix $\Sigma_{i,j}$. For completeness, we present the determinant calculation:

$$\begin{aligned}\det(\Sigma_{i,j}) = \operatorname{Var}(x_i x_j)(\operatorname{Var}(x_i)\operatorname{Var}(x_j) - \operatorname{Cov}(x_i, x_j)^2) - \operatorname{Var}(x_i)\operatorname{Cov}(x_i x_j, x_j)^2 \\ - \operatorname{Var}(x_j)\operatorname{Cov}(x_i x_j, x_i)^2 + 2\operatorname{Cov}(x_i, x_j)\operatorname{Cov}(x_i x_j, x_j)\operatorname{Cov}(x_i x_j, x_i) \quad .\end{aligned}$$

Same outcome unit ($i = j$) Note that the 3-by-3 covariance matrix defining the system of linear equation above will have zero determinant, as $x_i = x_j$. Nevertheless, we show that a solution to the linear system may still be obtained under certain conditions on the distribution of the exposure x_i. Observe that the $b_{i,i}$ and $c_{i,i}$ terms are redundant, as $x_i = x_j$. By taking $c_{i,i} = 0$, we reduce the 3-by-3 linear system in Proposition B.1 to the following 2-by-2 linear system:

$$\begin{bmatrix} \operatorname{Var}(x_i^2) & \operatorname{Cov}(x_i, x_i^2) \\ \operatorname{Cov}(x_i^2, x_i) & \operatorname{Var}(x_i) \end{bmatrix} \begin{bmatrix} a_{i,i} \\ b_{i,i} \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

Observe that the matrix in this linear system is the covariance matrix between the exposure x_i and its square x_i^2. Recall that in the main body, we defined this 2-by-2 matrix to be $\Sigma_{i,i}$. Note that a solution exists when the $\det(\Sigma_{i,i}) > 0$, which is to say that the exposure x_i and its square x_i^2 are not perfectly correlated. In this case, the solution to the coefficients may be obtained as

$$a_{i,i} = \frac{\operatorname{Var}(x_i)}{\det(\Sigma_{i,i})}, \quad b_{i,i} = \frac{\operatorname{Cov}(x_i, x_i^2)}{\det(\Sigma_{i,i})}, \quad c_{i,i} = 0$$

where $\det(\Sigma_{i,i}) = \operatorname{Var}(x_i)\operatorname{Var}(x_i^2) - \operatorname{Cov}(x_i, x_i^2)^2$.

B.2. Unbiasedness of variance estimator (Theorem 5.1)

In this section, we present proofs of Proposition B.1 and Theorem 5.1, which establish unbiasedness of the proposed variance estimator under certain conditions on the exposure distribution.

Proposition B.1. *Under the linear response assumption, $\widehat{C}_{i,j} = Y_i(\boldsymbol{z})Y_j(\boldsymbol{z}) \times R_{i,j}(x_i, x_j)$ is an unbiased estimator for the individual covariance term $\operatorname{Cov}(\hat{\tau}_i, \hat{\tau}_j)$ if the coefficients $a_{i,j}, b_{i,j}, c_{i,j}$ in the weighting function $S_{i,j}(x_i, x_j)$ satisfy the system of linear equations:*

$$\Sigma_{i,j} \begin{bmatrix} a_{i,j} \\ b_{i,j} \\ c_{i,j} \end{bmatrix} = \begin{bmatrix} \operatorname{Var}(x_i x_j) & \operatorname{Cov}(x_i, x_i x_j) & \operatorname{Cov}(x_j, x_i x_j) \\ \operatorname{Cov}(x_i x_j, x_i) & \operatorname{Var}(x_i) & \operatorname{Cov}(x_j, x_i) \\ \operatorname{Cov}(x_i x_j, x_j) & \operatorname{Cov}(x_i, x_j) & \operatorname{Var}(x_j) \end{bmatrix} \begin{bmatrix} a_{i,j} \\ b_{i,j} \\ c_{i,j} \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \quad .$$

Proof. If $\operatorname{Cov}(x_i, x_j) = 0$, then by the linear response assumption $\operatorname{Cov}(\hat{\tau}_i, \hat{\tau}_j) = 0$. In this case, the weighting function is identically zero and thus $\widehat{C}_{i,j} = 0$, which is trivially unbiased. The remainder of the proof focuses on the case where $\operatorname{Cov}(x_i, x_j) \neq 0$ and the weighting function $R_{i,j}(x_i, x_j) = Q_{i,j}(x_i, x_j) -$

$S_{i,j}(x_i, x_j)$. Observe that the expectation of the individual covariance estimator $\widehat{C}_{i,j}$ is equal to

$$\begin{aligned}\mathbb{E}\Big[\widehat{C}_{i,j}\Big] &= \mathbb{E}\Big[Y_i(\boldsymbol{z})Y_j(\boldsymbol{z})R_{i,j}(x_i,x_j)\Big] \\ &= \mathbb{E}\Big[Y_i(\boldsymbol{z})Y_j(\boldsymbol{z})Q_{i,j}(x_i,x_j)\Big] - \mathbb{E}\Big[Y_i(\boldsymbol{z})Y_j(\boldsymbol{z})S_{i,j}(x_i,x_j)\Big] \ .\end{aligned} \tag{3}$$

By construction of the weighting function $Q_{i,j}(x_i, x_j)$, we can compute the expectation of the first term in (3) as

$$\mathbb{E}\Big[Y_i(\boldsymbol{z})Y_j(\boldsymbol{z})Q_{i,j}(x_i,x_j)\Big] = \mathbb{E}\left[Y_i(\boldsymbol{z})Y_j(\boldsymbol{z})\Big(\frac{x_i - \mathbb{E}[x_i]}{\mathrm{Var}(x_i)}\Big)\Big(\frac{x_j - \mathbb{E}[x_j]}{\mathrm{Var}(x_j)}\Big)\right] = \mathbb{E}\big[\hat{\tau}_i\hat{\tau}_j\big] \ .$$

Next, we evaluate the expectation of the second term in (3). Before doing so, observe that $\mathbb{E}[S_{i,j}(x_i, x_j)] = 0$ by construction. Moreover, the coefficients used in the $S_{i,j}(x_i, x_j)$ weighting function satisfy the system of linear equations by assumption, which is equivalent to the following three equations:

- $\mathbb{E}[x_i S_{i,j}(x_i, x_j)] = 0$
- $\mathbb{E}[x_j S_{i,j}(x_j, x_j)] = 0$
- $\mathbb{E}[x_i x_j S_{i,j}(x_i, x_j)] = 1$

Using these four equations together with the linear response assumption, we evaluate the expectation of the second term in (3) as

$$\begin{aligned}&\mathbb{E}\Big[Y_i(\boldsymbol{z})Y_j(\boldsymbol{z})S_{i,j}(x_i,x_j)\Big] \\ &\quad= \mathbb{E}\Big[(\beta_i x_i + \alpha_i)(\beta_j x_j + \alpha_j)S_{i,j}(x_i,x_j)\Big] \\ &\quad= \beta_i\beta_j\mathbb{E}\Big[x_i x_j S_{i,j}(x_i,x_j)\Big] + \alpha_i\alpha_j\mathbb{E}\Big[S_{i,j}(x_i,x_j)\Big] \\ &\qquad+ \beta_i\alpha_j\mathbb{E}\Big[x_i S_{i,j}(x_i,x_j)\Big] + \alpha_i\beta_j\mathbb{E}\Big[x_j S_{i,j}(x_i,x_j)\Big] \\ &\quad= \beta_i\beta_j \\ &\quad= \mathbb{E}[\hat{\tau}_i]\mathbb{E}[\hat{\tau}_j] \ ,\end{aligned}$$

where the last inequality follows from the unbiasedness of the individual treatment effect estimators $\hat{\tau}_i$ and $\hat{\tau}_j$. Thus, substituting these two calculations into (3) yields the desired result:

$$\mathbb{E}\Big[\widehat{C}_{i,j}\Big] = \mathbb{E}\big[\hat{\tau}_i\hat{\tau}_j\big] - \mathbb{E}[\hat{\tau}_i]\mathbb{E}[\hat{\tau}_j] = \mathrm{Cov}(\hat{\tau}_i, \hat{\tau}_j) \qquad \square$$

We are now ready to prove Theorem 5.1, which establishes unbiasedness of the variance estimator.

Theorem 5.1. *Under Assumption 4 and the linear response assumption, the variance estimator of the* ERL *point estimator is unbiased, i.e.* $\mathbb{E}[\widehat{\mathrm{Var}}(\hat{\tau})] = \mathrm{Var}(\hat{\tau})$.

Proof. Note that Assumption 4, together with the linear response assumption, ensure that the conditions of Proposition B.1 hold for every pair $i, j \in [n]$ so that $\mathbb{E}[\widehat{C}_{i,j}] = \mathrm{Cov}(\hat{\tau}_i, \hat{\tau}_j)$. Using this fact, we may calculate the expectation of the variance estimate $\widehat{\mathrm{Var}}(\hat{\tau})$ as

$$\begin{aligned}\mathbb{E}[\widehat{\mathrm{Var}}(\hat{\tau})] &= \mathbb{E}\left[\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} Y_i(\boldsymbol{z})Y_j(\boldsymbol{z})R_{i,j}(x_i, x_j)\right] \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}\Big[Y_i(\boldsymbol{z})Y_j(\boldsymbol{z})R_{i,j}(x_i, x_j)\Big] \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{Cov}(\hat{\tau}_i, \hat{\tau}_j) = \mathrm{Var}(\hat{\tau}) \ . \end{aligned}$$

□

B.3. Consistency of variance estimator (Theorem 5.3)

In this section, we present the proof of Theorem 5.3, which establishes consistency of the normalized variance estimator. The main parts of the proof are to show that the individual covariance estimators $\widehat{C}_{i,j}$ are sufficiently uncorrelated and that they have small variance.

To establish a bound on the correlation between covariance estimators $\widehat{C}_{i,j}$, we use the formalism of the dependency graph. For each pair of outcome units $i, j \in [n]$, define $a_{i,j} \triangleq \mathrm{Cov}(\hat{\tau}_i, \hat{\tau}_j) - \widehat{C}_{i,j}$ to be the error of the individual covariance estimator. Let $\mathcal{A}_V = \{a_{i,j} : i, j \in [n]\}$ be the set of individual covariance estimator errors. For each variable $a_{i,j}$, we define the *dependency neighborhood* as

$$\mathcal{I}_V(i,j) \subset \mathcal{A}_V \text{ such that } a_{i,j} \text{ is jointly independent of the variables } \mathcal{A}_V \backslash \mathcal{I}_V(i,j)\,.$$

Unlike the dependency graph in Appendix A, this dependency graph is indexed over pairs of integers $i, j \in [n]$. Additionally, quantities associated with this dependency graph are denoted by a subscript V.

The following lemma bounds the maximum degree of the dependency graph and is based on a similar counting argument as that used to prove Lemma A.1.

Lemma B.2. *The dependency degree of individual covariance estimator errors is at most* $D_V \leq 4(kd_d d_o)^2$.

Proof. The first part of this proof is to establish a necessary condition for an individual covariance error $a_{k,\ell}$ to be in the dependency neighborhood of $a_{i,j}$, i.e. $a_{k,\ell} \in \mathcal{I}_V(i,j)$. We begin by re-writing the exposures under a cluster design. Recall that the exposures are defined as $x_i = \sum_{j=1}^{m} w_{i,j} z_j$. For each cluster $C \in \mathcal{C}$, define $w_{i,C} = \sum_{j \in C} w_{i,j}$ and define z_C to be the ± 1 cluster treatment assignment variable which is 1 if diversion units in C are treated and -1 otherwise. If $w_{i,C} \neq 0$, then we say that cluster C is *incident* to outcome unit i. Define $S(i) = \{z_C : w_{i,C} \neq 0\}$ to be the cluster treatment assignments which

influence the exposure x_i. Under the cluster design, the exposure for outcome unit i may be written as

$$x_i = \sum_{C \in \mathcal{C}} w_{i,C} z_C = \sum_{C \in S(i)} w_{i,C} z_C \ .$$

By the linear exposure-response assumption, the individual covariance estimator error $a_{i,j}$ is a function of the exposures x_i and x_j. Moreover, $a_{i,j}$ is a function of the cluster treatment assignment variables in $S(i,j) \triangleq S(i) \cup S(i)$. Let us denote this relationship by writing $a_{i,j} = g_{i,j}(S(i,j))$, where $g_{i,j}$ is a function of the cluster treatment variables $z_C \in S(i,j)$. Let $B \subset V_o \times V_o$ be a collection of pairs of outcome units. We remark that joint independence of cluster treatment assignments implies joint independence of individual covariance errors:

$$S(i,j) \perp\!\!\!\perp \{S(k,\ell) : (k,\ell) \in B\} \Rightarrow a_{i,j} \perp\!\!\!\perp \{a_{k,\ell} : (k,\ell) \in B\} \ .$$

Under an independent cluster design, the cluster treatment assignments $S(i,j)$ are jointly independent of the cluster treatment assignments $\{S(k,\ell) : (k,\ell) \in B\}$ when the corresponding sets of clusters are disjoint, i.e. $S(i,j) \cap (\cup_{(k,\ell)\in B} S(k,\ell)) = \emptyset$. Thus, the individual covariance estimate $a_{i,j}$ is jointly independent of the collection of individual treatment effect estimates $\{a_{k,\ell} : (k,\ell) \in B\}$ when outcome units i and j are not incident to any cluster that is incident to an outcome unit in B. In other words, $a_{k,\ell} \in \mathcal{I}_V(j,k)$ only if one of the outcome units i,j and one of the outcome units k,ℓ are incident to a common cluster.

Fix a pair of outcome units $i,j \in V_o$. The remainder of the proof is a simple counting argument which uses this necessary condition to establish that $|\mathcal{I}_V(i,j)| \leq (2kd_d d_o)^2$. In particular, we will count the number of outcome units that are incident to one of the clusters that are incident to i and j. Because the degrees of outcome unit i and j are at most d_o, they are incident to at most $2d_o$ clusters. Each of these clusters has at most k diversion units, by Assumption 2. Because the degree of all diversion units is at most d_d, the number of pairs of outcome units which are incident to at least one of these clusters is at most

$$\binom{2kd_d d_o}{2} \leq (2kd_d d_o)^2 \ .$$

Thus, we have established that

$$D = \max_{i,j \in V_o} |\mathcal{I}_V(i,j)| \leq 4(kd_d d_o)^2 \ . \qquad \square$$

Next, we derive a bound on the variance of the individual covariance estimators. Because these individual estimators are unbiased, this yields a bound on their mean squared error. Recall that $\Delta = \min_{i,j\in[n]} \det(\Sigma_{i,j})$ is defined to be the smallest non-degeneracy measure amongst all distinct pairs ($i \neq j$) of exposures and single exposures ($i = j$). We begin by establishing that the coefficients of the weighting function are bounded in magnitude.

Lemma B.3. *For each pair of outcome units $i, j \in [n]$, the absolute values of the coefficients in the weighting function $S_{i,j}(x_i, x_j)$ are bounded in magnitude as* $\max\{|a_{i,j}|, |b_{i,j}|, |c_{i,j}|\} \leq 2/\Delta$.

Proof. Recall that the coefficients $a_{i,j}$, $b_{i,j}$, and $c_{i,j}$ in the weighting function $S_{i,j}(x_i, x_j)$ are of the form

$$a_{i,j} = \frac{\tilde{a}_{i,j}}{\det(\Sigma_{i,j})}, b_{i,j} = \frac{\tilde{b}_{i,j}}{\det(\Sigma_{i,j})}, \text{ and } c_{i,j} = \frac{\tilde{c}_{i,j}}{\det(\Sigma_{i,j})} \ ,$$

where $\tilde{a}_{i,j}$, $\tilde{b}_{i,j}$, and $\tilde{c}_{i,j}$ depend on statistics of the joint distribution of the exposures x_i and x_j. We will now show that $\max\{|\tilde{a}_{i,j}|, |\tilde{b}_{i,j}|, |\tilde{c}_{i,j}|\} \leq 2$. We focus only on the case of distinct exposures ($i \neq j$), as the case of a single exposure ($i = j$) follows in an identical way. First, observe that

$$\tilde{a}_{i,j} = \operatorname{Var}(x_i)\operatorname{Var}(x_j) - \operatorname{Cov}(x_i, x_j)^2 \leq \operatorname{Var}(x_i)\operatorname{Var}(x_j) \leq 1 \ .$$

Similarly using the triangle inequality, Cauchy-Schwarz inequality, and the fact that the exposures are supported on $[-1, 1]$, we have that

$$\begin{aligned}
|\tilde{b}_{i,j}| &= |\operatorname{Cov}(x_i, x_j)\operatorname{Cov}(x_i x_j, x_j) - \operatorname{Var}(x_j)\operatorname{Cov}(x_i x_j, x_i)| \\
&\leq |\operatorname{Cov}(x_i, x_j)\operatorname{Cov}(x_i x_j, x_j)| + |\operatorname{Var}(x_j)\operatorname{Cov}(x_i x_j, x_i)| \\
&\leq \sqrt{\operatorname{Var}(x_i)\operatorname{Var}(x_j)\operatorname{Var}(x_i x_j)\operatorname{Var}(x_j)} + \operatorname{Var}(x_j)\sqrt{\operatorname{Var}(x_i x_j)\operatorname{Var}(x_i)} \\
&= 2\operatorname{Var}(x_j)\sqrt{\operatorname{Var}(x_i)\operatorname{Var}(x_i x_j)} \\
&\leq 2 \ ,
\end{aligned}$$

where the last inequality follows from the fact that random variables on $[0, 1]$ have variance at most 1. The bound on $|\tilde{c}_{i,j}|$ is identical. This establishes that $\max\{|\tilde{a}_{i,j}|, |\tilde{b}_{i,j}|, |\tilde{c}_{i,j}|\} \leq 2$. Thus, we have that $\max\{|a_{i,j}|, |b_{i,j}|, |c_{i,j}|\} \leq 2/\det(\Sigma_{i,j}) \leq 2/\Delta$. □

Using Lemma B.3 together with previously proved lower bounds on the variance of an exposure (Lemma A.2), we obtain the following bound on the variance of the individual covariance estimators:

Lemma B.4. *The variance of an individual covariance estimator is bounded by*

$$\operatorname{Var}(\widehat{C}_{i,j}) \leq CM^4\left(\left(\frac{d_o}{p(1-p)}\right)^4 + \frac{1}{\Delta^2}\right)$$

for some absolute constant C.

Proof. Recall that for two random variables X and Y, we have the following inequality: $\operatorname{Var}(X+Y) \leq (\sqrt{\operatorname{Var}(X)} + \sqrt{\operatorname{Var}(Y)})^2$. Applying this inequality to the individual covariance estimator, we obtain

$$\operatorname{Var}(\widehat{C}_{i,j}) = \operatorname{Var}\Big(Y_i(\boldsymbol{z})Y_j(\boldsymbol{z})R_{i,j}(x_i, x_j)\Big)$$

$$= \mathrm{Var}\Big(Y_i(\boldsymbol{z})Y_j(\boldsymbol{z})Q_{i,j}(x_i,x_j) - Y_i(\boldsymbol{z})Y_j(\boldsymbol{z})S_{i,j}(x_i,x_j)\Big)$$
$$\leq \left(\sqrt{\mathrm{Var}\Big(Y_i(\boldsymbol{z})Y_j(\boldsymbol{z})Q_{i,j}(x_i,x_j)\Big)} + \sqrt{\mathrm{Var}\Big(Y_i(\boldsymbol{z})Y_j(\boldsymbol{z})S_{i,j}(x_i,x_j)\Big)}\right)^2$$

Our next goal is to bound each of the terms appearing above. The variance in the first term may be bounded as

$$\begin{aligned}
&\mathrm{Var}(Y_i(\boldsymbol{z})Y_j(\boldsymbol{z})Q_{i,j}(x_i,x_j)) \\
&= \mathrm{Var}\left(Y_i(\boldsymbol{z})Y_j(\boldsymbol{z})\Big(\frac{x_i - \mathbb{E}[x_i]}{\mathrm{Var}(x_i)}\Big)\Big(\frac{x_j - \mathbb{E}[x_j]}{\mathrm{Var}(x_j)}\Big)\right) \\
&= \frac{1}{\big(\mathrm{Var}(x_i)\,\mathrm{Var}(x_j)\big)^2}\,\mathrm{Var}(Y_i(\boldsymbol{z})Y_j(\boldsymbol{z})(x_i - \mathbb{E}[x_i])(x_j - \mathbb{E}[x_j])) \\
&\leq \Big(\frac{d_o}{p(1-p)}\Big)^4 \mathrm{Var}(Y_i(\boldsymbol{z})Y_j(\boldsymbol{z})(x_i - \mathbb{E}[x_i])(x_j - \mathbb{E}[x_j])) \\
&\leq \Big(\frac{d_o}{p(1-p)}\Big)^4 \cdot \big(2M^2\big)^2
\end{aligned}$$

,

where the first inequality follows from using Lemma A.2 and the second inequality follows from the bound $|Y_i(\boldsymbol{z})Y_j(\boldsymbol{z})(x_i - \mathbb{E}[x_i])(x_j - \mathbb{E}[x_j])| \leq M^2$. to lower bound the variance of the exposures and the upper bound on the potential outcomes, i.e. $|Y_i(\boldsymbol{z})| \leq M$ for all $\boldsymbol{z} \in \{0,1\}^n$.

We now seek to bound the variance appearing in the second term. Note that the magnitude of the term inside the variance may be bounded as

$$\begin{aligned}
&|Y_i(\boldsymbol{z})Y_j(\boldsymbol{z})S_{i,j}(x_i,x_j)| \\
&\leq |Y_i(\boldsymbol{z})Y_j(\boldsymbol{z})| \cdot |S_{i,j}(x_i,x_j)| \\
&= |Y_i(\boldsymbol{z})Y_j(\boldsymbol{z})| \cdot |a_{i,j}(x_ix_j - \mathbb{E}[x_ix_j]) + b_{i,j}(x_i - \mathbb{E}[x_i]) + c_{i,j}(x_j - \mathbb{E}[x_j])| \\
&\leq |Y_i(\boldsymbol{z})Y_j(\boldsymbol{z})| \cdot \big(|a_{i,j}| \cdot |x_ix_j - \mathbb{E}[x_ix_j]| + |b_{i,j}| \cdot |x_i - \mathbb{E}[x_i]| + |c_{i,j}| \cdot |x_j - \mathbb{E}[x_j]|\big) \\
&\leq M^2 \cdot 3 \cdot (2/\Delta) \cdot 2 \\
&= 12\frac{M^2}{\Delta} \quad ,
\end{aligned}$$

where the final inequality follows from the bound on the potential outcomes, the bound on the weighting coefficients given in Lemma B.3, and the fact that the exposures take values in $[0,1]$. Thus, the variance in the second term is at most

$$\mathrm{Var}\Big(Y_i(\boldsymbol{z})Y_j(\boldsymbol{z})S_{i,j}(x_i,x_j)\Big) \leq \Big(24 \cdot \frac{M^2}{\Delta}\Big)^2 \quad .$$

Plugging these two bounds into the bound on $\mathrm{Var}(\widehat{C}_{i,j})$, and using the Arithmetic-Geometric Inequality, we obtain

$$\mathrm{Var}(\widehat{C}_{i,j}) \leq \left(2M^2\Big(\frac{d_o}{p(1-p)}\Big)^2 + 24 \cdot \frac{M^2}{\Delta}\right)^2 \leq CM^4\left(\Big(\frac{d_o}{p(1-p)}\Big)^4 + \frac{1}{\Delta^2}\right) \quad . \ \square$$

Finally, we are ready to prove Theorem 5.3, which establishes consistency rates for the variance estimator. At a high level, we will combine Lemmas B.2 and B.2, which show that the individual covariance estimators are sufficiently uncorrelated and have small variance.

Theorem 5.3. *Under Assumptions 1, 2, and 4, the mean squared error of the normalized variance estimator is bounded as*

$$\mathbb{E}\Big[\big(n \cdot \mathrm{Var}(\hat{\tau}) - n \cdot \widehat{\mathrm{Var}}(\hat{\tau})\big)^2\Big] = \mathcal{O}\Big(\frac{1}{n} \cdot \Big(d_d^3 d_o^7 + \frac{1}{\Delta^2}\Big)\Big) \ .$$

Thus, the normalized variance estimator is consistent if $d_d^3 d_o^7 = o(n)$ *and* $\Delta = \omega(n^{-1/2})$.

Proof. By unbiasesness of the variance estimator together a decomposition of its variance, we have that

$$\begin{aligned}
&\mathbb{E}\Big[\big(n \cdot \mathrm{Var}(\hat{\tau}) - n \cdot \widehat{\mathrm{Var}}(\hat{\tau})\big)^2\Big] \\
&= n^2 \cdot \mathrm{Var}\Big(\widehat{\mathrm{Var}}(\hat{\tau})\Big) = n^2 \cdot \mathrm{Var}\left(\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\widehat{C}_{i,j}\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{\ell=1}^{n}\mathrm{Cov}(\widehat{C}_{i,j},\widehat{C}_{k,\ell}) \ . \qquad (4)
\end{aligned}$$

We now discuss which terms are zero in the sum. Recall that when $\mathrm{Cov}(x_i, x_j) = 0$, then $\widehat{C}_{i,j} = 0$. Moreover, $\mathrm{Cov}(x_i, x_j) = 0$ for all $j \notin \mathcal{I}(i)$. Thus, $\mathrm{Cov}(\widehat{C}_{i,j}, \widehat{C}_{k,\ell}) = 0$ for any $j \notin \mathcal{I}(i)$. Additionally, the individual covariance estimators $\widehat{C}_{i,j}$ and $\widehat{C}_{k,\ell}$ are uncorrelated if $(k,l) \notin \mathcal{I}_V(i,j)$. Thus, we may simplify terms as

$$= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\in\mathcal{I}(i)}\sum_{(k,\ell)\in\mathcal{I}_V(i,j)}\mathrm{Cov}(\widehat{C}_{i,j},\widehat{C}_{k,\ell}) \ . \qquad (5)$$

Next, we use Cauchy-Schwarz inequality on the covariances together with the upper bound the variances $\mathrm{Var}(\widehat{C}_{i,j})$ provided by Lemma B.4 to obtain

$$\begin{aligned}
&\leq \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\in\mathcal{I}(i)}\sum_{(k,\ell)\in\mathcal{I}_V(i,j)}\sqrt{\mathrm{Var}(\widehat{C}_{i,j})\,\mathrm{Var}(\widehat{C}_{k,\ell})} \\
&\leq \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\in\mathcal{I}(i)}\sum_{(k,\ell)\in\mathcal{I}_V(i,j)} CM^4\left(\Big(\frac{d_o}{p(1-p)}\Big)^4 + \frac{1}{\Delta^2}\right) \ .
\end{aligned}$$

Finally, we use the maximum dependency degree bounds $\max_{i\in[n]}|\mathcal{I}(i)| \triangleq D \leq kd_d d_o$ and $\max_{i,j\in[n]}|\mathcal{I}_V(i,j)| \triangleq D_V \leq 4(kd_d d_o)^2$ and rearrange terms to obtain

$$\leq \frac{1}{n^2} nDD_V \cdot CM^4\left(\Big(\frac{d_o}{p(1-p)}\Big)^4 + \frac{1}{\Delta^2}\right)$$

$$\leq \frac{4C}{n}(d_d d_o)^3 \cdot M^4 \left(\left(\frac{d_o}{p(1-p)} \right)^4 + \frac{1}{\Delta^2} \right) .$$

Under Assumptions 1 and 2, the magnitude of the potential outcomes M, the size of the clusters k, and the term $1/p(1-p)$ are constants in the asymptotic sequence. Thus, mean squared error of the normalized variance estimator is bounded as

$$\mathbb{E}\Big[\big(n \cdot \mathrm{Var}(\hat{\tau}) - n \cdot \widehat{\mathrm{Var}}(\hat{\tau}) \big)^2 \Big] = \mathcal{O}\Big(\frac{1}{n} \cdot \Big(d_d^3 d_o^7 + \frac{1}{\Delta^2} \Big) \Big) . \qquad \square$$

B.4. Asymptotic validity of confidence intervals (Theorem 4.3)

In this section, we present the proof of Corollary 5.4 which establishes asymptotic validity of the Wald-type confidence intervals using the variance estimator.

Lemma B.5. *Under Assumptions 1, 2, 3, and 4 and further supposing that* $d_o^3 d_d^7 = o(n)$ *and* $\Delta = \omega(n^{-1/2})$*, the ratio of the variance estimator and the true estimator converges to 1 in probability:* $\frac{\mathrm{Var}(\hat{\tau})}{\widehat{\mathrm{Var}}(\hat{\tau})} \xrightarrow{p} 1$.

Proof. This may be shown by applying the continuous mapping theorem to the result of Theorem 5.3, as Assumption 3 bounds the normalized variance away from zero. However, we take a more elementary approach using Chebyshev's inequality.

Let $\epsilon > 0$ be given. Chebyshev's inequality states that $\Pr(|X - \mu| \geq k\sigma)$ for any random variable X with mean μ and standard deviation σ:. For random variables with positive mean, rearranging terms yields $\Pr(|\frac{X}{\mu} - 1| > \epsilon) \leq \frac{\sigma^2}{\epsilon^2 \mu^2}$. Applying Chebyshev's inequality together with the bound on the mean squared error of the variance estimator (Theorem 5.3) and Assumption 3, we have

$$\begin{aligned}
&\Pr\left(\left| \frac{\mathrm{Var}(\hat{\tau})}{\widehat{\mathrm{Var}}(\hat{\tau})} - 1 \right| > \epsilon \right) \\
&\quad \leq \frac{\mathbb{E}\Big[\big(\mathrm{Var}(\hat{\tau}) - \widehat{\mathrm{Var}}(\hat{\tau}) \big)^2 \Big]}{\epsilon^2 \, \mathrm{Var}(\hat{\tau})^2} && \text{(Chebyshev's Inequality)} \\
&\quad = \frac{n^2}{n^2} \cdot \frac{\mathbb{E}\Big[\big(\mathrm{Var}(\hat{\tau}) - \widehat{\mathrm{Var}}(\hat{\tau}) \big)^2 \Big]}{\epsilon^2 \, \mathrm{Var}(\hat{\tau})^2} \\
&\quad = \frac{\mathbb{E}\Big[\big(n \cdot \mathrm{Var}(\hat{\tau}) - n \cdot \widehat{\mathrm{Var}}(\hat{\tau}) \big)^2 \Big]}{\epsilon^2 (n \cdot \mathrm{Var}(\hat{\tau}))^2} \\
&\quad \leq \frac{1}{\epsilon^2} \cdot \mathcal{O}\Big(\frac{1}{n} \cdot \Big(d_d^3 d_o^7 + \frac{1}{\Delta^2} \Big) \Big) && \text{(Theorem 5.3, Assumption 3)} \\
&\quad = \frac{1}{\epsilon} \cdot o(1) \ ,
\end{aligned}$$

where the final inequality follows from the assumptions that $d_o^3 d_d^7 = o(n)$ and $\Delta = \omega(n^{-1/2})$. This establishes that the ratio of the variance estimator and the true estimator converges to 1 in probability. □

We are now ready to prove Corollary 5.4, which establishes asymptotic validity of the Wald-based confidence intervals using the proposed variance estimator. For completeness, we restate the corollary below.

Corollary 5.4. *Under Assumptions 1-4 and further supposing that* $d_o^4 d_d^{10} = o(n)$ *and* $\Delta = \omega(n^{-1/2})$*, the Wald-type confidence interval using the proposed variance estimator is asymptotically valid:*

$$\lim_{n\to\infty} \Pr\Big(\tau \in \Big[\hat{\tau} \pm \Phi^{-1}(1-\alpha/2)\sqrt{\widehat{\mathrm{Var}}(\hat{\tau})}\Big]\Big) = 1-\alpha \ .$$

Proof. Define the random variable $Z = \frac{\tau-\hat{\tau}}{\sqrt{\mathrm{Var}(\hat{\tau})}}$. By Theorem 4.3, Z converges in distribution to a standard normal, $Z \xrightarrow{d} \mathcal{N}(0,1)$. Define $Z' = \frac{\tau-\hat{\tau}}{\sqrt{\widehat{\mathrm{Var}}(\hat{\tau})}}$ and observe that

$$Z' = \frac{\tau-\hat{\tau}}{\sqrt{\widehat{\mathrm{Var}}(\hat{\tau})}} = \frac{\tau-\hat{\tau}}{\sqrt{\mathrm{Var}(\hat{\tau})}} \cdot \frac{\sqrt{\mathrm{Var}(\hat{\tau})}}{\sqrt{\widehat{\mathrm{Var}}(\hat{\tau})}} = Z \cdot \sqrt{\frac{\mathrm{Var}(\hat{\tau})}{\widehat{\mathrm{Var}}(\hat{\tau})}}$$

By Lemma B.5, the ratio of the variance and the variance estimator converges to 1 in probability. Thus, by Slutsky's theorem, $Z' \xrightarrow{d} \mathcal{N}(0,1)$.

Now, we evaluate the probability of coverage in the limit. By rearranging terms, we can re-write the coverage probability in terms of the tails of Z' as follows:

$$\begin{aligned}
&\lim_{n\to\infty} \Pr\Big(\tau \in \Big[\hat{\tau} \pm \Phi^{-1}(1-\alpha/2)\sqrt{\widehat{\mathrm{Var}}(\hat{\tau})}\Big]\Big) \\
&\quad = \lim_{n\to\infty} \Pr\Big(\Phi^{-1}(1-\alpha/2) \le \frac{\tau-\hat{\tau}}{\sqrt{\widehat{\mathrm{Var}}(\hat{\tau})}} \le \Phi^{-1}(1-\alpha/2)\Big) \\
&\quad = \lim_{n\to\infty} \Pr\Big(\Phi^{-1}(1-\alpha/2) \le Z' \le \Phi^{-1}(1-\alpha/2)\Big) \\
&\quad = \lim_{n\to\infty} \Pr\Big(\Phi^{-1}(1-\alpha/2) \le Z' \le \Phi^{-1}(\alpha/2)\Big) \ , \qquad (6)
\end{aligned}$$

where the last equality follows from symmetry of the normal distribution. Let F_n be the cumulative distribution function of Z'. By the convergence of Z' in distribution to a standard normal, we have that

$$\begin{aligned}
&= \lim_{n\to\infty} F_n\Big(\Phi^{-1}(1-\alpha/2)\Big) - F_n\Big(\Phi^{-1}(\alpha/2)\Big) \\
&= \Phi\Big(\Phi^{-1}(1-\alpha/2)\Big) - \Phi\Big(\Phi^{-1}(\alpha/2)\Big) \\
&= (1-\alpha/2) - (\alpha/2) \\
&= 1-\alpha \ .
\end{aligned}$$

□

Appendix C: Exposure-Design and correlation clustering

In this section, we prove the relationship between Exposure-Design, its reformulation Corr-Clust, the previously proposed correlation clustering design of [49], and other correlation clustering variants. A summary of the results are:

- In Section C.1, we show that the Exposure-Design may be reformulated as the clustering problem, Corr-Clust.
- In Section C.2, we compare Exposure-Design to the correlation clustering-based design presented in [49]. In particular, we prove that their design is equivalent to Exposure-Design when the trade-off parameter is set as $\phi = 1/(n-1)$ and no constraint is placed on cluster sizes, i.e. $k = m$.
- In Section C.3, we compare Corr-Clust to other correlation clustering variants. In particular, we prove that (unconstrained) Corr-Clust may be viewed as an instance of the weighted maximization correlation clustering considered by [14, 58] but with a possibly large additive constant which prevents an approximation-preserving reduction.

To begin, we demonstrate how to re-write the Corr-Clust objective using matrix notation. Let $\omega_{i,j} \in \mathbb{R}$ be the weights for pairs $i, j \in [m]$ and let Ω be the m-by-m matrix whose (i,j)th entry is $\omega_{i,j}$. For a partition $\mathcal{C}$ of the indices $[m]$, let $Z_{\mathcal{C}}$ be the m-by-m matrix where the (i,j)th entry is 1 if i and j are in the same cluster of $\mathcal{C}$ and 0 otherwise. Then, we may express the Corr-Clust objective as

$$\sum_{C_r \in \mathcal{C}} \sum_{i,j \in C_r} \omega_{i,j} = \sum_{i=1}^{n} \sum_{j=1}^{n} \omega_{i,j} [Z_{\mathcal{C}}]_{(i,j)} = \operatorname{tr}(\Omega \; Z_{\mathcal{C}}) \ .$$

Throughout the remainder of the section, it will be useful to write the Corr-Clust objective using this matrix notation.

C.1. Reformulating Exposure-Design *as* Corr-Clust

We are now ready to prove Proposition 7.1, which we restate here for completeness.

Proposition 7.1. *For each pair of diversion units $i, j \in V_d$, define the value $\omega_{i,j} \in \mathbb{R}$ as*

$$\omega_{i,j} = (1+\phi) \sum_{k=1}^{m} w_{k,i} w_{k,j} - \phi \Big(\sum_{k=1}^{m} w_{k,i} \Big) \Big(\sum_{k=1}^{m} w_{k,j} \Big) \ , \tag{2}$$

where $w_{k,i}$ is the weight of the edge between the kth outcome unit and the ith diversion unit. Exposure-Design *is equivalent to the following clustering problem:*

$$\max_{\text{clusterings } \mathcal{C}} \quad \sum_{C_r \in \mathcal{C}} \sum_{i,j \in C_r} \omega_{i,j} \ . \tag{Corr-Clust}$$

Proof. Recall that the objective of Exposure-Design is defined as

$$\sum_{i=1}^{n} \mathrm{Var}(x_i) - \phi \sum_{i \neq j} \mathrm{Cov}(x_i, x_j) \ ,$$

where the expectation in the variance and covariance terms is taken with respect to the random assignment vector $\boldsymbol{z} \in \{0,1\}^m$, which is drawn from the cluster design given by $\mathcal{C}$. Recall that the exposures are given by $\boldsymbol{x} = \boldsymbol{W}\boldsymbol{z}$. Using matrix notation, we can more compactly represent this objective as

$$\begin{aligned}
&\sum_{i=1}^{n} \mathrm{Var}(x_i) - \phi \sum_{i \neq j} \mathrm{Cov}(x_i, x_j) \\
&\quad = \mathrm{tr}\Big(\big(\boldsymbol{I} - \phi(\boldsymbol{1}\boldsymbol{1}^\intercal - \boldsymbol{I}) \big) \, \mathrm{Cov}(\boldsymbol{x}) \Big) && \text{(rewriting in terms of tr)} \\
&\quad = \mathrm{tr}\Big(\big((1+\phi)\boldsymbol{I} - \phi \boldsymbol{1}\boldsymbol{1}^\intercal \big) \, \mathrm{Cov}(\boldsymbol{x}) \Big) && \text{(rearranging terms)} \\
&\quad = \mathrm{tr}\Big(\big((1+\phi)\boldsymbol{I} - \phi \boldsymbol{1}\boldsymbol{1}^\intercal \big) \, \mathrm{Cov}(\boldsymbol{W}\boldsymbol{z}) \Big) && \text{(definition of exposure)} \\
&\quad = \mathrm{tr}\Big(\big((1+\phi)\boldsymbol{I} - \phi \boldsymbol{1}\boldsymbol{1}^\intercal \big) \boldsymbol{W} \, \mathrm{Cov}(\boldsymbol{z}) \boldsymbol{W}^\intercal \Big) && \text{(property of covariance)} \\
&\quad = \mathrm{tr}\Big(\boldsymbol{W}^\intercal \big((1+\phi)\boldsymbol{I} - \phi \boldsymbol{1}\boldsymbol{1}^\intercal \big) \boldsymbol{W} \, \mathrm{Cov}(\boldsymbol{z}) \Big) && \text{(cyclic property of trace)}
\end{aligned}$$

Because $\boldsymbol{z}$ is drawn from an independent cluster design, the (i,j)th entry of the covariance matrix $\mathrm{Cov}(\boldsymbol{z})$ is $1/2$ if diversion units i and j are in the same cluster and 0 otherwise. Thus, by the observation above, this clustering objective is a correlation clustering where the weights are given by the matrix

$$\Omega = \boldsymbol{W}^\intercal \big((1+\phi)\boldsymbol{I} - \phi \boldsymbol{1}\boldsymbol{1}^\intercal \big) \boldsymbol{W} \ .$$

By inspection, we have that the (i,j)th entry of this matrix Ω is

$$\omega_{i,j} = (1+\phi) \sum_{k=1}^{n} w_{k,i} w_{k,j} - \phi \Big(\sum_{k=1}^{n} w_{k,i} \Big) \Big(\sum_{k=1}^{n} w_{k,j} \Big) \ ,$$

as desired. □

C.2. An instance of Exposure-Design *when* $\phi = 1/(n-1)$

Now we demonstrate that the correlation clustering objective proposed in [49] is a special case of Exposure-Design when $\phi = 1/(n-1)$ and no constraint is placed on cluster sizes, i.e. $k = m$. Before giving the formal statement, we re-introduce the clustering objective in that paper; that is,

$$\max_{\text{clusterings } \mathcal{C}} \mathbb{E}\left[\sum_{i=1}^{n} \left(x_i - \Big(\frac{1}{n} \sum_{j=1}^{n} x_j \Big) \right)^2 \right] \ , \qquad \text{(Exposure-Spread)}$$

where the expectation is with respect to the treatment vector $\boldsymbol{z} \in \{\pm 1\}^m$ drawn according to the independent cluster design given by $\mathcal{C}$. The quantity in the expectation is a measure of the spread of the exposures. We remark that in [49], the exposures are called "doses" and the quantity in the expectation is referred to as the "empirical dose variance".

Proposition C.1. *Up to additive and multiplicative constants,* Exposure-Spread *is equivalent to* Exposure-Design *when the trade-off parameter is set to* $\phi = 1/(n-1)$.

Proof. Let us denote the exposure spread by

$$Q = \sum_{i=1}^{n} \left(x_i - \left(\frac{1}{n} \sum_{j=1}^{n} x_j \right) \right)^2 ,$$

Note that the exposure spread is equal to the ℓ_2 norm of the *de-meaned* exposure vector $\bar{\boldsymbol{x}} = (\bar{x_1}, \bar{x_2}, \dots \bar{x_n})$, where

$$\bar{x_i} = x_i - \left(\frac{1}{n} \sum_{j=1}^{n} x_j \right) .$$

The entire de-meaned exposure vector may be written as $\bar{\boldsymbol{x}} = (\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T)\boldsymbol{x}$. Using the fact that this matrix is a projection and that the exposure vector is $\boldsymbol{x} = \boldsymbol{W}\boldsymbol{z}$, we can write the exposure spread as

$$\begin{aligned} Q = \|\bar{\boldsymbol{x}}\|^2 = \|(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T)\boldsymbol{x}\|^2 = \boldsymbol{x}^\intercal (\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T)^2 \boldsymbol{x} \\ = \boldsymbol{x}^\intercal (\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T)\boldsymbol{x} = \boldsymbol{z}^\intercal \boldsymbol{W}^\intercal (\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T)\boldsymbol{W}\boldsymbol{z} \ . \end{aligned}$$

Finally, the expectation of the exposure spread may be written as

$$\begin{aligned} \mathbb{E}[Q] &= \mathbb{E}\left[\boldsymbol{z}^\intercal \boldsymbol{W}^\intercal (\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T)\boldsymbol{W}\boldsymbol{z} \right] && \text{(from above)} \\ &= \mathbb{E}\left[\operatorname{tr}\left(\boldsymbol{z}^\intercal \boldsymbol{W}^\intercal (\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T)\boldsymbol{W}\boldsymbol{z} \right) \right] && \text{(trace of a scalar)} \\ &= \mathbb{E}\left[\operatorname{tr}\left(\boldsymbol{W}^\intercal (\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T)\boldsymbol{W}\boldsymbol{z}\boldsymbol{z}^\intercal \right) \right] && \text{(cyclic property of trace)} \\ &= \operatorname{tr}\left(\boldsymbol{W}^\intercal (\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T)\boldsymbol{W}\mathbb{E}[\boldsymbol{z}\boldsymbol{z}^\intercal] \right) && \text{(linearity of trace)} \\ &= \operatorname{tr}\left(\boldsymbol{W}^\intercal (\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T)\boldsymbol{W} \operatorname{Cov}(\boldsymbol{z}) \right) + c \ , \end{aligned}$$

where the value c in the last line is $c = \operatorname{tr}\left(\boldsymbol{W}^\intercal (\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T)\boldsymbol{W}\mathbb{E}[\boldsymbol{z}]\mathbb{E}[\boldsymbol{z}]^\intercal \right)$, which follows from $\operatorname{Cov}(\boldsymbol{z}) = \mathbb{E}[\boldsymbol{z}\boldsymbol{z}^\intercal] - \mathbb{E}[\boldsymbol{z}]\mathbb{E}[\boldsymbol{z}]^\intercal$ and linearity of trace. Moreover, when

the probability of treatment assignment p is fixed, this value c is a constant with respect to the clustering being chosen.

Observe that by setting $\phi = 1/(n-1)$ and multiplying by a factor $(n-1)/n$, the Exposure-Design objective becomes

$$\frac{n-1}{n} \cdot \operatorname{tr}\Big(\boldsymbol{W}^\intercal\big((1+\frac{1}{n-1})\boldsymbol{I} - \frac{1}{n-1}\mathbf{1}\mathbf{1}^\intercal\big)\boldsymbol{W}\operatorname{Cov}(\boldsymbol{z})\Big)$$
$$= \operatorname{tr}\Big(\boldsymbol{W}^\intercal(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T)\boldsymbol{W}\operatorname{Cov}(\boldsymbol{z})\Big) \ .$$

Thus, the Exposure-Spread objective is equivalent (up to additive and multiplicative constants) to the Exposure-Design objective when $\phi = 1/(n-1)$. □

C.3. Comparison to other correlation clustering variants

Recall that we defined the objective of the correlation clustering variant Corr-Clust as

$$\sum_{C_r \in \mathcal{C}} \sum_{i,j \in C_r} \omega_{i,j} \ ,$$

where $\omega_{i,j}$ is defined for each pair of diversion units $i, j \in V_d$ as

$$\omega_{i,j} = (1+\phi)\sum_{k=1}^{n} w_{k,i}w_{k,j} - \phi\Big(\sum_{k=1}^{n} w_{k,i}\Big)\Big(\sum_{k=1}^{n} w_{k,j}\Big) \ ,$$

and $w_{k,i}$ is the weight of the edge between the kth outcome unit and the ith diversion unit. Observe that the term $\omega_{i,j}$ can take positive or negative values.

The maximization weighted correlation clustering variant considered by [14, 58] is defined as follows. Let $G = (V, E)$ be a graph where each edge $e = (i, j) \in E$ has two *non-negative* weights: $w_{in}(i, j)$ and $w_{out}(i, j)$. Given a clustering $\mathcal{C}$, an edge $e = (i, j)$ is said to be *in-cluster* if i and j are in the same cluster and *out-cluster* otherwise. The objective function for a given clustering is given by

$$\sum_{\substack{\text{in-cluster} \\ \text{edges } e}} w_{in}(e) + \sum_{\substack{\text{out-cluster} \\ \text{edges } e}} w_{out}(e) \qquad \text{(Corr-Clust-CS)}$$

We now show that the Corr-Clust objective may be written as an instance of the Corr-Clust-CS objective, but with the addition of a large additive constant. Again, we stress that this reduction is primarily for aesthetic comparison purposes because the appearance of the large additive constant prevents any meaningful approximation-preserving reduction.

Proposition C.2. *Our formulation* Corr-Clust *may be viewed as an instance of* Corr-Clust-CS *with a large additive constant. More precisely, let* $w_{in}(i, j) = \max\{0, \omega_{i,j}\}$ *and* $w_{out}(i, j) = \min\{0, \omega_{i,j}\}$. *For a clustering* $\mathcal{C}$, *we have that the objectives are related by*

$$\sum_{C_r \in \mathcal{C}} \sum_{i,j \in C_r} \omega_{i,j} - \sum_{i=1}^{n}\sum_{j=1}^{n} \min\{0, \omega_{i,j}\} = \sum_{\substack{\text{in-cluster} \\ \text{edges } e}} w_{in}(e) + \sum_{\substack{\text{out-cluster} \\ \text{edges } e}} w_{out}(e)$$

Proof. For each pair of diversion units i, j, define $\omega^+_{i,j} = \max\{0, \omega_{i,j}\}$ and $\omega^-_{i,j} = -\min\{0, \omega_{i,j}\}$. Observe that $\omega_{i,j} = \omega^+_{i,j} + \omega^-_{i,j}$ and so we can re-distribute the following sum as

$$\sum_{C_r \in \mathcal{C}} \sum_{i,j \in C_r} \omega_{i,j} = \sum_{C_r \in \mathcal{C}} \sum_{i,j \in C_r} \left(\omega^+_{i,j} + \omega^-_{i,j}\right) = \sum_{C_r \in \mathcal{C}} \sum_{i,j \in C_r} \omega^+_{i,j} + \sum_{C_r \in \mathcal{C}} \sum_{i,j \in C_r} \omega^-_{i,j} \ .$$

Subtracting the (instance-dependent) constant $\sum_{i=1}^n \sum_{j=1}^n \min\{0, \omega_{i,j}\}$ from both sides and rearranging yields

$$\begin{aligned}
&\sum_{C_r \in \mathcal{C}} \sum_{i,j \in C_r} \omega_{i,j} - \sum_{i=1}^n \sum_{j=1}^n \min\{0, \omega_{i,j}\} \\
&\qquad = \sum_{C_r \in \mathcal{C}} \sum_{i,j \in C_r} \omega^+_{i,j} + \sum_{C_r \in \mathcal{C}} \sum_{i,j \in C_r} \omega^-_{i,j} - \sum_{i=1}^n \sum_{j=1}^n \min\{0, \omega_{i,j}\} \\
&\qquad = \sum_{C_r \in \mathcal{C}} \sum_{i,j \in C_r} \omega^+_{i,j} - \sum_{C_r \neq C'_r \in \mathcal{C}} \sum_{\substack{i \in C_r \\ j \in C'_r}} \omega^-_{i,j} \\
&\qquad = \sum_{C_r \in \mathcal{C}} \sum_{i,j \in C_r} \omega^+_{i,j} + \sum_{C_r \neq C'_r \in \mathcal{C}} \sum_{\substack{i \in C_r \\ j \in C'_r}} (-\omega^-_{i,j}) \\
&\qquad = \sum_{\substack{\text{in-cluster} \\ \text{edges } e}} w_{in}(e) + \sum_{\substack{\text{out-cluster} \\ \text{edges } e}} w_{out}(e) \ .
\end{aligned}$$

Finally, observe that for each pair (i, j), the values $w_{in}(i, j)$ and $w_{out}(i, j)$ are non-negative so that the final equation is a valid objective function for the Corr-Clust-CS formulation. □

Acknowledgments

We thank P.M. Aronow, Kay Brodersen, Nick Doudchenko, Ramesh Johari, Khashayar Khosravi, Sebastien Lahaie, Vahan Nanumyan, Georgia Papadogeorgou, Lewis Rendell, Johan Ugander, and C. M. Zigler for stimulating discussions which helped shape this work.

References

[1] Angrist, J. D. (1998). Estimating the labor market impact of voluntary military service using social security data on military applicants. *Econometrica*, 66(2):249–288.

[2] Aral, S. and Walker, D. (2011). Creating social contagion through viral product design: A randomized trial of peer influence in networks. *Management science*, 57(9):1623–1639.

[3] Aronow, P. M. and Samii, C. (2013). Conservative variance estimation for sampling designs with zero pairwise inclusion probabilities. *Survey Methodology*, 39(1):231–241.

[4] Aronow, P. M. and Samii, C. (2015). Does regression produce representative estimates of causal effects? *American Journal of Political Science*, 60(1):250–267.

[5] Aronow, P. M. and Samii, C. (2017). Estimating average causal effects under general interference, with application to a social network experiment. *The Annals of Applied Statistics*, 11(4):1912–1947. MR3743283

[6] Aydin, K., Bateni, M., and Mirrokni, V. (2019). Distributed balanced partitioning via linear embedding. *Algorithms*, 12(8):162. MR3998661

[7] Baird, S., Bohren, J. A., McIntosh, C., and Özler, B. (2018). Optimal Design of Experiments in the Presence of Interference. *The Review of Economics and Statistics*, 100(5):844–860.

[8] Bansal, N., Blum, A., and Chawla, S. (2002). Correlation clustering. In *Proceedings of the 43rd Symposium on Foundations of Computer Science*, FOCS'02, page 238. IEEE Computer Society. MR3363423

[9] Basse, G. W. and Airoldi, E. M. (2018). Limitations of design-based causal inference and A/B testing under arbitrary and network interference. *Sociological Methodology*, 48(1):136–151.

[10] Basse, G. W., Soufiani, H. A., and Lambert, D. (2016). Randomization and the pernicious effects of limited budgets on auction experiments. In *Artificial Intelligence and Statistics*, pages 1412–1420. PMLR.

[11] Blake, T. and Coey, D. (2014). Why marketplace experimentation is harder than it seems: The role of test-control interference. In *Proceedings of the fifteenth ACM conference on Economics and computation*, pages 567–582.

[12] Bramoullé, Y., Djebbari, H., and Fortin, B. (2009). Identification of peer effects through social networks. *Journal of Econometrics*, 150(1):41–55. MR2525993

[13] Chamberlain, G. (1984). Panel data. In Griliches, Z. and Intriligator, M. D., editors, *Handbook of Econometrics*, volume 2, pages 1247–1318. Elsevier. MR0726759

[14] Charikar, M., Guruswami, V., and Wirth, A. (2005). Clustering with qualitative information. *J. Comput. Syst. Sci.*, 71(3):360–383. MR2168358

[15] Chin, A. (2019a). Central limit theorems via Stein's method for randomized experiments under interference. arXiv:1804.03105.

[16] Chin, A. (2019b). Regression adjustments for estimating the global treatment effect in experiments with interference. *Journal of Causal Inference*, 7(2). MR4350071

[17] Cohen, E. and Lewis, D. D. (1999). Approximating matrix multiplication for pattern recognition tasks. *Journal of Algorithms*, 30(2):211 – 252. MR1671828

[18] Doudchenko, N., Zhang, M., Drynkin, E., Airoldi, E., Mirrokni, V., and Pouget-Abadie, J. (2020). Causal inference with bipartite designs. *arXiv preprint arXiv:2010.02108*.

[19] Eckles, D., Karrer, B., and Ugander, J. (2016a). Design and analysis of ex-

periments in networks: Reducing bias from interference. *Journal of Causal Inference*, 5(1). MR4323809

[20] Eckles, D., Kizilcec, R. F., and Bakshy, E. (2016b). Estimating peer effects in networks with peer encouragement designs. *Proceedings of the National Academy of Sciences*, 113(27):7316–7322.

[21] Einav, L., Kuchler, T., Levin, J. D., and Sundaresan, N. (2011). Learning from seller experiments in online markets. Technical report, National Bureau of Economic Research.

[22] Elsner, M. and Schudy, W. (2009). Bounding and comparing methods for correlation clustering beyond ilp. In *Proceedings of the Workshop on Integer Linear Programming for Natural Langauge Processing*, ILP'09, page 19–27. Association for Computational Linguistics.

[23] Fattorini, L. (2006). Applying the Horvitz-Thompson criterion in complex designs: A computer-intensive perspective for estimating inclusion probabilities. *Biometrika*, 93(2):269–278. MR2278082

[24] Fradkin, A. (2015). Search frictions and the design of online marketplaces. *Work. Pap., Mass. Inst. Technol.*

[25] Fradkin, A. (2017). Search, matching, and the role of digital marketplace design in enabling trade: Evidence from airbnb. Available at SSRN: https://ssrn.com/abstract=2939084.

[26] Goldberger, A. S. (1991). *A Course in Econometrics*. Harvard University Press, Cambridge. MR1133625

[27] Gupta, S., Kohavi, R., Tang, D., Xu, Y., Andersen, R., Bakshy, E., Cardin, N., Chandran, S., Chen, N., Coey, D., et al. (2019). Top challenges from the first practical online controlled experiments summit. *ACM SIGKDD Explorations Newsletter*, 21(1):20–35.

[28] Halloran, M. E. and Hudgens, M. G. (2016). Dependent happenings: A recent methodological review. *Current Epidemiology Reports*, 3(4):297–305.

[29] Harshaw, C., Sävje, F., Spielman, D., and Zhang, P. (2021). Balancing covariates in randomized experiments with the gram-schmidt walk design. arXiv:1911.03071.

[30] He, R. and McAuley, J. (2016). Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *proceedings of the 25th international conference on world wide web*, pages 507–517.

[31] Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960. MR0867618

[32] Holtz, D., Lobel, R., Liskovich, I., and Aral, S. (2020). Reducing interference bias in online marketplace pricing experiments. *arXiv preprint arXiv:2004.12489*.

[33] Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685. MR0053460

[34] Hudgens, M. G. and Halloran, M. E. (2008). Toward causal inference with interference. *Journal of the American Statistical Association*, 103(482):832–842. MR2435472

[35] Imbens, G. W. and Rubin, D. B. (2015). *Causal inference in statistics,*

social, and biomedical sciences. Cambridge University Press. MR3309951

[36] Johari, R., Li, H., and Weintraub, G. (2020). Experimental design in two-sided platforms: An analysis of bias. In *Proceedings of the 21st ACM Conference on Economics and Computation*, page 851.

[37] Kempton, R. (1997). Interference between plots. In *Statistical methods for plant variety evaluation*, pages 101–116. Springer.

[38] Li, X., Ding, P., and Rubin, D. B. (2018). Asymptotic theory of rerandomization in treatment–control experiments. *Proceedings of the National Academy of Sciences*, 115(37):9157–9162. MR3856113

[39] Liu, M., Mao, J., and Kang, K. (2020). Trustworthy online marketplace experimentation with budget-split design. *arXiv preprint arXiv:2012.08724*.

[40] Lock Morgan, K. and Rubin, D. B. (2012). Rerandomization to improve covariate balance in experiments. *Annals of Statistics*, 40(2):1263–1282. MR2985950

[41] Manski, C. F. (1991). Regression. *Journal of Economic Literature*, 29(1):34–50.

[42] Manski, C. F. (1993). Identification of endogenous social effects: The reflection problem. *The Review of Economic Studies*, 60(3):531–542. MR1236836

[43] McAuley, J., Targett, C., Shi, Q., and Van Den Hengel, A. (2015). Image-based recommendations on styles and substitutes. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pages 43–52.

[44] Miratrix, L. W., Sekhon, J. S., and Yu, B. (2013). Adjusting treatment effect estimates by post-stratification in randomized experiments. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(2):369–396. MR3021392

[45] Narain, R. (1951). On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Studies*, (3):169–175. MR0045354

[46] Neyman, J. (1923). Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych*, 10:1–51.

[47] Offer-Westort, M. and Dimmery, D. (2021). Experimentation for homogenous policy change. arXiv:2101.12318.

[48] Ogburn, E. L., Sofrygin, O., Diaz, I., and van der Laan, M. J. (2020). Causal inference for social network data. arXiv:1705.08527.

[49] Pouget-Abadie, J., Aydin, K., Schudy, W., Brodersen, K., and Mirrokni, V. (2019). Variance reduction in bipartite experiments through correlation clustering. In *Advances in Neural Information Processing Systems*, pages 13309–13319.

[50] Reiley, D. H. (2006). Field experiments on the effects of reserve prices in auctions: More magic on the internet. *The RAND Journal of Economics*, 37(1):195–211.

[51] Robinson, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica*, 56(4):931–954. MR0951762

[52] Ross, N. (2011). Fundamentals of Stein's method. *Probability Surveys*, 8:210–293. MR2861132

[53] Sävje, F. (2021). Causal inference with misspecified exposure mappings. arXiv:2103.06471.

[54] Sävje, F., Aronow, P. M., and Hudgens, M. G. (2021). Average treatment effects in the presence of unknown interference. *The Annals of Statistics*, 49(2):673–701. MR4255103

[55] Sinclair, B., McConnell, M., and Green, D. P. (2012). Detecting spillover effects: Design and analysis of multilevel experiments. *American Journal of Political Science*, 56(4):1055–1069.

[56] Sloczynski, T. (2020). Interpreting ols estimands when treatment effects are heterogeneous: Smaller groups get larger weights. *Review of Economics and Statistics*, in press.

[57] Struchiner, C. J., Halloran, M. E., Robins, J. M., and Spielman, A. (1990). The behaviour of common measures of association used to assess a vaccination programme under complex disease transmission patterns—a computer simulation study of malaria vaccines. *International journal of epidemiology*, 19(1):187–196.

[58] Swamy, C. (2004). Correlation clustering: Maximizing agreements via semidefinite programming. SODA'04, page 526–527. Society for Industrial and Applied Mathematics. MR2291092

[59] Toulis, P. and Kao, E. (2013). Estimation of causal peer influence effects. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28, pages 1489–1497.

[60] Wilks, S. (1932). Certain Generalizations in the Analysis of Variance. *Biometrika*, 24(3-4):471–494.

[61] Zigler, C. M. and Papadogeorgou, G. (2021). Bipartite causal inference with interference. *Statist. Sci.*, 36(1):109–123. MR4194206