Research Article

Johan Ugander* and Hao Yin

Randomized graph cluster randomization

https://doi.org/10.1515/jci-2022-0014
received February 24, 2022; accepted January 28, 2023

Abstract: The global average treatment effect (GATE) is a primary quantity of interest in the study of causal inference under network interference. With a correctly specified exposure model of the interference, the Horvitz–Thompson (HT) and Hájek estimators of the GATE are unbiased and consistent, respectively, yet known to exhibit extreme variance under many designs and in many settings of interest. With a fixed clustering of the interference graph, graph cluster randomization (GCR) designs have been shown to greatly reduce variance compared to node-level random assignment, but even so the variance is still often prohibitively large. In this work, we propose a randomized version of the GCR design, descriptively named randomized graph cluster randomization (RGCR), which uses a random clustering rather than a single fixed clustering. By considering an ensemble of many different clustering assignments, this design avoids a key problem with GCR where the network exposure probability of a given node can be exponentially small in a single clustering. We propose two inherently randomized graph decomposition algorithms for use with RGCR designs, *randomized 3-net* and *1-hop-max*, adapted from the prior work on multiway graph cut problems and the probabilistic approximation of (graph) metrics. We also propose weighted extensions of these two algorithms with slight additional advantages. All these algorithms result in network exposure probabilities that can be estimated efficiently. We derive structure-dependent upper bounds on the variance of the HT estimator of the GATE, depending on the metric structure of the graph driving the interference. Where the best-known such upper bound for the HT estimator under a GCR design is exponential in the parameters of the metric structure, we give a comparable upper bound under RGCR that is instead polynomial in the same parameters. We provide extensive simulations comparing RGCR and GCR designs, observing substantial improvements in GATE estimation in a variety of settings.

Keywords: causal inference under interference, global average treatment effect, network effects, spillovers, social networks

MSC 2020: 62D05

1 Introduction

Interest in the design and analysis of randomized experiments under interference has accelerated in recent years [1–6], motivating work on efficient estimators of the global average treatment effect (GATE) [7–9]. GATE estimation seeks to understand the difference between placing *all* units in treatment vs placing *all* units in control, a natural estimand capturing the full average treatment effect net of all "network effects." A major motivation for studying the GATE comes from experiments run on online social networking platforms [10–12] and online marketplaces [13–17], where the interactions are either between social relations or between marketplace competitors. In these settings, a platform designer typically has full control over

* **Corresponding author: Johan Ugander,** Management Science and Engineering, Stanford University, Stanford, CA 94305, California, United States, e-mail: jugander@stanford.edu
Hao Yin: Institute for Computational and Mathematical Engineering, Stanford University, Stanford, CA 94305, California, United States, e-mail: yinh@stanford.edu

treatment assignments and is specifically interested in understanding which condition, when assigned to all units, has the best average outcome.

In the case of a binary intervention, a so-called A/B test of treatment versus control, the GATE is defined as the difference between the average of outcomes when all individuals are exposed to the treatment condition vs when all individuals are exposed to control. Formally, let $\boldsymbol{Z} \in \{0, 1\}^n$ be a length-n vector representing the treatment assignment of a population of n individuals, where the value of 1 and 0 correspond to treatment and control, respectively. Let $Y_i(\boldsymbol{Z} = \boldsymbol{z})$ be the ith individual's outcome or response; the mean outcome of all units to $\boldsymbol{Z}$ is expressed as follows:

$$\mu(\boldsymbol{z}) \triangleq \frac{1}{n}\sum_{i=1}^{n} Y_i(\boldsymbol{Z} = \boldsymbol{z}),$$

and the GATE is then $\tau \triangleq \mu(\mathbf{1}) - \mu(\mathbf{0})$. For work relating this potential outcomes formulation of interference to causal diagrams, see [18,19].

Exact measurement of the GATE is not possible because the scenarios $\boldsymbol{z} = \mathbf{1}$ and $\boldsymbol{z} = \mathbf{0}$ are strongly counterfactual: it is not possible to simultaneously observe the entire population in treatment and the entire population in control. The GATE is typically estimated through randomized experiments, but to connect the outcome of a randomized experiment with the GATE, assumptions are required to make $\mu(\boldsymbol{z})$ identifiable, e.g., the *no interference* [20] or *stable unit treatment value assumption* (SUTVA) [21]. However, in many situations, there is unavoidable interference between individuals, in the sense that their outcome depends directly on the treatment or outcome of others. In the presence of interference, estimators derived under the SUTVA assumptions are generally biased [3,22]. A variety of alternative assumptions have been made in attempts to bring reasonable power to potential outcome inferences under interference, including monotonicity assumptions on the individual treatment effect [7,11,23,24]. In this work, we do not require a monotonicity assumption for our results to hold, but instead commit to an exposure model framework [23,25,26].

In prior efforts to estimate the GATE, a promising approach has been to replace the SUTVA assumption with a less restrictive *exposure model* [3,23,27]. An exposure model identifies, for each unit i, the condition when the unit has the same response as if all units are assigned to treatment or control. We use $E_i^{\boldsymbol{z}}$ to denote the events – defined by subsets of global assignment vectors, to be formally specified later on – where node i responds as if exposed to global treatment ($\boldsymbol{z} = \mathbf{1}$) or global control ($\boldsymbol{z} = \mathbf{0}$). For network experiments, $E_i^{\mathbf{1}}$ and $E_i^{\mathbf{0}}$ then capture conditions under which we consider i to be "network exposed to treatment" vs "network exposed to control." Throughout this work, we will focus our attention on the *full-neighborhood exposure model*, discussed further in Section 2.2.

The Horvitz–Thompson (HT) estimator [28] of the mean outcome $\mu(\boldsymbol{z})$, $\boldsymbol{z} \in \{\mathbf{0}, \mathbf{1}\}$, is

$$\hat{\mu}(\boldsymbol{z}) = \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbb{1}[E_i^{\boldsymbol{z}}] \cdot Y_i}{\mathbb{P}[E_i^{\boldsymbol{z}}]}, \tag{1}$$

where Y_i is the *observed outcome* of node i, which equals $Y_i(\mathbf{1})$ (or $Y_i(\mathbf{0})$) when node i is network exposed to treatment (or control). Consequently, the HT estimator for the GATE is $\hat{\tau} = \hat{\mu}(\mathbf{1}) - \hat{\mu}(\mathbf{0})$. Aronow and Samii [3] have shown that, assuming the exposure model is not misspecified, a standard consistency assumption on the potential outcomes [29] and that the probability of every node being network exposed to treatment and control is positive, then the estimators $\hat{\mu}(\mathbf{1})$, $\hat{\mu}(\mathbf{0})$, and $\hat{\tau}$ are unbiased. Here, an exposure model being properly specified means that it correctly specifies the graph-based conditions for each node to respond as if exposed to global treatment or control and that the interference graph is fully known (cf. recent interference work with unknown network structure [30]). While we focus our analysis of GATE estimation on the HT estimator, some of our results extend to the related self-normalized Hájek estimator [31].

Under independent node-level Bernoulli(p) randomization – where units are assigned for treatment with probability p and control with probability $(1 - p)$ – the variance of the HT GATE estimator quickly blows up if there are units i for which the exposure conditions $E_i^{\mathbf{1}}$ or $E_i^{\mathbf{0}}$ require many independent assignments to all come up heads or all come up tails. For exposure models such as full-neighborhood exposure, where a unit and all of its network neighbors must be assigned to treatment together, $\mathbb{P}[E_i^{\mathbf{1}}]$ and/or $\mathbb{P}[E_i^{\mathbf{0}}]$ then quickly become much smaller.

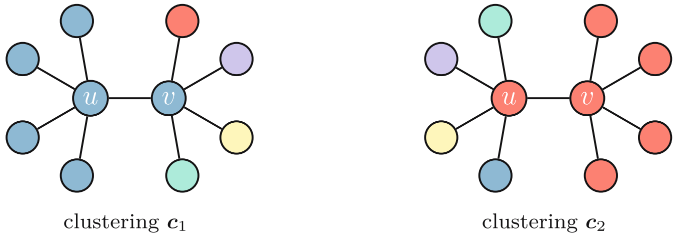
The *graph cluster randomization* (GCR) [27] experimental design scheme was proposed to combat this issue. Given a fixed clustering of the graph, i.e., the set of nodes has been partitioned into disjoint *clusters*, GCR jointly assigns all nodes within each cluster into either treatment or control together. This randomization design can be viewed as a correlation imposed on the way in which assignment vectors $\boldsymbol{Z} \in \{0, 1\}^n$ are drawn, correlating neighbors in the graph with the goal of broadly increasing the collections of probabilities $\mathbb{P}[E_i^{\mathbf{1}}]$ and $\mathbb{P}[E_i^{\mathbf{0}}]$, for all nodes i, for a given exposure model. GCR can be shown to achieve a considerable variance reduction in certain settings compared to independent assignment. Eckles et al. [7] evaluated GCR for GATE estimation and showed that it reduces bias, variance, and mean squared error (MSE) in scenarios where there is a strong direct treatment effect and network spillover. However, they found that it often still exhibits considerable MSE, which can then exceed the MSE of independent assignment when spillover effects are small.

The GCR scheme operates using a prespecified fixed clustering assignment. A known problem with GCR is that, with a single fixed clustering, a node can be "unlucky" in how its neighborhood was divided, meaning that it is adjacent to many clusters, and thus, its network exposure probability is small. Such small exposure probability greatly inflates the variance of the HT GATE estimator $\hat{\tau}$. Therefore, even though GCR has been shown to theoretically give considerable variance reductions compared with node-level randomization, the variance can still be much larger. Another disadvantage of GCR is the incompatibility with *complete randomization* at the cluster level due to a violation of the positivity assumption required by both the HT and Hájek GATE estimators.

1.1 Main contributions

We propose an extension of the GCR scheme whereby the graph cluster randomization is itself based on a randomized clustering. We descriptively call this scheme *randomized graph cluster randomization* (*RGCR*). We find that RGCR can greatly reduce the variance of the HT GATE estimator in both theory and extensive simulations, compared to ordinary GCR. Further simulations using the Hájek GATE estimator, while lacking theoretical support, show that it too benefits from RGCR (vs GCR) and is often preferable to the HT estimator for a given design. Most importantly, we find that these variance reductions are considerable enough to bring RGCR into the realm of being "useful" in many situations where GCR would fail to deliver a GATE estimate with actionable MSE.

The intuition that motivates using a random clustering partition is illustrated in Figure 1. Essentially, when averaging across different clustering assignments, the distribution of individual network exposure probabilities $\mathbb{P}[E_i^{\boldsymbol{z}}]$ will be less skewed because different nodes will be "unlucky" in different clusterings. Averaging across many clusterings washes out extremely small probabilities, greatly reducing GATE estimator variance.



clustering $\boldsymbol{c}_1$ clustering $\boldsymbol{c}_2$

Figure 1: An illustration of variance reduction with RGCR, considering two different clusterings $\boldsymbol{c}_1$ and $\boldsymbol{c}_2$, where colors denote clusters. Consider GCR with clusters assigned to treatment or control with probability $p = 1/2$, the full-neighborhood exposure probabilities of nodes u and v are either 2^{-1} and 2^{-5} (under $\boldsymbol{c}_1$) or 2^{-5} and 2^{-1} (under $\boldsymbol{c}_2$) respectively, contributing $2 + 2^5 = 34$ to the variance of the HT estimator of the GATE. In contrast, when randomizing evenly between $\boldsymbol{c}_1$ and $\boldsymbol{c}_2$, the exposure probabilities of both u and v become $(2^{-1} + 2^{-5})/2$, contributing $2/(2^{-1} + 2^{-5}) * 2 \approx 7.5$ to the variance.

One can consider two approaches to randomized graph clustering. First, consider employing a uniform mixture of K graph clusterings, each obtained via a (potentially different) black box clustering algorithm. In this setting, we can compute the exposure probabilities simply by averaging the exposure probabilities across clusterings. That said, computing many clusterings of a large graph can be very computationally expensive. As a more appealing approach, we consider employing inherently randomized graph clustering algorithms where it is potentially tractable to consider the exposure probabilities when integrating over the full randomness of the algorithm. Even though the exact computation of the full-neighborhood exposure probabilities can be NP-hard, we are able to construct Monte Carlo estimators of the probabilities with relative errors bounded at a reasonable computational cost. The Monte Carlo estimation procedure we employed is practically equivalent to generating K clusterings from the randomized algorithms and then averaging, but we do not need to store all K clusterings at any point.

To make our theoretical and simulation analyses concrete, in this article, we consider two specific randomized graph clustering algorithms. One algorithm is the *randomized 3-net* clustering, whose deterministic version has been analyzed in the causal inference literature [7,27]. The other algorithm, called *1-hop-max*, is very similar to 3-net, but it is more amenable to theoretical analysis.

1.1.1 Variance reduction

Our main theoretical result is to show that, under a structural assumption known as a restricted growth condition, RGCR delivers qualitatively better bounds on the HT variance compared to GCR (which is already known to be qualitatively better than independent randomization). Specifically, consider a graph of n nodes with largest degree $d_{\max}$ and restricted growth coefficient κ. Here, the restricted growth coefficient κ measures the growth rate of the r-hop neighborhood with the neighborhood radius r. Theoretically we have $\kappa \le 1 + d_{\max}$, though we expect κ to be substantially smaller than $d_{\max}$ on real-world networks that exhibit clustering structures. Let the probability of each cluster being assigned to the treatment group be p, previous results [27] have shown that the variance of the HT estimator of $\hat{\mu}(\mathbf{1})$ under GCR with a fixed 3-net clustering is upper bounded by

$$\mathbf{Var}[\hat{\mu}(\mathbf{1})] \le \frac{1}{n} d_{\max} \kappa^5 p^{-\kappa^6} \Theta(1), \tag{2}$$

polynomial in the maximum degree $d_{\max}$ but exponential in κ. In this work, we show that under RGCR with a randomized 1-hop-max clustering, we can upper bound the variance by

$$\mathbf{Var}[\hat{\mu}(\mathbf{1})] \le \frac{1}{n} d_{\max}^2 \kappa^4 p^{-1} \Theta(1), \tag{3}$$

polynomial in both κ and $d_{\max}$. The bounds on the variance of the HT GATE estimator $\hat{\tau}$ are analogous to these bounds for the mean outcome $\hat{\mu}(\mathbf{1})$. The exponential difference between these two variance bounds, (2) under GCR vs (3) under RGCR, is striking both in settings of a fixed modest κ and in settings where κ is on the order of $d_{\max}$. Per our analysis in Appendix A, the latter setting is empirically quite common.

The pairing of RGCR with 1-hop-max (instead of randomized 3-net) is for analytical convenience: the two algorithms are very similar, but once randomized, the distribution of clusterings produced by the randomized 3-net algorithm are not as amenable to analysis. For comparison, nonrandomized GCR with a single fixed 1-hop-max clustering has a HT variance upper bound of $\frac{1}{n} d_{\max} \kappa^3 p^{-d_{\max}} \Theta(1)$, exponential in the max degree (and thus worse than 3-net when κ is modest). A summary of our variance bounds for HT estimators is given in Table 1 in Section 4, which also shows slightly improved bounds based on weighted variations of both algorithms.

1.1.2 Simulation analyses

Beyond theoretical results, we provide an extensive simulation-based analysis of various RGCR schemes on both synthetic and real-world networks. We first show dramatic variance reduction for the HT GATE

Table 1: A summary of bounds from Section 4 and Appendix B pertaining to the HT estimator of the GATE under various randomization designs. The RGCR results apply for *both* independent and complete randomization, while the GCR bounds do *not* hold for complete randomization because of possible positivity violations (Assumption 1). Each variance bound is up to a Θ(1) multiplicative constant

Clustering algorithm	**Scheme**	$\mathbb{P}[\boldsymbol{E}_i^1]$ **(lower bound)**	**Var**$[\hat{\boldsymbol{\mu}}(\mathbf{1})]$ **(upper bound)**
—	i.i.d.	$p^{d_{\max}+1}$	$\frac{1}{n}d_{\max}\kappa p^{-d_{\max}}$
3-net	GCR	p^{κ^6}	$\frac{1}{n}d_{\max}\kappa^5 p^{-\kappa^6}$
	RGCR	$\frac{p}{(d_{\max}+1)\kappa}$	—
$\boldsymbol{w}^*$-weighted	RGCR	$\frac{p}{\lambda^*}$	—
1-hop-max	GCR	$p^{d_{\max}+1}$	$\frac{1}{n}d_{\max}\kappa^3 p^{-d_{\max}}$
	RGCR	$\frac{p}{(d_{\max}+1)\kappa}$	$\frac{1}{n}d_{\max}^2\kappa^4 p^{-1}$
$\boldsymbol{w}^*$-weighted	RGCR	$\frac{p}{\lambda^*}$	$\frac{1}{n}\lambda^* d_{\max}\kappa^3 p^{-1}$

estimator used RGCR compared to GCR, bringing a useless variance ($\approx 10^{50}$) down to a potentially useful variance ($\approx 10^0$). We then vary many aspects of the simulation to understand the efficacy of RGCR-based experiments for GATE estimation. Specifically, we vary the structure of the underlying network, we vary the randomized clustering algorithm, we vary the possible weighting used in the algorithm, we vary the use of HT vs Hájek estimation, and we vary whether randomization is independent or complete.

1.2 Other contributions

1.2.1 Bounded geometry of empirical social networks

While bounded geometry assumptions play a central role in the previous theoretical analysis of GCR [27] and other recent works [32,33], the empirical growth rates of r-balls in social networks have not been well documented. The average degree, degree distribution, and path-length distribution of large-scale social networks have all been the subject of extensive empirical investigations [34–36], with the path length distribution being the central object of study in the large literature on "degrees of separation" inspired by Travers and Milgram [37]. Less attention has been given to the empirical structure of neighborhood sizes at different distances, though some intuition for the relationship between friend counts and friend-of-friend counts can be derived from the prior works [36,38,39].

Our empirical analysis, given in Appendix A, documents that for Facebook college social networks, κ is typically on the order of 25–50% of $d_{\max}$. As an aside, recall that the coefficient κ describes a worst-case coefficient. We observe that κ is typically pushed up by a few bad nodes where, e.g., a degree-1 node u is connected to a high degree node, making $|B_2(u)|/|B_1(u)|$ very high, and thus, κ is high for the graph as a whole. It is possible that new paths forward for studying estimators (and limit theorems [32]) on social networks may be more suited to an alternative formulation of restricted growth, not yet formulated.

1.2.2 Curse of large clusters

The GCR scheme suffers from large variance when nodes are connected to many clusters. A naive solution to this specific problem would be to partition the network into only a few, say K, clusters, where $K = O(1)$ is prespecified and independent of the size of the network. However, such an approach fails when the nodes' outcome exhibits homophily or some other global drift pattern such that nodes at a short distance have similar response outcomes. If there is a significant difference in the response of nodes in different clusters, and only a few clusters, then the observed difference $\hat{\tau} = \hat{\mu}(\mathbf{1}) - \mu(\mathbf{0})$ will be sensitive to this cluster-level variation, with additional variance incurred that does not then decay with the size of the network.

For the RGCR scheme, in Section 5, we show that this large cluster issue persists, and using a random clustering with large clusters (of size $\Theta(n)$, so $K = O(1)$) prohibits the HT estimator variance from converging to zero even as $n \to \infty$.

Specifically, we analyze a ring network with homophilous responses where the optimal balanced K-partitions are obvious. When selecting one of the optimal balanced K-partition uniformly at random in RGCR,-partition uniformly at random in RGCR, we show that

$$\mathbf{Var}[\hat{\tau}] \to \Omega\left(\frac{1}{K}\right)$$

as $n \to \infty$. Therefore, if $K = O(1)$, then we have $\mathbf{Var}[\hat{\tau}] = \Omega(1)$ even with $n \to \infty$. This result provides an important insight on the choice of random clustering used in RGCR scheme: the number of clusters in the output random clustering should increase with the number of individuals to let $\mathbf{Var}[\hat{\tau}] \to 0$ as $n \to \infty$, a necessary condition on the random clustering strategy. Consequently, various graph clustering algorithms such as spectral partitioning [40–42], balanced label propagation [43], or reLDG [10,44] are good clustering strategies for RGCR only if the number of output clusters grows with the size of the network.

1.2.3 A response model with network homophily

A specific innovation in our simulations is a rich graph-aware response model, exhibiting both degree-correlated responses and homophily in responses. Specifically, if two nodes have short graph distance, their responses tend to be close, resembling responses in many real-world settings [45] not captured in typical response models used in beyond-SUTVA simulations. Note that a failure to capture homophily in the response model can result in preferring a random clustering algorithms that generates few large clusters, concealing the issue of large clusters as discussed earlier and presented more fully in Section 5.

1.3 Paper roadmap

The remainder of this article is organized as follows. After preliminary definitions and related works in Section 2, we formally propose the RGCR scheme in Section 3. In Section 4, we develop key theoretical properties of RGCR (e.g., variance reduction) under HT estimation, with a focus on the two families of random clustering algorithms we consider in this work, the 3-net and 1-hop-max algorithms, as well as their weighted variants. We also discuss the bias of the related Hájek estimator under RGCR. In Section 5, we formalize a theory for the curse of large clusters, which provides a necessary condition on the random clustering algorithm for the variance to converge to zero as a network grows large. In Section 6, we provide extensive simulation results comparing different RGCR and GCR schemes. Section 7 concludes.

2 Preliminaries and related work

2.1 Network preliminaries

Throughout this work, we will consider interference in network settings as modeled by an undirected, unweighted network $G = (V, E)$, dubbed the *interference graph*, where the node set $V = \{1, 2, \ldots, n\}$ represents the units/individuals and E is the collection of edges that encode pairwise response dependencies that underly the interference. For each individual i, let N_i be the set of its neighbors on the network, and $d_i \triangleq |N_i|$ be its *degree*. We use $d_{\max} = \max_{i \in V} d_i$ to denote the maximum degree of all nodes in the network. A natural distance between a pair of nodes i and j on network G is the *shortest path distance* denoted as $\mathbf{dist}(i, j)$, i.e., the length of the shortest path connecting them. With a positive integer radius $r > 0$, we use $B_r(i) = \{j \in V | \mathbf{dist}(i, j) \leq r\}$ to denote the r-hop neighborhood of node i. For example, with $r = 1$, $B_1(r)$ contains node i itself and all its neighbors, and thus, $|B_1(i)| = d_i + 1$.

2.1.1 Restricted growth conditions

The conceptual notion of a (graph) metric with bounded geometry is very useful for considering the design of good clustering algorithms for social networks, as social networks arguably exhibit a version of bounded growth. As a motivating empirical observation, due to apparent tendencies toward clustering, the size of social network neighborhoods $|B_r(i)|$ tend to grow slower than $(d_{\max})^r$ in r [36].

There are two ways to operationalize this empirical tendency. First, by borrowing a definition from the literature on metric approximation [46], one could consider experimental designs that perform well under a condition of *bounded growth*, whereby there is a constant $\eta > 0$ such that

$$|B_{2r}(i)| \leq \eta |B_r(i)|, \quad \forall r \geq 1,$$

for all nodes i. Second, the original GCR work identified and developed results under a less restrictive metric property of *restricted growth* [27], which assumes there is a constant $\kappa > 0$ such that

$$|B_{r+1}(i)| \leq \kappa |B_r(i)|, \quad \forall r \geq 1,$$

for all nodes i. Notice that bounded growth implies restrictive growth. The constants η and κ here are called the *bounded growth* and *restrictive growth coefficients*, respectively. We emphasize that both of these definitions start at a radius of $r \geq 1$, and thus, we do not require any relationships to hold between B_0 and B_1 (otherwise we would have $\kappa = 1 + d_{\max}$), and it can be easily verified that $\kappa \leq d_{\max}$. Our goal, building on the initial analysis of GCR, is to exploit degree bounds and/or restricted growth structure to design algorithms that work provably well when $d_{\max}$ and/or κ are modest. We do not perform any analysis under bounded growth conditions ("η"), owing to the well-known fact that social networks have a very limited effective diameter [47], with the majority of node pairs appearing within a hop distance of six [35], limiting the utility of a bounded relationship between B_{2r} and B_r.

We note that a separate approach to causal inference under network interference has recently assumed metric growth conditions of a slightly different variety [33]. That work follows the recent work on limit theorems for network-dependent random variables where growth conditions appear as part of sufficient conditions [32].

2.1.2 Clustering

Throughout this work, we make broad use of the idea of a decomposition of a graph into *clusters*. A *clustering* is a partition of all nodes in the network into some nonoverlapping clusters, which is also referred as a *partition*. We denote a clustering as a vector $\mathbf{c} = [c_1, \ldots, c_n] \in \mathbb{R}^n$ such that nodes i and j belong to the same cluster if and only if $c_i = c_j$. Ideally clusters are internally densely connected while relatively separated from the rest of the network, though our definitions require no such thing.

The randomized clustering algorithms we analyze in this work stem from the literature on probabilistic approximations of graph metrics. Randomized graph decompositions have a rich history [48] originally driven by interests in distributed graph computations [49,50]. The algorithm we call *1-hop-max* is closely related to the Calinescu-Karloff-Raban (CKR) partitioning algorithm [51], developed as an approach to the 0-extension problem [52], a metric generalization of the multiway cut problem on graphs [53]. Our 1-hop-max algorithm runs the CKR algorithm with centers (or "terminals") selected at random, as is also done in the closely related Fakcharoenphol-Rao-Talwar (FRT) algorithm for metric approximation [54], and with a fixed radius of one. The other algorithm we consider, *randomized 3-net clustering*, comes from the related literature on metric approximation in bounded geometries [55] with applications to nearest neighbor search [46]. Graph cluster randomization with a fixed 3-net clustering was previously analyzed in the original work on GCR [27]. In the randomized setting of RGCR, we find 1-hop-max more amenable to theoretical analysis, while simulations indicate that RGCR with 1-hop-max and randomized 3-net do comparably well in diverse settings.

2.2 GATE estimation under exposure models

In many online and social settings, the presence of interference introduces bias in the estimation of global average treatment effects if a no-interference assumption, e.g., SUTVA, is incorrectly specified. More relaxed assumptions than SUTVA can be made that, if correct, can enable reasonable inference. As a first example, the class of *constant treatment response* (CTR) [23] assumptions identify, for each individual i, an *effective treatment mapping* g_i that captures equivalence classes of the global assignment vectors $\boldsymbol{z}$: if $g_i(\boldsymbol{z}_1) = g_i(\boldsymbol{z}_2)$ for two global assignments $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$, then $Y_i(\boldsymbol{z}_1) = Y_i(\boldsymbol{z}_2)$. SUTVA is a special case of CTR with $g_i(\boldsymbol{z}) = z_i$, i.e., where each individual's response depends only on the treatment assignment of itself.

The *neighborhood treatment response* [3] assumption is a CTR assumption that allows some treatment-based spillover effect: for any two global assignments $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$, $g_i(\boldsymbol{z}_1) = g_i(\boldsymbol{z}_2)$ if $\boldsymbol{z}_1[B_1(i)] = \boldsymbol{z}_2[B_1(i)]$, i.e., an individual's response depends only on the treatment assignment of itself and its neighbors. Consequently, individuals generate the same response as under the global treatment ($\boldsymbol{z} = \mathbf{1}$) assignment (a condition termed *network exposed to treatment*) if they and all their neighbors are assigned to the treatment group; similarly, they generate the same response as under the global control ($\boldsymbol{z} = \mathbf{0}$) assignment (a condition termed *network exposed to control*) if they and all their neighbors are assigned to the control group. Ugander et al. termed this pair of network exposure conditions as the *full-neighborhood exposure model*, and other more relaxed neighborhood exposure models have also been discussed [23,27].

In this work, we focus on the full-neighborhood exposure model due to it being the most restrictive neighbor exposure model. It greatly simplifies our theoretical analysis, relative to other more complicated exposure models, while still providing conclusions that generalize, at least at the level of intuition, to more relaxed neighborhood exposure models. Throughout this work, we use the events $E_i^{\boldsymbol{z}}$ specifically for full-neighborhood exposure, letting $E_i^{\boldsymbol{z}}$ denote the event (a subset of the global assignment vectors in $\{0,1\}^n$) where node i is network exposed to treatment ($\boldsymbol{z} = \mathbf{1}$) or control ($\boldsymbol{z} = \mathbf{0}$).

Both the HT and Hájek estimators require the following positivity assumption on the network exposure probabilities in order to be well defined.

Assumption 1. At every node i and for both $\boldsymbol{z} \in \{\mathbf{1}, \mathbf{0}\}$, the network exposure probability is positive: $\mathbb{P}[E_i^{\boldsymbol{z}}] > 0$.

Aronow and Samii have shown that assuming the exposure model is properly specified and a standard consistency assumption on the potential outcomes applies, the HT estimators are unbiased. They derive the variance of the HT estimators under these assumptions [3]. Specifically, the variance of the HT estimator of the mean outcome, $\hat{\mu}(\boldsymbol{z})$, is

$$\mathbf{Var}[\hat{\mu}(\boldsymbol{z})] = \frac{1}{n^2}\left[\sum_{i=1}^{n}\left(\frac{1}{\mathbb{P}[E_i^{\boldsymbol{z}}]} - 1\right)Y_i(\boldsymbol{z})^2 + \sum_{i=1}^{n}\sum_{j=1, j\neq i}^{n}\left(\frac{\mathbb{P}[E_i^{\boldsymbol{z}} \cap E_j^{\boldsymbol{z}}]}{\mathbb{P}[E_i^{\boldsymbol{z}}]\mathbb{P}[E_j^{\boldsymbol{z}}]} - 1\right)Y_i(\boldsymbol{z})Y_j(\boldsymbol{z})\right], \tag{4}$$

for $\boldsymbol{z} = \mathbf{1}, \mathbf{0}$, and the variance of GATE estimator is then

$$\mathbf{Var}[\hat{\tau}] = \mathbf{Var}[\hat{\mu}(\mathbf{1})] + \mathbf{Var}[\hat{\mu}(\mathbf{0})] - 2 \cdot \mathbf{Cov}[\hat{\mu}(\mathbf{1}), \hat{\mu}(\mathbf{0})], \tag{5}$$

where the covariance is

$$\mathbf{Cov}[\hat{\mu}(\mathbf{1}), \hat{\mu}(\mathbf{0})] = \frac{1}{n^2}\left[\sum_{i=1}^{n}\sum_{j=1, j\neq i}^{n}\left(\frac{\mathbb{P}[E_i^{\mathbf{1}} \cap E_j^{\mathbf{0}}]}{\mathbb{P}[E_i^{\mathbf{1}}]\mathbb{P}[E_j^{\mathbf{0}}]} - 1\right)Y_i(\mathbf{1})Y_j(\mathbf{0}) - \sum_{i=1}^{n} Y_i(\mathbf{1})Y_i(\mathbf{0})\right]. \tag{6}$$

2.3 Graph cluster randomization

The full-neighborhood exposure probabilities $\mathbb{P}[E_i^{\boldsymbol{z}}]$, as well as the joint exposure probabilities $\mathbb{P}[E_i^{\boldsymbol{z}_1} \cap E_j^{\boldsymbol{z}_2}]$, are properties of the experimental design. With node-level independent randomization, where we assign each node into the treatment or control group independently, the full-neighborhood exposure probability is exponential to the degree, and thus, it can be extremely small. The variance of HT estimator is a monotone decreasing function in any single exposure probability, meaning that small probabilities beget large variances. As a result, the HT estimator variance (5) can be exponentially large in the max degree $d_{\max}$ and not practical [27].

To overcome the issue of exponential variance under independent assignment, Ugander et al. proposed to randomize at the cluster level, the GCR scheme [27]: with a clustering $\boldsymbol{c}$ of the network, one can jointly assign all nodes in each cluster into the treatment or control group. A definition of HT estimator for $\mu(\boldsymbol{z})$ was given in the introduction, but restating it more formally in the context of GCR,

$$\hat{\mu}_{\boldsymbol{c}}(\boldsymbol{z}) = \frac{1}{n}\sum_i \frac{\mathbb{1}[E_i^{\boldsymbol{z}}]\cdot Y_i}{\mathbb{P}[E_i^{\boldsymbol{z}}|\boldsymbol{c}]}, \tag{7}$$

where the subscript indicates that this estimator is based on GCR with clustering $\boldsymbol{c}$. Under this design, the exposure probability of each node is exponential not in its degree, but in the number of clusters intersecting with its 1-hop neighborhood and thus should reduce the variance in the HT estimators if a reasonable clustering is in use. Specifically, Ugander et al. show that if the clustering is generated from the 3-net clustering algorithm, and the graph satisfies the restricted growth condition with coefficient κ, then the variance is upper bounded by a linear function of the maximum degree of the graph equation (2).

Despite significant variance reduction compared with node-level independent randomization, the GCR scheme has one main disadvantage: the variance is still enormous when there are small exposure probabilities. With a single fixed clustering of the network, a node may be "unlucky" (Figure 1) and directly connect to many clusters. For such a node to be network exposed to treatment or control, all the adjacent clusters have to be assigned into the treatment or control group respectively, making the exposure probability exponentially small.

A naive solution to this issue would be to partition the network into only a few clusters, so each node can be adjacent to at most the number of clusters in the clustering. However, this solution is problematic due to two concerns. First, partitioning the network into few but large clusters makes the estimated result very sensitive to homophily, as discussed in Section 5.

Second, with just a few clusters, independent randomization at the cluster level may cause significant imbalance in treatment/control assignment. For example, with a bisection of the network, if each cluster is assigned independently into the treatment group with probability 1/2, then there is a 25% chance that both clusters (and consequently all nodes in the network) are assigned into the treatment group, and we collect no information about the control condition. To maintain balance with two clusters, one would need to assign the clusters to opposite conditions (treatment, control), the method of complete randomization. However, a secondary disadvantage of the GCR scheme is that it is incompatible with complete randomization at the cluster level due to potential violation of the positivity assumption (Assumption 1). For example, with GCR with few clusters and complete randomization, a node connected to all the clusters will always have some neighbors in treatment and some in control, making it impossible for that node to be full-neighborhood exposure to either treatment or control.

3 Randomized GCR

We now present the RGCR experimental design and its aligned analysis. Formally, let $\mathcal{P}$ be a *random clustering generator*, i.e., an algorithm whose output $\boldsymbol{C}$ is a clustering of the input graph, and the output is random. Without ambiguity of notation, we also use $\mathcal{P}$ to denote the distribution of the randomly generated clustering, i.e., $\mathcal{P}(\boldsymbol{c})$ is the probability of the clustering $\boldsymbol{c}$ being generated. Both the design and analysis of the RGCR scheme are tailored to the random clustering generator $\mathcal{P}(\cdot)$, or equivalently, the resulting distribution of random clusterings.

3.1 Design

With a random clustering generator $\mathcal{P}$, the experimental design is based on a two-step process. First, we realize a clustering $\boldsymbol{c}$ from the random clustering $\boldsymbol{C}$. Second, like in the GCR scheme, we perform treatment/control assignment at the cluster level, jointly assigning all nodes within each cluster of $\boldsymbol{c}$ into the treatment group with probability p, or into control otherwise.

In the second step of the aforementioned cluster-level randomization, GCR assign each cluster using *independent randomization*. For RGCR, besides independent randomization, we also consider *complete randomization*, where we further introduce stratification. In the case of $p = 1/2$, we first stratify the clusters of $\boldsymbol{c}$ into pairs, by size (measured by the number of nodes): the two largest clusters are a pair, the third and fourth largest clusters are a pair, and so on. We then assign each pair of clusters together, with one into the treatment and the other into the control group. Complete randomization guarantees an equal number of clusters in treatment and control, thereby roughly balancing the number of individuals as well. Stratification further tightens this balance. Complete randomization with other values of p is implemented analogously.

Balance guarantees are especially important when the clustering contains only few clusters. For example, in the case of a clustering formed by a graph bisection, under independent randomization, the probability that both clusters are assigned into the treatment group or both assigned into the control group is 0.5, an unpleasant scenario where we collect information about only the treatment group or only the control group. In contrast, with complete randomization we always have one cluster assigned to the treatment group and the other to the control group. Moreover, complete randomization may increase $\mathbb{P}[E_i^{\mathbf{1}} \cap E_j^{\mathbf{0}}]$ for distant nodes, which increases the covariance of $\hat{\mu}(\mathbf{1})$ and $\hat{\mu}(\mathbf{0})$ and thus further reduces variance according to equations (6) and (5). Such variance reduction is also observed in our simulations in Section 6.

Under GCR, complete randomization can violate the positivity assumption. For example, if a node i is adjacent to a pair of clusters that are determined to be oppositely assigned into the treatment and control group, then it is impossible for node i to be full neighborhood exposed to treatment or control, i.e., $\mathbb{P}[E_i^{\mathbf{1}}] = \mathbb{P}[E_i^{\mathbf{0}}] = 0$. Without positivity, the HT estimator (equation (7)) is ill-defined. For RGCR, Section 4.2 shows that as a consequence of Theorem 4.2, RGCR using our randomized 3-net and 1-hop max clustering algorithms always satisfies node-level positivity for the full-neighborhood exposure condition (and related fractional conditions).

3.2 Analysis

With both independent or complete randomization, the exposure probability of each node i conditioned on the generated clustering $\boldsymbol{C} = \boldsymbol{c}$, i.e., $\mathbb{P}[E_i^{\boldsymbol{z}} | \boldsymbol{C} = \boldsymbol{c}]$, can be computed as in the GCR scheme. While we focus on full-neighborhood exposure throughout this work, we note that this observation applies to, e.g., partial neighborhood exposure conditions [27] as well. In the analysis phase of a RGCR experiment, we use the exposure probabilities unconditional on the clustering in use, which only depends on the clustering distribution $\mathcal{P}$. Formally, since the random clustering in use is generated from the distribution $\mathcal{P}$, the network exposure probability of each node i, by to the law of total expectation, is

$$\mathbb{P}[E_i^{\boldsymbol{z}}|\mathcal{P}] = \sum_{\boldsymbol{c}} \mathcal{P}(\boldsymbol{c})\mathbb{P}[E_i^{\boldsymbol{z}}|\boldsymbol{c}] = \mathbb{E}\boldsymbol{c} \sim \mathcal{P}[\mathbb{P}[E_i^{\boldsymbol{z}}|\boldsymbol{c}]]. \tag{8}$$

Consequently, the HT estimators are as follows:

$$\hat{\mu}_{\mathcal{P}}(\boldsymbol{z}) = \frac{1}{n}\sum_i \frac{\mathbb{1}[E_i^{\boldsymbol{z}}]\cdot Y_i}{\mathbb{P}[E_i^{\boldsymbol{z}}|\mathcal{P}]}, \tag{9}$$

where $\boldsymbol{z} = \mathbf{0}$ or $\boldsymbol{z} = \mathbf{1}$, and $\hat{\tau}_{\mathcal{P}} = \hat{\mu}_{\mathcal{P}}(\mathbf{1}) - \hat{\mu}_{\mathcal{P}}(\mathbf{0})$ is the HT estimator of the GATE τ. Here, the subscript $\mathcal{P}$ emphasizes that the estimator is based on a distribution of clusterings. The Hájek estimators for $\mu(\mathbf{0})$, $\mu(\mathbf{1})$, and τ are analogous.

3.3 Putting design and analysis together

There are a number of important challenges in going from using a single fixed clustering to using a random clustering. Most concretely, in the design phase, one needs to be able to efficiently generate a single random clustering to launch an experiment. As a complementary challenge in the analysis phase, HT and Hájek estimators require per-node unconditional exposure probabilities, which may be more or less difficult to compute, depending on the randomized clustering algorithm used. We discuss and compare properties of different random clustering strategies in the following section, showing that both randomized 3-net and 1-hop-max are good algorithms in these regards.

4 Properties of RGCR

In this section, we analyze the properties of the RGCR scheme. We focus on the Horvitz–Thompson (HT) estimator due to its theoretical amenability, while some important insights on the Hájek estimator are discussed at the end.

Since the RGCR scheme requires a randomized clustering strategy, we consider two initial algorithms: randomized 3-net, a randomized version of the 3-net algorithm considered in the original analysis of the GCR scheme, and 1-hop-max, a new randomized clustering algorithm similar to 3-net but more easily amenable to a rigorous analysis. In Appendix B, we consider the weighted versions of these two algorithms, which introduce node-level flexibility and can effectively balance the exposure probabilities of high- and low-degree nodes, addressing an additional imbalance found in the two basic algorithms. The goal of this section is to provide an analysis of how RGCR can lead to considerable variance reduction when compared with the vanilla GCR scheme based on a single clustering. All but the simplest proofs are removed to Appendix C.

We summarize the results of this section in Table 1 and highlight some important observations. First, for each clustering algorithm, by using GCR with a single fixed clustering, the variance of the HT estimator is upper bounded by an exponential function of either $d_{\max}$ or κ. Note that both quantities can be large in real-world networks, resulting in the huge variance in the original GCR scheme. In contrast, with RGCR, the variance is upper bounded by a polynomial function of $d_{\max}$ and κ. If the graph has bounded degree $d_{\max}$ but the growth is not further restricted, then we still have $\kappa < d_{\max}$. Therefore, the RGCR scheme can significantly reduce the estimator variance compared with GCR, both with and without restricted growth.

Second, we highlight that variance reduction is achieved primarily by obtaining a much larger exposure probabilities, which are the inverse weights in the HT estimator and play a similar role in the Hájek estimator. With a fixed clustering, a node can be at the boundary of a cluster (Figure 1), making it adjacent to many clusters and furnishing an exponentially small exposure probability. However, with RGCR, such exponentially small probabilities are "washed out" by averaging across clusterings.

Finally, for each random clustering algorithm considered, complete randomization is valid for RGCR, i.e., positivity (Assumption 1) is satisfied. In contrast, the positivity assumption is generally violated in GCR with complete randomization. The results for RGCR summarized in Table 1 apply for both independent and complete randomization, while those for GCR apply only for independent randomization.

Beside extensive analysis on the HT estimator, we also present some key properties of the Hájek estimator under the GCR and RGCR schemes. Compared to the HT estimator, the Hájek estimator enjoys much lower variance due to the self-normalization, while a potential drawback of introducing bias. How bad can this bias be? We find that the Hájek estimator is unbiased under GCR and RGCR if the individual treatment effect $\tau_i = Y_i(\mathbf{1}) - Y_i(\mathbf{0})$ is constant across all nodes. In practice, treatment effects can be expected to be nonconstant, which motivates us to consider simulations in Section 6 of nonconstant individual treatment effects to study the bias of the Hájek estimator.

4.1 Randomized 3-net and 1-hop-max algorithms

We now outline the two random clustering algorithms we consider in detail. For notation brevity, our analysis is always conditioned on the distribution of random clusterings in focus, unless stated otherwise.

4.1.1 Randomized 3-net

The first algorithm in consideration is the 3-net clustering, which is used in the original analysis of the graph cluster randomization scheme [27]. Here, we assume that a 3-net clustering is generated from a random ordering of all nodes, and thus, its output is random, while such randomness was not exploited in any part of the analysis of vanilla GCR, which was conditional on a single clustering outputted by the algorithm.

Algorithm 1. 3-net clustering.

```
Input: Graph $G = (V, E)$
Output: Graph clustering $\boldsymbol{c} \in \mathbb{R}^n$
$\pi \leftarrow$ generate a uniformly random total ordering of all nodes
$S \leftarrow \varnothing$, unmark all nodes
for $i \in \pi$ do
    if $i$ is unmarked then
        $S \leftarrow S \cup \{i\}$
        for $j \in B_2(i)$ do
            mark node $j$ if it is unmarked yet
for $i \in V$ do
    $c_i \leftarrow \arg\min\{j \in S, j \to \mathbf{dist}(i, j)\}$, i.e., the id of the node in $S$ with shortest graph distance to $i$ (arbitrary tie breaking)
return $\boldsymbol{c}$
```

Algorithm 2. 1-hop-max clustering.

```
Input: Graph $G = (V, E)$
Output: Graph clustering $\boldsymbol{c} \in \mathbb{R}^n$
for $i \in V$ do
    $X_i \leftarrow \mathcal{U}(0, 1)$
for $i \in V$ do
    $c_i \leftarrow \max([X_j \text{ for } j \in B_1(i)])$
return $\boldsymbol{c}$
```

Formally, the randomized 3-net clustering algorithm is given in Algorithm 1, which consists of three major steps. First, we generate a total ordering of all nodes sampled uniformly over all permutations. Second, construct a maximal distance-3 independent set of the network (line 2–7) using a greedy algorithm proceeding according to the total ordering generated in line 1. We call each node in the independent set a *seed* node. Next we assign every node in the network to the seed node with smallest graph distance, with ties broken by some arbitrary rule. These steps return a clustering partition.

In the returned clustering, since the seed nodes form a distance-3 independent set, any 1-hop neighbors of a seed node will be assigned to the seed. Therefore, the seeds nodes are guaranteed to be in the interior of a cluster, not connecting to any nodes in a different cluster. Consequently, the returned clustering consists of node-neighborhood clusters known to form relatively good clusters (in terms of edges cut) in real-world networks [56,57].

A potential disadvantage of 3-net clustering algorithm is the parallel runtime. Even though parallel algorithms have been developed for the random maximal independent set problem [49,58], the runtime still increases with the size of the network, and thus, it is generally slow to sample a random 3-net clustering on a very large network, even by more complicated means.

4.1.2 1-hop-max

As a second algorithm for RGCR, we propose *1-hop-max*, given in Algorithm 2. This algorithm consists of two steps. First, every node i independently generates a random number from the uniform distribution on $(0, 1)$. Second, for every node i, find the maximum of the generated numbers within node i's 1-hop neighborhood. The unique numbers define the clustering: nodes with the same 1-hop-maximum form a single cluster.

Similar to the 3-net algorithm, the clustering returned by the 1-hop-max algorithm contains neighborhood-like clusters: every cluster is associated with a center node. On the other hand, the 1-hop-max algorithm has a much faster parallel runtime. Formally, we have the following result in terms of the *depth* (i.e., length of longest chain in the computation dependency graph) [59], a key constraint in parallel computing.

Theorem 4.1. *The* 1*-hop-max algorithm, Algorithm* 2, *has* $O(\log(d_{\max}))$ *depth.*

Both algorithms require $O(m)$ work, but the above depth compares favorably to the 3-net algorithm's depth of at least $O(\log(n)^2)$, according to the analysis in [58], which finds a maximal independent set, not a distance-3 maximal independent set.

4.2 Network exposure probabilities

The network exposure probabilities $\mathbb{P}[E_i^{\mathbf{z}}|\mathcal{P}]$ of these algorithms are key parts of the HT and Hájek GATE estimators under RGCR.

4.2.1 Bounds

Before discussing how to compute or estimate these probabilities, we first show a simple but useful lower bound of the full neighborhood exposure probabilities when using 3-net or 1-hop-max random clustering generator. This result is crucial in both the analysis of a Monte Carlo method for estimating the probabilities in Section 4.2.2 and the variance analysis in Section 4.3.

Theorem 4.2. *Using either 3-net or 1-hop-max random clustering on a graph with restricted growth coefficient* κ, *using either independent or complete randomization at the cluster level, the full-neighborhood exposure probabilities for any node i satisfy*

$$\mathbb{P}[E_i^{\mathbf{1}}|\mathcal{P}] \geq \frac{p}{|B_2(i)|} \geq \frac{p}{(1+d_{\max})\kappa}, \quad \mathbb{P}[E_i^{\mathbf{0}}|\mathcal{P}] \geq \frac{1-p}{|B_2(i)|} \geq \frac{1-p}{(1+d_{\max})\kappa}.$$

A detailed proof is given in Appendix C, while the high-level idea is as follows. If a node i is ranked first within $B_2(i)$ in a 3-net clustering algorithm (or generated the largest number in the 1-hop-max algorithm), which happens with probability $1/|B_2(i)|$, then all its 1-hop neighbors are guaranteed to be in the same cluster as node i, and thus, it is definitely network exposed to either treatment or control.

Several remarks are in order on the aforementioned result. First, this lower bound is much more favorable than an analogous lower bound for the GCR scheme. With GCR, 3-net clustering, and independent randomization (but not complete randomization), we have $\mathbb{P}[E_i^{\mathbf{1}}|\mathcal{G}] \geq p^{d_{\max}}$ in general and $\mathbb{P}[E_i^{\mathbf{1}}|\mathcal{G}] \geq p^{\kappa^6}$ under a restricted growth condition [27]. That lower bound is exponentially small in the restrictive growth parameter κ. In real-world networks, κ can be of magnitude of 100, making the exposure probabilities impossibly small. In contrast, with RGCR, the exposure probability is lowered bounded by a polynomial function of $d_{\max}$ and κ.

As a second remark, these lower bounds also hold when we consider a partial neighborhood exposure model. If a node is full-neighborhood exposed, it must also be partial-neighborhood exposed, and thus, the partial-neighborhood exposure probability of each node is no lower than that for full-neighborhood exposure.

As a third remark, another significant implication of Theorem 4.2 is that it provides a positive lower bound on the node-level exposure probabilities, making complete randomization feasible. Note that complete randomization is not feasible for the GCR scheme due to violation of the positivity assumption. However, for RGCR scheme, according to Theorem 4.2, even under complete randomization, the exposure probability of each node is guaranteed to be positive.

The exposure probability lower bound in Theorem 4.2 is obtained by solely considering scenario when a node generates the largest number in its 2-hop neighborhood. Actually, one can obtain an improved lower bound from more careful consideration on node's ranking among its 2-hop neighborhood.

Theorem 4.3. *With 1-hop-max random clustering algorithm and independent randomization at the cluster level, if* $|B_2(i)| - d_i \geq 1/(1-p)$, *then the full-neighborhood exposure probabilities for any node i satisfy*

$$\mathbb{P}[E_i^{\mathbf{1}}|\mathcal{P}] \geq \frac{1}{|B_2(i)|} \cdot \frac{p}{1-p}, \quad \mathbb{P}[E_i^{\mathbf{0}}|\mathcal{P}] \geq \frac{1}{|B_2(i)|} \cdot \frac{1-p}{p}.$$

The proof of this result involves a more carefully analysis, and for $p = 1/2$, the difference between the lower bounds in Theorems 4.3 and 4.2 is merely a factor of 2.

4.2.2 Computation and estimation

Computing the exact network exposure probabilities can be challenging as it potentially requires considering an exponential number of different clusterings in equation (8). More formally, Theorem 4.4 shows that with 3-net clustering, computation of the exact exposure probability for a single node is NP-hard. The proof is given in Appendix C.

Theorem 4.4. *For the 3-net random clustering algorithm, using either independent or complete randomization at the cluster level, exact computation of the full-neighborhood exposure probability for a node in an arbitrary graph is NP-hard.*

Note that even though we do not have an analogous rigorous proof for the 1-hop-max clustering strategy, we expect the analogous exposure probability computations to also be NP-hard.

Despite the difficulty of exactly computing the probabilities, they can be efficiently estimated using a relatively straight-forward Monte Carlo method with theoretical guarantees. The procedure begins by generating K clusterings $\{\mathbf{c}^{(k)}\}_{k=1}^{K}$ from our randomized clustering algorithm and compute the exact exposure probability of each node under each clustering. The estimator of the exposure probability is then

$$\hat{\mathbb{P}}[E_i^{\mathbf{z}}|\mathcal{P}] = \frac{1}{K}\sum_{k=1}^{K}\mathbb{P}[E_i^{\mathbf{z}}|\mathbf{c}^{(k)}]. \tag{10}$$

We then have the following result on the mean square error (MSE) of relative error in this Monte Carlo estimator.

Theorem 4.5. *For either 3-net or 1-hop-max random clustering algorithm, and with K Monte-Carlo trials and any node i, the relative error of the Monte-Carlo estimator is upper bounded in MSE as follows*:

$$\mathbb{E}\left[\frac{\hat{\mathbb{P}}[E_i^{\mathbf{1}}|\mathcal{P}] - \mathbb{P}[E_i^{\mathbf{1}}|\mathcal{P}]}{\mathbb{P}[E_i^{\mathbf{1}}|\mathcal{P}]}\,\middle|\,\mathcal{P}\right]^2 \le \frac{|B_2(i)|}{Kp}.$$

The proof is given in Appendix C, which is obtained from the fact that the ground-truth exposure probability is bounded away from 0 as is shown in Theorem 4.2.

Given this MSE guarantee, it is natural to use the estimated exposure probabilities as the inverse weights in an HT estimator. A potential issue is the possible violation of the positivity assumption under complete randomization. Even though the ground-truth exposure probability is bounded away from 0 per Theorem 4.2, the estimated probability (equation (10)) can be zero: it is possible that for some node i, $\mathbb{P}[E_i^{\mathbf{1}}|\mathbf{c}^{(k)}] = 0$ for all the generated clusterings, and thus, $\hat{\mathbb{P}}[E_i^{\mathbf{1}}|\mathcal{P}] = 0$. A zero probability in the denominator would make the HT estimate ill defined. A fix to this issue is to use stratified sampling in generating the clustering samples.

To stratify our Monte Carlo estimator, we generate Kn samples $\{\mathbf{c}^{(k,i)}\}$ with $k \in \{1, 2, \ldots, K\}$ and $i \in \{1, 2, \ldots, n\}$ such that, if the 3-net clustering is in use, then the clustering $\mathbf{c}^{(k,i)}$ is based on a random node ordering conditional on node i being ranked first among all nodes. Analogously, if the 1-hop-max clustering is in use, then in the generation of clustering $\mathbf{c}^{(k,i)}$, node i generates the largest X_i among all nodes. Consequently, under clustering $\mathbf{c}^{(k,i)}$, node i is guaranteed to be the center of a cluster and thus, $\mathbb{P}[E_i^{\mathbf{1}}|\mathbf{c}^{(k,i)}] = p$. Now the network exposure probability of node i is estimated as follows:

$$\hat{\mathbb{P}}[E_i^{\mathbf{z}}|\mathcal{P}] = \frac{1}{nK}\sum_{k=1}^{K}\sum_{j=1}^{n}\mathbb{P}[E_i^{\mathbf{z}}|\mathbf{c}^{(k,j)}].$$

In total, each node i is "favored" exactly K times among the Kn samples, and we have

$$\hat{\mathbb{P}}[E_i^{\mathbf{1}}|\mathcal{P}] \ge \frac{p}{n} > 0.$$

Besides a guarantee of positivity in the estimated exposure probabilities, this stratified sampling technique is also effective at reducing variance in the estimation. Therefore, when computationally feasible to sample at least n clustering samples, this stratified sampling method should be strictly preferred over independent sampling.

As a final but important note on probability computation and estimation, we point out that the potential computational bottleneck of generating K clusterings when using RGCR should not pose practical concerns. First, we highlight that the exposure probabilities are needed only in the analysis phase but not the design phase. To launch an experiment, it suffices to generate a single clustering from a randomized algorithm and use it in assigning individuals to treatment or control; after the experiment has been launched, we can later sample other random clusterings to estimate the exposure probabilities. Second, we note that the estimated exposure probabilities can be shared across experiments as long as the interference network remains unchanged. In practice, with hundreds of A/B testings running at the same time, practitioners only need to estimate the exposure probabilities once.

4.3 Variance of HT estimators

We now analyze the variance of the Horvitz–Thompson (HT) estimator with RGCR. We show that, with 1-hop-max clustering, the variance is upper bounded by a polynomial function in both the maximum degree $d_{\max}$ and the restricted growth parameter κ, which also decays as $n \to \infty$.

We first present a useful property of the randomized 1-hop-max clustering algorithm, the local dependence, which distinguishes it from 3-net clustering.

Lemma 4.1. *With 1-hop-max random clustering algorithm, for any node i, the joint distribution of $\mathbf{C}_{B_1(i)}$, i.e., the clusterings of all nodes in $B_1(i)$, depends only on the structure of the graph induced on the node set $B_2(i)$.*

Proof. Since the clustering of every node is $C_j = \max\{X_{j'} : j' \in B_1(j)\}$ with $X_{j'} \sim \mathcal{U}(0, 1)$, the joint distribution of $[C_j]_{j \in B_1(i)}$ depends only on the structure of the graph induced on the node set $B_2(i)$ and is independent of the rest of network. □

With this local dependence property, now we present the following result on the variance of mean-outcome HT estimator.

Theorem 4.6. *For RGCR with a 1-hop-max clustering, if every node's responses are within* $[0, \bar{Y}]$, *then*

$$\mathbf{V\ ar}[\hat{\mu}_{\mathcal{P}}(\mathbf{1})] \leq \frac{\bar{Y}^2}{n^2} \sum_{i=1}^{n} \frac{|B_4(i)|}{\mathbb{P}[E_i^{\mathbf{1}}|\mathcal{P}]}$$

for both independent and complete cluster-level randomization.

As an intuition for this result, by local dependence, we have that the full-neighborhood exposure events $E_i^{\mathbf{1}}$ and $E_j^{\mathbf{1}}$ of two nodes i and j become independent (or negatively correlated) events if their graph distance is sufficiently big. This observation limits many cross-terms of the variance formula, equation (4), yielding an upper bound. A formal proof is given in Appendix C. As a consequence of Theorem 4.6, we have the following main result.

Theorem 4.7. *For RGCR with 1-hop-max clustering on a graph with maximum degree $d_{\max}$ and restricted growth coefficient κ, If every node's responses are within* $[0, \bar{Y}]$, *then*

$$\mathbf{Var}[\hat{\mu}_{\mathcal{P}}(\mathbf{1})] \leq \frac{1}{n} \cdot \bar{Y}^2 (d_{\max} + 1)^2 \kappa^4 p^{-1},$$

for both independent and complete cluster-level randomization.

Proof. From Theorem 4.6, we have

$$\mathbf{Var}[\hat{\mu}_{\mathcal{P}}(\mathbf{1})] \leq \frac{\bar{Y}^2}{n^2} \sum_{i=1}^{n} \frac{|B_4(i)|}{\mathbb{P}[E_i^{\mathbf{1}}|\mathcal{P}]} \leq \frac{\bar{Y}^2}{n^2} \sum_{i=1}^{n} \left[\frac{|B_2(i)|}{p} \cdot |B_4(i)| \right] \leq \frac{1}{n} \cdot \bar{Y}^2 (1 + d_{\max})^2 \kappa^4 p^{-1},$$

where the second inequality is due to $\mathbb{P}[E_i^{\mathbf{1}}|\mathcal{P}] \geq \frac{p}{B_2(i)}$ and the third inequality is due to $|B_r(i)| \leq (1 + d_{\max})\kappa^{r-1}$. □

This upper bound is to be compared with equation (2), the variance upper bound when using a single fixed clustering, which is exponential to the restrictive growth coefficient κ. In contrast, if a random graph clustering is used, the upper bound is a polynomial function of κ. This result provides a strong theoretical justification of variance reduction from using random graph partitioning in GCR.

From the variance of the mean outcome estimator $\hat{\mu}_{\mathcal{P}}(\mathbf{1})$, we can obtain the following variance upper bound on the GATE estimator $\hat{\tau}_{\mathcal{P}}$.

Theorem 4.8. *For RGCR with 1-hop-max clustering on a graph with maximum degree $d_{\max}$ and restricted growth coefficient κ, if every node's responses are within* $[0, \bar{Y}]$, *then*

$$\mathbf{Var}[\hat{\tau}_{\mathcal{P}}] \leq \frac{2}{n} \cdot \bar{Y}^2 (d_{\max} + 1)^2 \kappa^4 (p^{-1} + (1 - p)^{-1}),$$

for both independent and complete cluster-level randomization.

All of our analysis thus far has been nonasymptotic (finite-n) results. As such, we have not assumed that $d_{\max}$ or κ are fixed in n. As a corollary of Theorem 4.8 then, we have the following sufficient condition for convergence of the HT estimator of the GATE, which extends beyond the regime of bounded-degree graphs.

Theorem 4.9. *Let G_n be a sequence of graphs on n nodes with maximum degree $d_{\max}$ and restricted growth coefficient κ both possibly dependent on n. Let all responses be within $[0, \bar{Y}]$. Then for RGCR with 1-hop-max, a fixed cluster-level randomization probability p, and either:*

- *κ fixed, $d_{\max} = o(n^{1/2})$ or*
- *κ, $d_{\max} = o(n^{1/6})$,*

we have $\mathbf{Var}[\hat{\tau}_{\mathcal{P}}] \to 0$ *as $n \to \infty$, for both independent and complete cluster-level randomization.*

If κ is fixed, then the analogous sufficient condition for GCR (from equation (2)) requires $d_{\max}$ to be only $o(n)$. But if $d_{\max}$ and κ are of similar order – as Appendix A suggests that they often are empirically in social networks – the analogous GCR sufficient condition requires $d_{\max}$ to be $o(\log n)$, a significantly stronger requirement than the $o(n^{1/6})$ requirement under RGCR.

The proof of the variance upper bound in Theorem 4.8 does not apply to RGCR under a randomized 3-net clustering. An analogous analysis breaks down because local dependence (Lemma 4.1) does not hold for the 3-net clustering algorithm. Specifically, the distribution of $\boldsymbol{C}_{B_1(i)}$ depends on the structure of the whole network. Despite the lack of strictly local dependence, we still expect randomized 3-net clustering to undergo similar variance reduction when using randomized clustering versus a single fixed clustering. In Section 6, we see in simulations that the variance of RGCR with 3-net clustering is much lower than with GCR, and it is in fact lower than that of RGCR with 1-hop-max clustering.

4.4 Weighted randomized 3-net and 1-hop-max clusterings

A drawback of both the 3-net and 1-hop-max clustering algorithms (and shared by many other existing approaches to clustering) is an implicit disadvantage for high-degree nodes: compared to low-degree nodes they are invariably connected to many more clusters and thus have much smaller exposure probabilities. This phenomenon is supported by Theorem 4.2, where we showed a exposure probability lower bound that decreases with the size of its two-hop neighborhood. Per Theorem 4.6, the smallest exposure probabilities (and thus, those for high degree nodes) dominate the variance in HT estimators.

To offset the outsized contribution of high-degree nodes to the variance, Appendix B develops and analyzes weighted versions of both random clustering algorithms, introducing additional node-level flexibility to adjust and balance the exposure probability of nodes. In particular, we can choose to prioritize high-degree nodes in these weighted clustering algorithms. A high-level summary of the properties of the weighted algorithms is included in Table 1 at the start of this section, with details given in Appendix B.

4.5 HT vs Hájek estimation

While we focus our analysis of GATE estimation on HT estimators, some of our results extend to the related Hájek estimator [31], also called the self-normalized estimator [60,61], of the mean outcome

$$\tilde{\mu}(\boldsymbol{z}) = \left(\sum_{i=1}^{n} \frac{\mathbb{1}[E_i^{\boldsymbol{z}}]}{\mathbb{P}[E_i^{\boldsymbol{z}}]}\right)^{-1} \sum_{i=1}^{n} \frac{\mathbb{1}[E_i^{\boldsymbol{z}}] \cdot Y_i}{\mathbb{P}[E_i^{\boldsymbol{z}}]}, \tag{11}$$

with the Hájek GATE estimator taking the form $\tilde{\tau} = \tilde{\mu}(\mathbf{1}) - \tilde{\mu}(\mathbf{0})$. Notice that the Hájek and HT estimators utilize the same exposure probabilities for a given design. The Hájek estimator is typically biased but often preferable to the HT estimator under a strong bias–variance trade-off.

The Hájek estimator is much less amenable to theoretical analysis than the HT estimator, and so our analysis of the Hájek estimator of the GATE is much less extensive. Both estimators depend on the same exposure probabilities, so the general analysis of the exposure probabilities under randomized 3-net and 1-hop-max sheds light on the behavior of the Hájek estimator as well. Regardless of these theoretical difficulties, the Hájek estimator has many intuitive advantages as a GATE estimator, relative to the HT estimator. We catalog these intuitive advantages briefly and also contribute a possibly useful observation about the Hájek GATE estimator: it is unbiased when the individual treatment effect is constant.

As a first generic advantage of the Hájek GATE estimator over the HT estimator, the value of the Hájek estimator of a mean outcome, $\tilde{\mu}(\boldsymbol{z})$, is bounded within the range of all units' responses, due to the estimator having the form of a convex combination of the responses of all exposed units (weighted by the inverse exposure probability). As a result, when the responses are bounded, then the Hájek estimator variance is immediately bounded. In contrast, the value of the HT estimator can be far outside this range of responses, due to its sensitivity to extremely small exposure probabilities, and the HT variance can then be much, much larger as well.

As a second advantage, the variance of the Hájek estimator is invariant to a shift in unit responses: if every unit's response is increased or decreased (additively) by a constant, then the variance of Hájek estimator remains unchanged. This, again, is not a property of the HT estimator for the same estimand.

As a third advantage, for a given outcome $\boldsymbol{Z} = \boldsymbol{z}$, the Hájek estimator depends only on the relative value of network exposure probabilities of all nodes and is invariant to their absolute value. Specifically, for two sets of node-wise exposure probabilities $\{\mathbb{P}_1[E_i^{\boldsymbol{z}}]\}_{i=1}^n$ and $\{\mathbb{P}_2[E_i^{\boldsymbol{z}}]\}_{i=1}^n$, which may come from two different experiment designs, if there is a constant c such that $\mathbb{P}_1[E_i^{\boldsymbol{z}}] = c \cdot \mathbb{P}_2[E_i^{\boldsymbol{z}}]$ for every node i, then for a given outcome $\boldsymbol{Z} = \boldsymbol{z}$, the two sets of exposure probabilities yield the same Hájek estimator. This property might imply an advantage for the RGCR scheme compared with GCR in Hájek estimation, as the RGCR scheme yields a more uniform network exposure probability of all nodes: RGCR tends to increase small probabilities of "unlucky" nodes and decrease large probabilities of "lucky" nodes compared to a GCR scheme with a fixed clustering (Figure 1).

Compared with the HT estimator, a potential drawback of the Hájek estimator, widely known in the literature, is the potential issue of bias, i.e., $\mathbb{E}[\tilde{\mu}(\boldsymbol{z})] \neq \mu(\boldsymbol{z})$ and $\mathbb{E}[\tilde{\tau}] \neq \tau$. However, for GATE estimation in the setting where every node has the same individual treatment effect ($\tau_i \triangleq Y_i(\mathbf{1}) - Y_i(\mathbf{0})$), we observe that it is somewhat surprisingly an unbiased estimator for the GATE.

Theorem 4.10. *If the treatment effect of every node is constant across all nodes*, i.e., $\tau_i \equiv \tau$, *then using either GCR or RGCR scheme with* $p = 0.5$, *we have* $\mathbb{E}[\tilde{\tau}] = \tau$.

In practice, individual treatment effects τ_i are reasonably nonconstant across individuals, making the Hájek estimator potentially biased. In our simulations in Section 6, which feature nonconstant individual treatment effects, we find that this bias is modest in our settings and the overall MSE of the Hájek GATE estimator is broadly superior to that of the HT GATE estimator.

5 The curse of large clusters

In this section, we use a specific network and simple response model to study how the variance of RGCR can be affected by the network homophily. We conclude that in the setting, we investigate, if the number of clusters returned by the clustering algorithm is $O(1)$ in the size of the graph, a nonvanishing variance persists as part of the HT and Hájek estimators. We consider both independent and complete randomization.

Consider a ring-like network, the cycle graph with n nodes, where each node $i \in \{1, 2, \ldots, n\}$ is connected to nodes $i - 1$ and $i + 1$ (except for node 1 and n being connected). Further consider the following simple response model with network drift. For each node i,

$$Y_i(\boldsymbol{z}) = a + bh_i + \tau \frac{\sum_{j\in B_1(i)} z_i}{1 + d_i}, \tag{12}$$

where a, b, and τ are scalar constants.

In the second term, h_i is the homophily drift also used in our simulations in Section 6 and described in detail there. Informally, h_i is defined according to a natural disagreement minimization problem on the graph. On the cycle graph, this problem has the well-known closed-form solution

$$h_i = \sin\alpha_i, \quad \text{where } \alpha_i \triangleq \frac{i}{2\pi n}.$$

Here, α_i can be thought of as the angular (latent) position of node i along an evenly spaced cycle. The solution comes from basic properties of the cycle graph Laplacian, which is a symmetric circulant matrix. This h_i term then effectively models how nearby nodes generate similar responses while distant nodes generate different responses.

The third term in Section 5 represents a linear-in-means treatment effect, where we seek to estimate the GATE τ. We briefly note that this present model is simpler than the response model we consider in our simulations in Section 6, yet still sufficient to induce the curse of large clusters we seek to demonstrate.

With a constant $k > 0$ that divides n, an oracle clustering of this network into k clusters is the k-partition formed by breaking the "ring" into k equally sized connected arcs. Note that there are n/k such different oracle k-partitions.

We study the variance of the RGCR scheme with a random oracle k-partition, in the large-network scenario when $n \to \infty$. We have the following results on the HT estimator, with the proof given in Appendix C.

Theorem 5.1. *Suppose $p = 1/2$ and $k = o(n)$, then for the HT estimator $\hat{\tau}$ as $n \to \infty$,*

- *with independent randomization, we have*

$$\mathbf{Var}[\hat{\tau}] \to \frac{(2a+\tau)^2}{k} + \frac{b^2 k}{\pi^2}(1 - \cos(2\pi/k)) = [(2a+\tau)^2 + 2b^2]\cdot \Theta(1/k),$$

- *with complete randomization, we have*

$$\mathbf{Var}[\hat{\tau}] \to \frac{b^2k^2}{\pi^2(k-1)}(1 - \cos(2\pi/k)) = 2b^2 \cdot \Theta(1/k),$$

Theorem 5.1 yields two important insights. First, if the clustering algorithm generates a fixed number of clusters, then the variance of the HT GATE estimator, both for independent randomization and complete randomization, does not converge to 0 as $n \to \infty$. This is, in part or in full, due to the issue of network homophily, a phenomenon commonly observed in real-world networks whereby closely connected nodes share common behaviors [45]. In a large graph with few clusters, nodes in each cluster may generate different response than other clusters, obfuscate GATE estimation if we assign treatment/control at cluster level: the difference in the responses of different clusters might be unrelated to the treatment effect, but instead due to endogenous node properties captured in the network topology [62]. Therefore, in order for the variance of the estimator to vanish under RGCR, the clustering algorithm needs to generate an increasing number of clusters as the network grows large.

Second, the analysis also shows a separate deficit of independent randomization, at least in this model: the variance increases quadratically with the average response a, making the estimation sensitive to the scaling and shifting of the average responses. In contrast, complete randomization does not suffer from this issue, with a variance under this response model that is independent of a. Therefore, we recommend using complete randomization with RGCR whenever possible (when positivity is satisfied), a change from ordinary GCR where complete randomization typically does not satisfy positivity for any relevant exposure model.

We also note that the aforementioned complete randomization result for the HT estimator applies equally for the Hájek estimator, since these two estimators are asymptotically equivalent in this specific setting. Under complete randomization, and due to the fact that each cluster in the oracle k-partition contains the same number of nodes, we always have a constant number of nodes in the treatment and control groups. Moreover, due to the symmetry of the network, every node has the same exposure probability $\mathbb{P}[E_1^{\mathbf{z}}] \to 1/2$ in the limit of $n \to \infty$. Therefore, the denominator of the Hájek estimator concentrates at a constant n, making it equivalent to the HT estimator. In summary, we also have nonvanishing variance in the Hájek estimator if the number of clusters k is upper bounded in n as $n \to \infty$.

6 Simulation experiments

In this section, we evaluate the performance of the RGCR scheme in diverse simulations. After introducing the simulation setup in Sections 6.1 and 6.2, we examine the behavior of the HT estimator in Sections 6.3 and 6.5, and the Hájek estimator in Section 6.6. For each estimator, we first demonstrate significant variance reduction (as well as bias reduction for the Hájek estimator) under the RGCR scheme compared with GCR and then compare the bias, variance, and MSE under RGCR employing various random clustering algorithms.

As randomized clustering algorithms, we consider both randomized 3-net and 1-hop-max, both applied in unweighted, spectral-weighted, and degree-weighted forms. Note that for RGCR designs, we consider both independent and complete randomization, while for GCR, we only consider independent randomization (complete randomization is unattractive under GCR due to the potential violation of our positivity assumption). We find that the spectral- and degree-weighted variants of 3-net and 1-hop-max clusterings further reduce the variance of the HT estimator and the bias and variance of the Hájek estimator (compared with the unweighted clustering algorithms). In comparing complete randomization and independent randomization, we find that complete randomization leads to lower variance in the HT estimator and the two approaches have comparable bias and variance for the Hájek estimator.

These estimators require exposure probabilities, which are estimated with Monte Carlo methods introduced in Sections 4.2.2 and B.2. In Section 6.4, we demonstrate the high accuracy in estimation and visualize how the exposure probabilities vary under different random clustering algorithms. Specifically, we observe that applying the spectral- or degree-weighting scheme can increase the smallest exposure probabilities compared with the unweighted versions, offering an explanation of why the variance of the HT estimator as well as the bias and variance of the Hájek estimator is reduced.

Besides examining the bias, variance, and mean square error (MSE) in each network, we conclude this section by demonstrating how these quantities decay with the size n of the network. In Section 6.7, we show that the bias and variance of the Hájek estimator decays with a much higher rate under RGCR compared with GCR, which results in even more significant bias and variance reduction on large networks. These results highlight the broad favorability of the RGCR scheme in practice.

We provide code at https://github.com/hyin-stanford/RGCR-code that replicates all the analyses in this article.

6.1 Networks

We consider two interference networks across the experiments in this section. The first network is drawn from a variation on the small-world network model proposed by Kleinberg [63], itself a modification of small-world model proposed by Watts and Strogatz [64]. Besides the two properties of the Watts–Strogatz model of high clustering and short average pairwise distance, Kleinberg's small-world model is known for its navigability: individuals can find short chains from purely local information without centralized search [65].

The navigable small-world network is constructed from a periodic two-dimensional lattice: for each node, add a prespecified number of long edges, where the other end of each edge is randomly chosen on the network with probability proportional to the square of the inverse lattice distance, i.e.,

$$\mathbb{P}[\text{node } v \text{ is the end of random long edge from } u] \propto \mathbf{dist}_{\text{lattice}}(u, v)^{-2}.$$

The network we use is generated from a 96 × 96 lattice, where the number of long edges at each node is drawn from a power-law distribution [66] with exponent $\alpha = 2.3$. The resulting degree distribution is then heavy-tailed. We draw exactly one network from the model and fix it throughout the majority of our simulations. At a later point in the simulation discussion, we consider the effects of varying the network size within the framework of this model; we then sample a single graph for each lattice dimension.

Our second network for simulations is a snapshot the Facebook friendship network among Stanford students in 2005, included in the FB100 dataset [67]. Some basic properties of these two main networks are given in Table 2 with more detailed growth statistics given in Appendix A.

6.2 Response model

Our response model is intentionally more complicated than response models studied in pervious simulations of network interference; the added complications are intended to inject realism into the simulations. We propose that this model is "as simple as possible but not simpler," where removing any one of these components can mislead one to conclude that overly simplistic designs or analyses would work well in practice. We use the following response model throughout this simulation section:

$$Y_i(\mathbf{0}) = (a + b \cdot h_i + \sigma \cdot \varepsilon_i) \cdot \frac{d_i}{\bar{d}}, \tag{13}$$

$$Y_i(\boldsymbol{z}) = Y_i(\mathbf{0}) \cdot \left(1 + \delta z_i + \gamma \frac{\sum_{j \in N_i} z_j}{d_i}\right). \tag{14}$$

The model has the following components.

6.2.1 Parameters

The parameters a, b, and σ are constants, where a controls the shifting in average node's response, b controls the magnitude of homophily that results in a network drift effect (discussed below), and σ controls the noise level, where $\varepsilon_i \sim_{i.i.d.} N(0, 1)$ is independent of any other node attributes. In all our experiments, we use $a = 1$, $b = 0.5$, $\sigma = 0.1$, and let $\delta = \gamma = 0.5$ for the treatment effects.

6.2.2 Interference

Focusing first on the treatment effect, δz_i represents the direct effect and $\gamma \frac{\sum_{j \in N_i} z_j}{d_i}$ represents spillovers. With this response model, the full-neighborhood exposure model is properly specified, and we have

$$\tau_i = Y_i(\mathbf{1}) - Y_i(\mathbf{0}) = (\delta + \gamma) \cdot Y_i(\mathbf{0}), \tag{15}$$

Table 2: Basic properties of the two interference networks studied in our simulations. For more detailed growth statistics on these two networks, see Appendix A

Network	n	m	$\bar{d}$	d_{max}	κ
Small world	9,216	55,214	11.98	42	21.8
FB-stanford	11,586	568,309	98.10	1172	586.5

and the GATE τ becomes

$$\tau = \frac{1}{n}\sum_i \tau_i = (\delta + \gamma)\cdot \mu(\mathbf{0}). \tag{16}$$

6.2.3 Degree-correlated responses

The role of the degree d_i and average degree $\bar{d}$ induce a strong correlation between node degree and control response, a realistic phenomenon [68] that also injects heavy-tailed-ness into the response distribution whenever the degree distribution is heavy-tailed.

6.2.4 Multiplicative treatment effect

Instead of the more common additive treatment effect here the treatment effect is multiplicative at the node level. This multiplicative model, which has also been studied elsewhere [3], caries forward the correlation between degree and control response to cause a heterogeneous "individual" global treatment effect. Note that, according to Theorem 4.10, a heterogeneous treatment effect is required to reveal the bias in Hájek estimation. A multiplicative treatment effect can also be deemed natural because the different exposure levels incur the same *relative* change in a units' response.

6.2.5 Homophily

Our use of a network homophily term h_i is new to the literature on causal inference under interference, and we believe it provides an important missing piece for evaluating experimental designs under a more realistic response model. This term represents the network drift phenomenon where closely connected nodes usually have similar characteristics and consequently similar response, while distant nodes can be dissimilar. For example, in the United States, many behaviors are correlated with geography, while the network structure is also very obviously correlated with geography [43]. Failure to consider this network drift in the response incurs additional variance in the estimation under GCR: assigning all the East Coast users into the treatment while all West Coast users to control would make GATE estimation sensitive to any variations related to this network-level drift effect.

We can think of the homophily term h_i in the outcome model as arising due to a latent process that also gave rise to the network. That said, we wish to add outcome homophily to simulations on real-world networks where no latent process for generating the graph is known. To retrofit homophily into simulation on real-world graphs, we construct our h_i feature as solving the following disagreement minimization problem:

$$\begin{aligned} \min_{\boldsymbol{h}\in\mathbb{R}^n} \quad & \sum_{(i,j)\in E} (h_i - h_j)^2 \\ \text{subject to} \quad & \sum_{i=1}^{n} d_i h_i = 0, \\ & \max\{|h_i| : 1 \le i \le n\} = 1. \end{aligned} \tag{17}$$

Without the constraints, it is clear that any constant h_i would minimize the objective, but we are constructing h_i to be the scalar function that minimizes the disagreement across all edges, subject to their being nonzero disagreement. This minimization problem is a classic problem in the spectral graph theory, where the solution is the eigenvector associated with the second smallest eigenvalue of the normalized graph Laplacian matrix $D^{-1}L$ [69], i.e.,

$$D^{-1}L\boldsymbol{h} = \lambda_2\boldsymbol{h},$$

where D is the diagonal degree matrix, and $L = D - A$ is the unnormalized graph Laplacian. In the simple cycle network studied in Section 5, we saw that h_i has a simple closed-form solution $h_i = \sin\alpha_i$, and we can think of α_i here as an explicit latent position in a periodic one-dimensional space. For real-world graphs, h_i derived from the aforementioned disagreement minimization problem generalizes this latent homophily without requiring an explicit generative model that co-generates the network G and homophilous response $\boldsymbol{h}$.

A consequence of using this drift function h_i is that, as $n \to \infty$, we have $\mu(\mathbf{0}) \to a$. Such convergence is due to $\sum_i h_i d_i = 0$ as in equation (17), and $\frac{1}{n}\sum_i \varepsilon_i d_i \to 0$ since $\varepsilon_i \sim_{i.i.d.} N(0, 1)$, which is also independent of d_i. As a result, $\frac{1}{n}\sum_i a(d_i/\bar{d}) = a$. Consequently, we have

$$\tau = \mu(\mathbf{0}) \to 1,$$

as $n \to \infty$.

In Figure 2, we visualize the node homophily feature h_i in our heavy-tailed small-world network as well as the resulting response under global control $Y_i(\mathbf{0})$. Notice that adjacent nodes on the lattice have closer value in h_i, while distant nodes tend to have significantly different h_i, broadly mimicking the existence of a latent (in the small-world network, spatial) component to the response. The variation of h_i's along the lattice is not smooth due to the existence of a heavy-tailed number of long-range edges in the small-world network.

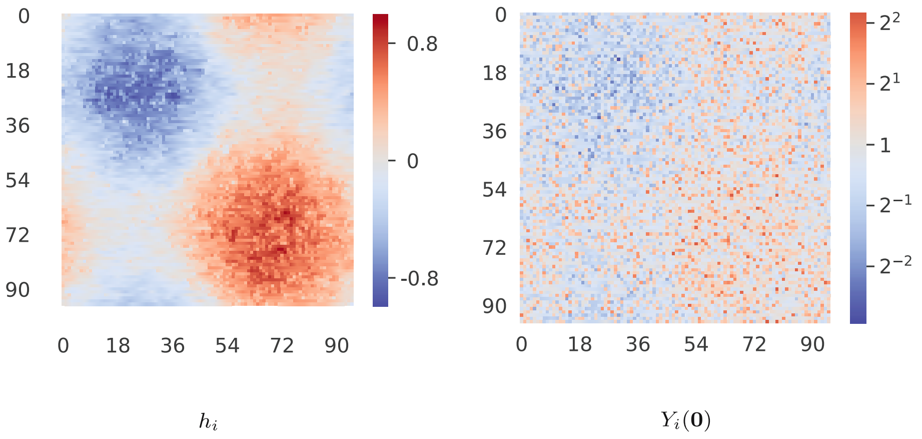
6.3 Variance reduction of $\hat{\tau}$

In this section, we demonstrate the significant variance reduction of the HT estimator under RGCR compared with the standard GCR scheme.

To make the benefits of randomization concrete, for each RGCR design, we study the variance of HT GATE estimators when mixing K clusterings, for varying values of K. Specifically, we first generate K random clusterings from the random clustering algorithm $\mathcal{P}$ and fix them. We then let the random clustering distribution $\mathcal{P}_K$ be the uniform distribution among the K generated clusterings, $\{\mathbf{c}^{(k)}\}_{k=1}^{K}$: in the design phase, we use a single clustering out of the K clusterings drawn uniformly at random; and in the analysis phase, the exposure probability of each node under $\mathcal{P}_K$ is then

$$\mathbb{P}[E_i^{\mathbf{z}} | \mathcal{P}_K] = \frac{1}{K}\sum_{k=1}^{K}\mathbb{P}[E_i^{\mathbf{z}} | \mathbf{c}^{(k)}]. \tag{18}$$

The benefit of analyzing mixtures of K fixed clusterings is two-fold. First, under this K-cluster design we can compute the exposure probabilities (equation (18)) exactly (compared to estimating them using a



h_i

$Y_i(\mathbf{0})$

Figure 2: Heatmap of the node homophily feature h_i (left) and the corresponding response under global control $Y_i(\mathbf{0})$ (right) of a heavy-tailed small-world network based on the periodic 96 × 96 lattice.

Monte Carlo procedure when considering the full clustering distribution $\mathcal{P}$). Second, it unifies the GCR and RGCR schemes: With $K = 1$, we are considering the standard GCR scheme with a single fixed clustering, while as $K \to \infty$, it approaches the RGCR scheme based on the random clustering algorithm $\mathcal{P}$.

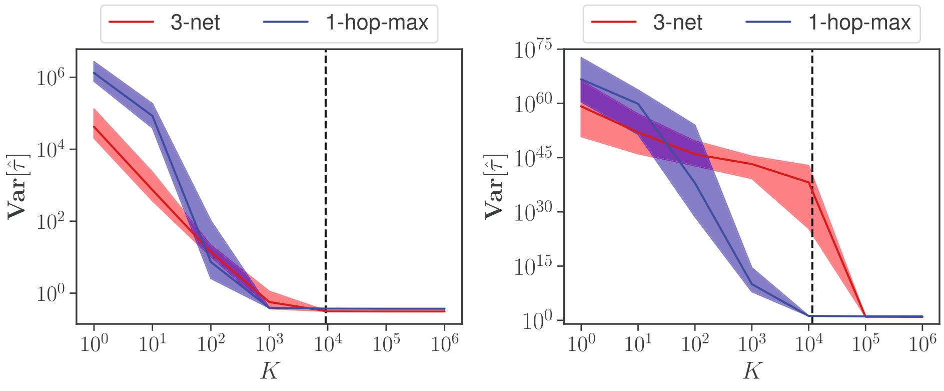
In our simulations, we contrast mixtures of $K \in \{1, 10, 10^2, \ldots, 10^6\}$ clusterings for both our heavy-tailed small-world network and the FB-Stanford friendship network. The variance of the GATE estimators depend on the K partitions being randomly generated, and thus, we repeat each partition generation process 400 time to obtain a distribution. We assign each cluster to the treatment group with probability $p = 0.5$, the symmetric assignment scenario where we have $\mathbb{P}[E_i^{\mathbf{1}}|\mathcal{P}] = \mathbb{P}[E_i^{\mathbf{0}}|\mathcal{P}]$.

Figure 3 shows the variance of the HT GATE estimator under each unweighted random clustering strategy with independent cluster-level assignment. Within each scheme, we observe enormous variance reduction from randomized clustering in both the synthetic small-world network and the Facebook friendship network. With $K = 1$, i.e., the standard GCR scheme with a fixed clustering, the variance is uselessly high due to the existence of many exponentially small exposure probabilities. As K increases, the variance of the HT GATE estimator decreases monotonically and significantly. Toward the limit of $K \to \infty$, which is an approximation to the RGCR scheme, the variance is reduced to a realistic level of around 10^0. The exact value of the variance depends strongly on both the response model and the size of the networks we use in our simulations. Consistent with the theory developed in Section 4, such variance reduction agrees with what one would expect from the exponentially small exposure probabilities being "washed out," gradually growing to probabilities that are only polynomially small.

6.4 Exposure probabilities

In Section 6.3, we approximate the RGCR scheme under the random cluster distribution $\mathcal{P}$ by using a uniform clustering from a collection of K clusterings from $\mathcal{P}$. In Sections 6.4.1 and 6.4.2, we examine the performance of the generic RGCR scheme. Recall that due to the infeasibility of exactly computing the exposure probabilities (Theorem 4.4), for a generic $\mathcal{P}$, we must rely on estimated probabilities obtained via Monte Carlo (Sections 4.2.2 and B.2).

We first validate the accuracy in this Monte Carlo procedure and then also examine the estimated probabilities, comparing them with the theory developed in Section 4. Again all simulations are conducted under the scenario where we assign each cluster into the treatment group with probability $p = 0.5$; thus, $\mathbb{P}[E_i^{\mathbf{1}}|\mathcal{P}] = \mathbb{P}[E_i^{\mathbf{0}}|\mathcal{P}]$, and we only need to estimate $\mathbb{P}[E_i^{\mathbf{1}}|\mathcal{P}]$.



Small World FB-Stanford

Figure 3: The distribution of **V ar**[$\hat{\tau}$] when mixing K random clusterings from the unweighted randomized 3-net and 1-hop-max algorithms in the heavy-tailed Small World network (left) and the FB-Stanford network (right). The number of nodes in each network is marked by a dashed line. We plot the median (solid line) as well as the 2.5 and 97.5% quantiles (shaded area) of the variance distribution from our simulations. For both algorithms and both networks, we observe enormous variance reduction from cluster mixing. Similar trends are also seen in the weighted clustering methods (not shown).

6.4.1 Accuracy in exposure probabilities estimation

We demonstrate the accuracy of our Monte Carlo procedure for a 32 × 32 instance of our heavy-tailed small-world network model using the unweighted 3-net clustering as an example. To measure the relative error of the probabilities, as well as how they decay with the number of stratified samples in Monte Carlo estimation, we conduct the estimation procedure with Kn stratified samples, where K ranges from $\{1, 2^1, \ldots, 2^7\}$. For each K, we repeat the estimation 10 times and estimate the relative standard deviation

$$\mathbf{rstd}(\hat{\mathbb{P}}_K[E_i^{\mathbf{1}}|\mathcal{P}]) = \frac{\mathbf{std}[\hat{\mathbb{P}}_K[E_i^{\mathbf{1}}|\mathcal{P}]]}{\mathbb{P}[E_i^{\mathbf{1}}|\mathcal{P}]}$$

of each node i with maximum likelihood, where the subscript K denotes that Kn samples were used. We then compute the average relative standard deviation of all nodes for each K and summarize the results in the left subfigure in Figure 4. We observe from the figure that the average relative standard deviation decays as

$$\widehat{\mathbf{rstd}}(\hat{\mathbb{P}}_K[E_i^{\mathbf{1}}|\mathcal{P}]) \propto K^{-1/2}.$$

When $K = 128$, the average relative standard deviation is around 1%.

Besides the relative standard deviation, we also examine the distribution of the relative error of a set of estimated exposure probabilities from $128n$ stratified samples,

$$\mathbf{err}_i = \frac{\hat{\mathbb{P}}[E_i^{\mathbf{1}}|\mathcal{P}] - \mathbb{P}[E_i^{\mathbf{1}}|\mathcal{P}]}{\mathbb{P}[E_i^{\mathbf{1}}|\mathcal{P}]},$$

where we use the average exposure probability across the 10 repetitions as the ground truth exposure probability of each node. The histogram of the relative errors at all nodes is shown in Figure 4 (right), where we see the relative errors are bounded within ±5% and mostly within ±2%. Analogous results for other networks and clustering algorithms (not shown) confirm a broadly satisfying accuracy for the estimated exposure probabilities. We use these estimated probabilities in place of the exact exposure probabilities in all our uses of the HT and Hájek GATE estimators.

6.4.2 Visualizing exposure probabilities

Figure 5 furnishes a scatterplot of the estimated exposure probabilities under each randomized clustering strategy with cluster-level independent randomization versus the size of the 2-ball at each node. In the first column, where we see the unweighted randomized 3-net and 1-hop-max schemes, the lower bound provided in Theorem 4.2 is verified (blue dashed line), as well as the slightly stronger lower bound for the unweighted 1-hop-max scheme from Theorem 4.3. These lower bounds fall off with $|B_2(i)|$, and thus nodes with larger 2-hop neighborhood can have lower exposure probabilities. Moreover, we observe that the lower bounds are more tight for nodes with a larger 2-neighborhood.

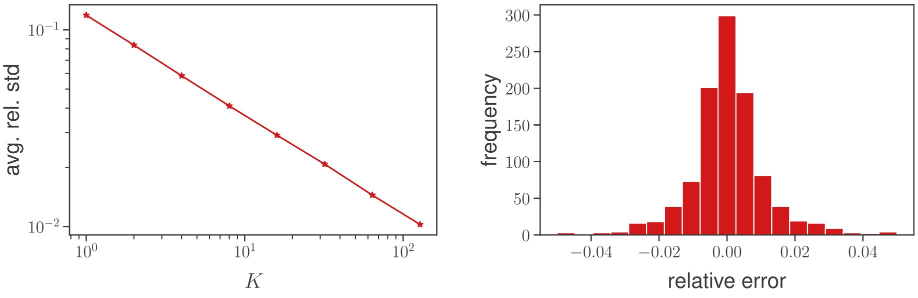


Figure 4: Left: Average relative standard deviation of the exposure probability estimator with Kn stratified samples in Monte Carlo estimation. Right: Histogram of the relative error in the estimated exposure probabilities of all nodes, where $128n$ stratified samples are used in Monte Carlo estimation. Both plots are constructed with the heavy-tailed small-world network.

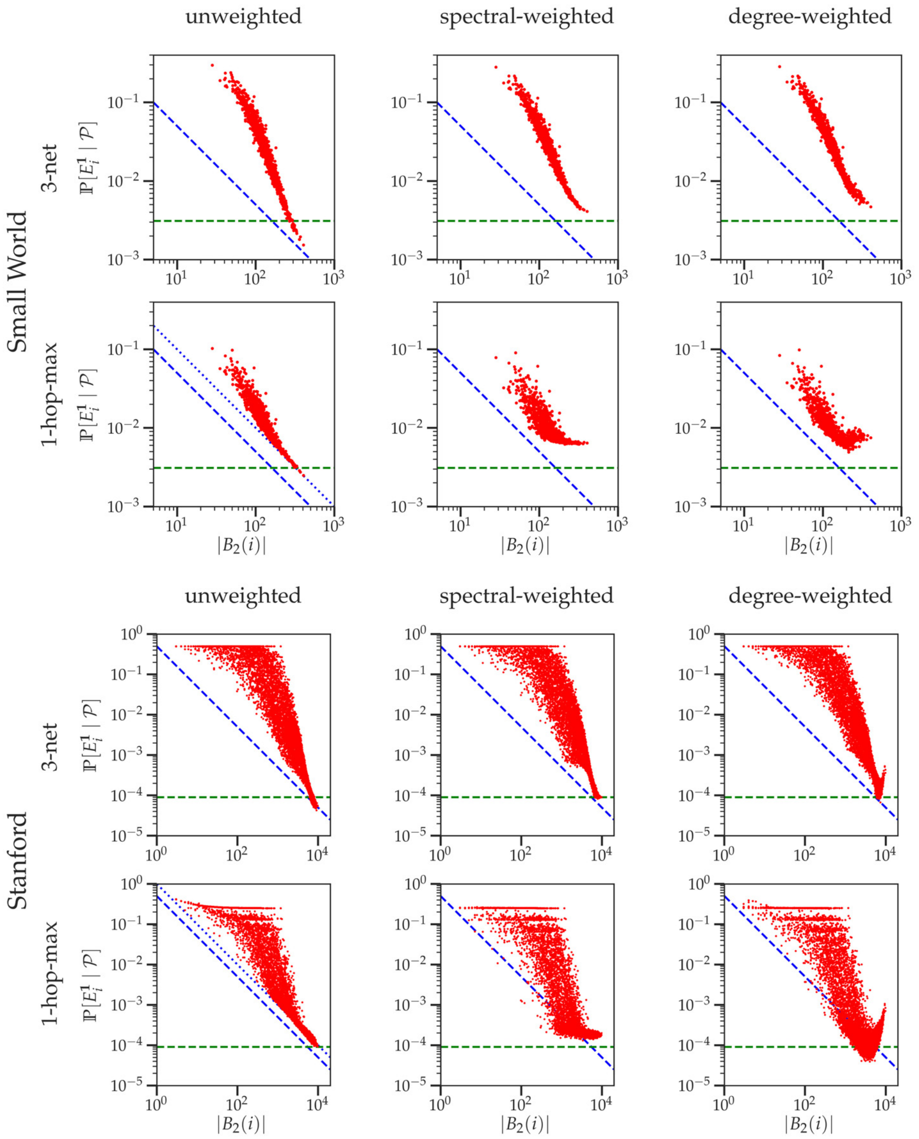


Figure 5: Scatterplot of the exposure probability $\mathbb{P}[E_i^1|\mathcal{P}]$ versus $|B_2(i)|$ at every node i in the heavy-tailed Small World network (top two rows) and the FB-Stanford network (bottom two rows), under each random clustering scheme with cluster-level independent randomization. The blue dashed line represents the exposure probability lower bound for unweighted 3-net and 1-hop-max schemes (Theorem 4.2). The blue dotted line represents the slightly improved lower bound for unweighted 1-hop-max (Theorem 4.3). The green dashed line represents the uniform lower bound for the spectral-weighted schemes (Theorem B.4).

The second and third columns of Figure 5 are associated with the weighted 3-net and 1-hop-max clustering strategies. In the second column, we consider the spectral weighting developed in Appendix B.3, which obey a uniform lower bound (green dashed line) on the exposure probability independent of $|B_2(i)|$. While the bound is uniform, the slack is not, and we observe that it is again more tight for nodes with larger $|B_2(i)|$. Besides spectral weighting, in the third column, we consider another scheme where each node is weighed by its degree. Compared with the spectral weighting, this degree weighting scheme improves the exposure probabilities at nodes with largest $|B_2(i)|$, while it may harm other nodes: note the dip below the green dashed line in the 1-hop-max scatterplots for the FB-Stanford network.

We further compare the exposure probability across different schemes in Figure 6. In the first two columns, we examine how the spectral weighting affects nodes' exposure probabilities associated the randomized 3-net and 1-hop-max clustering respectively. For 3-net, under both networks, spectral

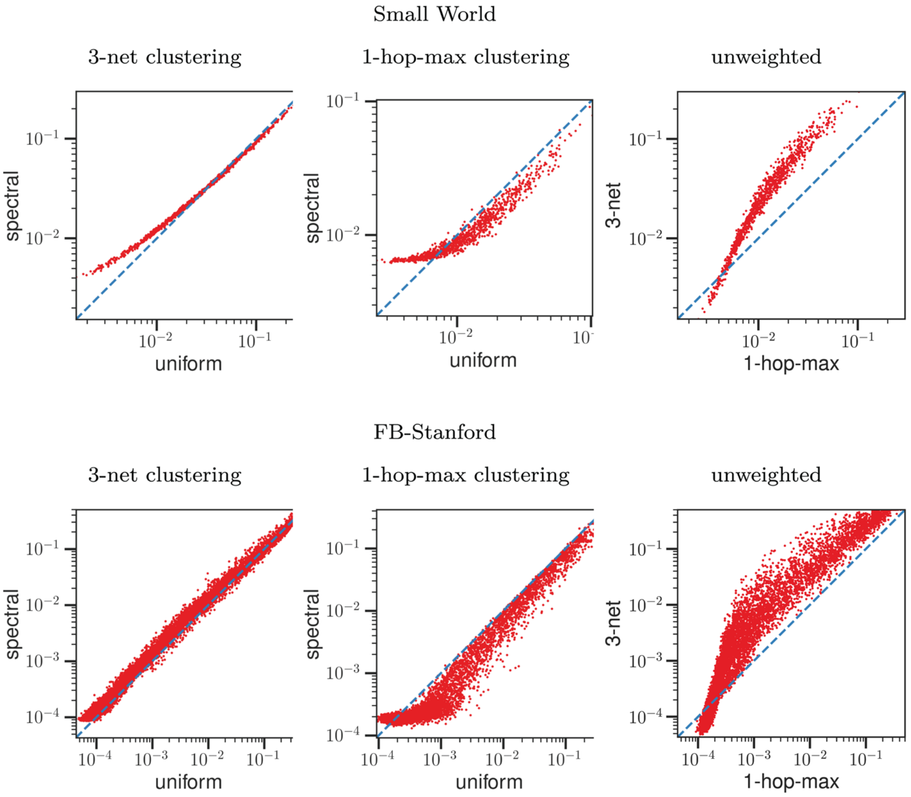


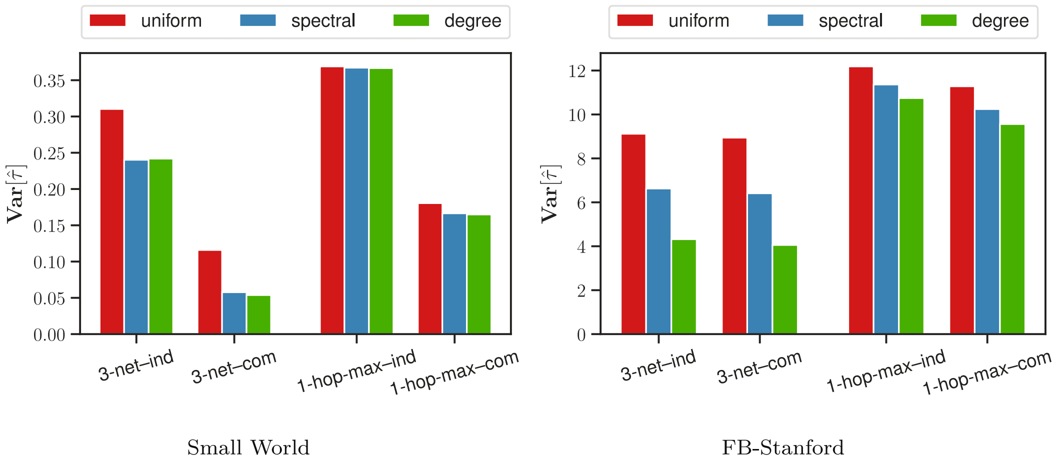
Figure 6: Scatterplot of the exposure probability $\mathbb{P}[E_i^{\mathbf{1}}|\mathcal{P}]$ of each node in the Small World network (first row) and FB-Stanford network (second row), with different random clustering strategies. The blue dashed line marks the scenario when the exposure probability under two schemes are same. In the first column, we compare the exposure probability at each node with uniform- (i.e., unweighted) and spectral-weighted random 3-net clustering, and similar comparison for 1-hop-max clustering is given in the second column. In the last column, we compare the exposure probabilities with unweighted 3-net and 1-hop-max clusterings.

weighting effectively increases the exposure probability of nodes whose probability is small under the unweighted 3-net, at a very small cost of decreasing some large exposure probabilities. In contrast, for 1-hop-max clustering, even though spectral weighting can increase the very small exposure probabilities seen in the unweighted scheme, it also significantly decreases the probability of many other nodes. Comparing 3-net clustering and 1-hop-max as in the last column, we observe that the exposure probabilities under 3-net are mostly higher than under 1-hop-max, though the smallest exposure probability under 1-hop-max is higher due to the improved lower bound theory in Theorem 4.3.

6.5 HT estimator variance

We now examine the variance of $\hat{\tau}$ under the RGCR scheme with each random clustering strategy. In both the Small World and FB-Stanford networks, we consider 3-net and 1-hop-max clusterings, each under uniform-, spectral-, and degree-weighting schemes. We also consider both independent and complete randomization at the cluster level.

The variances are obtained from equations (4) to (6), the exact ground-truth variance (available in simulations). In addition to the exposure probability of each node $\mathbb{P}[E_i^{\mathbf{1}}|\mathcal{P}]$, the variance formulae require the exposure probabilities of each pair of nodes, i.e., $\mathbb{P}[E_i^{\mathbf{z}_1} \cap E_j^{\mathbf{z}_2}|\mathcal{P}]$ for any node pair $i, j \in V$ and $\mathbf{z}_1, \mathbf{z}_1 \in \{\mathbf{1}, \mathbf{0}\}$. We also estimate these co-exposure probabilities using Monte Carlo estimation with Kn stratified samples, with $K = 128$ for the Small World network and $K = 16$ for the FB-Stanford network.



Small World

FB-Stanford

Figure 7: Variance of the HT GATE estimator under the RGCR scheme with various random clustering strategies. The suffix of each clustering method distinguishes independent (- -ind) or complete (- -com) randomization at the cluster level. The variance of these HT estimators under GCR, not shown, are all dramatically higher (comparable to results in Figure 3).

The results are shown in Figure 7, where we make four observations. First, for both networks and every random clustering and weighting scheme, complete randomization usually yields lower variance than independent randomization. Such variance reduction can be explained by the negative correlation introduced in the cluster-level assignment process, leading to larger values of $\mathbb{P}[E_i^{\mathbf{1}} \cap E_j^{\mathbf{0}} | \mathcal{P}]$ and a positive $\mathbf{Cov}[\hat{\mu}(\mathbf{1}), \hat{\mu}(\mathbf{0})]$. However, this reduction is much less substantial in the Facebook network, which we believe is due to its faster neighborhood growth structure as illustrated in Appendix A.

Second, in both networks and with both the 3-net and 1-hop-max random clustering strategy, the variance with the spectral-weighted scheme is lower than that of the unweighted version. Such reduction can be explained by the increase in the lowest exposure probabilities (see the first two columns of Figure 6). This variance reduction is less significant for 1-hop-max clustering, which is consistent with how the spectral-weighted scheme also decreases the exposure probabilities of most nodes (second column of Figure 6).

Third, we observe that degree weighting usually gives lower variance than the spectral-weighted scheme. According to the discussion begun in Appendix B.3, spectral weighting achieves a uniform lower bound on the exposure probability of every node, but this lower bound is less tight for nodes with smaller $|B_2(i)|$ (see Figure 5), and thus, it is not unexpected that a weighting favoring the nodes with large neighborhood, as degree-weighting does, would increase the minimum exposure probabilities of all nodes and reduce variance.

We conclude that for HT estimators, RGCR, with randomized 3-net and 1-hop-max are generally comparable. Between the two, randomized 3-net usually yields a modestly lower variance. According to Lemma 4.1, 1-hop-max has a local dependency property, and thus, the cross-node terms in the variance formulae (equations (4)–(6)) decay significantly in the distance between node pairs (and become zero when the distance is greater than 4). In contrast, randomized 3-net has no local dependency guarantee, and consequently, we do not have a nontrivial theoretical upper bound on the variance. However, in our simulations, the cross-node terms are also small, making the variance of the HT estimator under randomized 3-net even lower than using 1-hop-max.

In summary, for the HT GATE estimator for the response model and networks, we study:

- complete randomization yields lower variance than independent randomization,
- randomized 3-net clustering yields lower variance than 1-hop-max,
- spectral- and degree-weighting schemes yields lower variance than unweighted schemes.

6.6 Hájek estimator bias and variance

Unlike the HT estimator, the Hájek estimator does not have close-form formulae to compute the bias, variance, or MSE, and thus, we estimate these quantities via simulation. We therefore briefly describe how we evaluate performance via simulation. For GCR, since the estimation performance is associated with the specific clustering is use, we use the median bias, variance, and MSE across 1,000 randomly generated clusterings. Specifically for each clustering, we simulate the experimental procedure (assignment, outcome generation, and GATE estimation) and compute the sample bias, variance, and MSE and use as the proxy of the corresponding measure of the GCR scheme. For RGCR, we simulate the experimental procedure (random clustering generation, assignment, outcome generation, and GATE estimation) $50 * n$ times and analogously estimate each measures with the sample bias, sample variance, and sample MSE. As in Appendix B.2, here we use a similar idea of stratified sampling for the weighted 3-net clustering design: for each node, there are 50 times when it is ranked first among all nodes in generating the random clustering and guaranteed to be a seed node in the 3-net clustering. In the analysis phase, we use the exposure probabilities estimated for RGCR in Section 6.4. For GCR, we use the exact exposure probabilities associated with the clustering in use.

Figure 8 presents the bias, variance, and MSE of the Hájek estimator under GCR and RGCR, focusing on independent randomization. We make two main observations. First, the bias of the Hájek estimator under RGCR has been significantly reduced, compared with GCR. Recall that in our response model (Section 6.2), the ground-truth GATE is $\tau = 1.0$. The bias is reduced from 15% to less than 0.5% in the Small World network and from 93% to less than 20% in the FB-Stanford network.

One can intuitively interpret this bias reduction as follows. Under GCR, for nodes with exponentially small exposure probability (in the FB-Stanford network, it can be lower than 10^{-50}), they are almost never network exposed to treatment or control. As a result, their response Y_i are almost never revealed in the weighted averaging procedure of the Hájek estimator. In every experiment execution, the exposure nodes

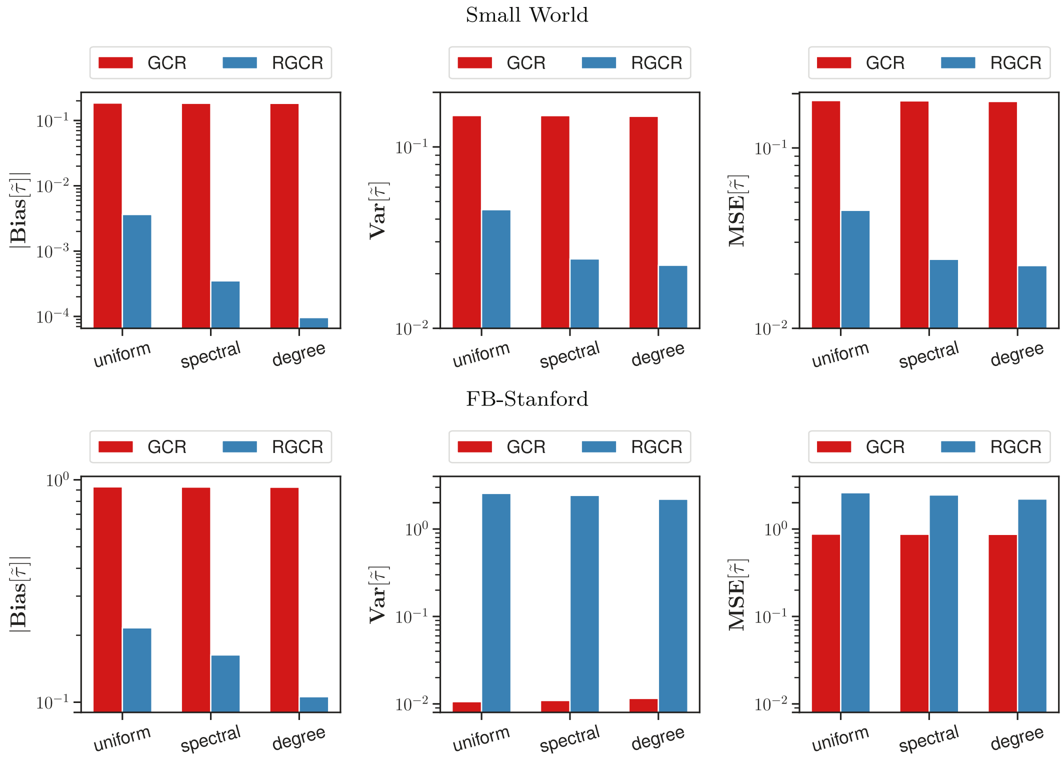


Figure 8: Bias, variance, and MSE of the Hájek GATE estimator under GCR and RGCR with 3-net clustering and independent randomization in the Small World (first row) and FB-Stanford (second row) networks.

are almost only those with large exposure probabilities and mostly those with small degree. A large degree node may also have a large exposure probability if it is at the center of a cluster; however, it is much less likely to be at the center of a cluster than a low-degree node, and even if it is exposed, its weight (the inverse of the exposure probability) is not larger than those low-degree exposed nodes. Therefore, the analysis procedure in GCR is internally biased against the large degree nodes, favoring the low-degree nodes, which have smaller individual treatment effects τ_i in our response model. Consequently, we see how the Hájek estimator can be severely biased downwards in such settings.

In contrast, for RGCR, the exposed nodes are not so strictly high exposure probabilities. For example, if a large degree node is at the center of a cluster in a randomly generated clustering, even though it has large *conditional* exposure probability under this clustering and thus being likely to be network exposure to treatment or control, its unconditional exposure probability can still be small. As a result, it is weighted more heavily in the weighted average procedure of Hájek estimation, making the estimator value shift toward the response of large degree nodes and thus less biased than that under GCR.

Alongside this understanding of Hájek bias, it is also expected to observe an increase in variance from RGCR, vs GCR, under Hájek estimation. In Figure 8, we see variance reduction from RGCR in the Small World network but an increased variance in the FB-Stanford network. Under GCR, since the estimator value is dominated by the response of low degree nodes, in our response model, the response of low-degree nodes have a much narrower range than the whole population, resulting in low variance (but overwhelming bias, we repeat).

Finally, we also compare the bias and MSE of the RGCR scheme with different random clustering strategies, which we also include complete randomization, and the results are given in Figure 9. In general, the benefits of complete randomization (over independent randomization) that we see for the HT estimator do not appear to carry over to Hájek estimation.

In summary, comparing with the GCR scheme, the Hájek estimator under the RGCR scheme has significantly lower bias but may have larger variance. Examining the MSE that trades off bias and variance, we see a lower MSE in the Small World network from RGCR (vs GCR), while we see a higher MSE in the FB-Stanford network under RGCR (vs GCR).

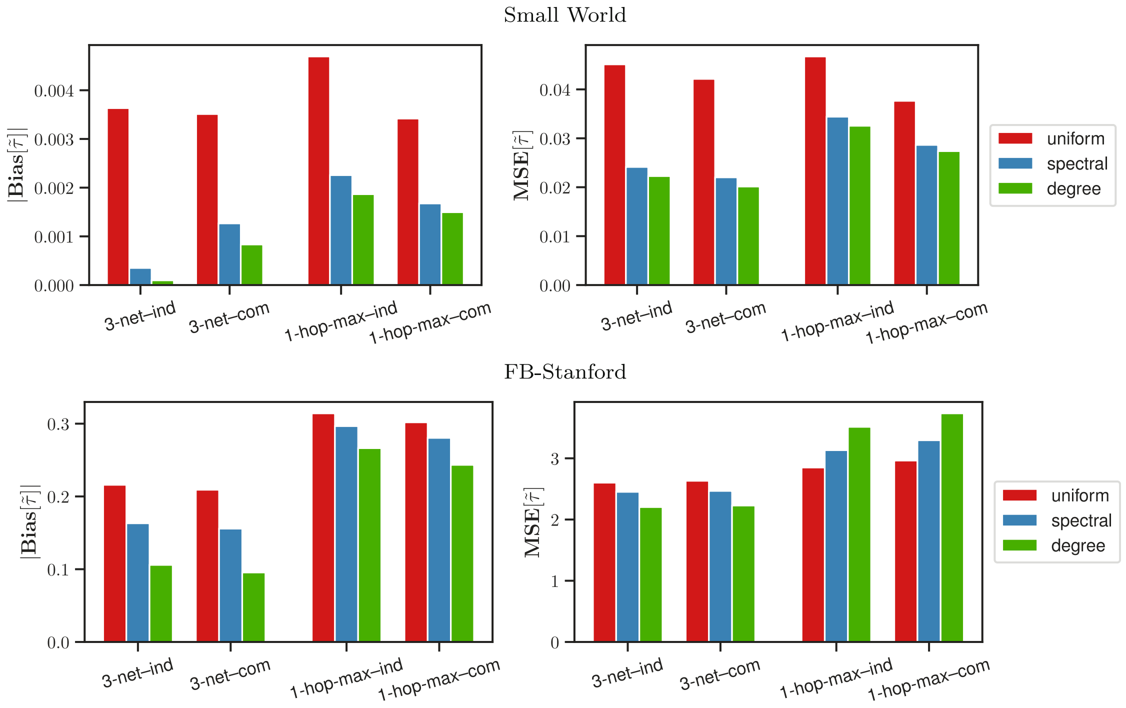


Figure 9: Bias and MSE of Hájek GATE estimator under RGCR with various clustering algorithms with both independent and complete randomization.

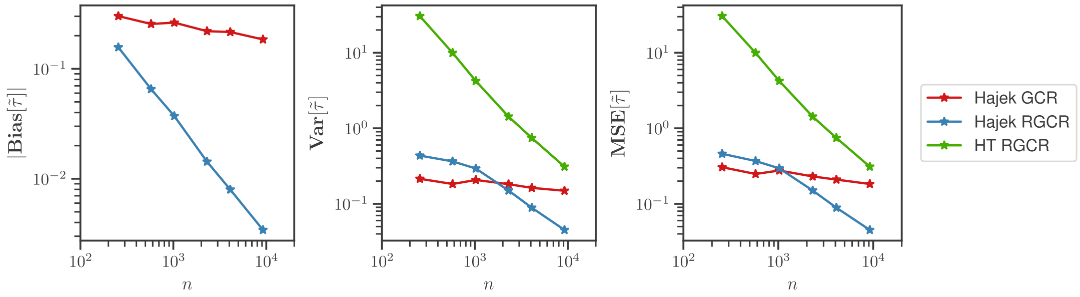


Figure 10: Bias (absolute value), variance, and MSE of the Hájek and HT GATE estimator under GCR and RGCR with unweighted 3-net clustering in a sequence of increasingly sized heavy-tailed small-world networks.

6.7 Variance, bias, and network size

Here, we examine how the bias, variance, and MSE change as a function of network size. Recall that the heavy-tailed small-world network we use throughout our earlier simulations is based on a periodic two-dimensional lattice of side length 96 (thus, $n = 96 * 96 = 9{,}216$). Here, we consider a series of heavy-tailed small-world networks in this family. Specifically, we generate a sequence of networks based on lattices of size $16 * 16$, $24 * 24$, $32 * 32$, $48 * 48$, $64 * 64$, and $96 * 96$, while the procedure (and parameters) for adding long-range edges remains fixed. With this sequence of networks, in Figure 10, we repeat the aforementioned simulation procedures and compare the bias, variance, and MSE under the GCR and RGCR schemes. This analysis examines both the Horvitz–Thompson (HT) and Hájek estimators.

From the first plot in Figure 10, we observe that the Hájek estimator under the RGCR scheme has consistently less bias than under the GCR scheme, across the range of network sizes we study. Moreover, the bias decays with the network size at a much higher rate for RGCR than for GCR. Recall that the HT estimator is unbiased (under both GCR and RGCR).

From the second plot, we observe that the variance of the Hájek estimator under RGCR also decays with higher rate than under GCR. We can explain the slow decay of the variance under GCR by observing that as the number of nodes n increases, so does the number of nodes with small exposure probabilities. It then becomes more likely that more nodes with small exposure probabilities have been network exposed. Note that variance of response among these nodes are large, and thus even though the expected number of exposed nodes increases proportional to n, due to the introduction of small exposure probability nodes, the variance of the Hájek estimator decays at a much smaller rate than n^{-1}. In contrast, the empirical rate of variance decay of RGCR is close to n^{-1}. Therefore, even though the variance under GCR might be lower in small networks, for the reason explained in Section 6.6, the RGCR scheme can significantly reduce variance in large networks. Meanwhile, the variance of the HT estimator under RGCR, while higher than that of the Hájek estimator with RGCR, also exhibits an empirical decay rate of n^{-1}. The variance of HT estimator under GCR is extremely high across all networks and is not presented in the plot.

Combining both bias and variance, we observe that the MSE of all three methods decays with the network size, while the decay rate is notably faster for the estimators based on the RGCR scheme. In summary, with a large interference network, the Hájek estimator with RGCR is preferred.

7 Conclusion

We developed RGCR as a scheme for the design and analysis of randomized experiments in the presence of interference. This scheme is an improvement on the graph clustering randomization (GCR) scheme in that it is based on a distribution of random clusterings instead of a single fixed clustering, with favorable consequences for the bias and variance of standard estimators. Compared to GCR, the RGCR scheme with proper random clustering generators enjoys significantly reduced variance for both the Horvitz–Thompson and the Hájek estimator of the GATE, and also supports complete randomization. We also discuss how the network drift pattern in node's responses, as is observed in real-world settings, plays an important role in the variance of GATE estimation, and propose a new response model exhibiting homophily in the form of network drift in responses, facilitating a more careful analysis of realistic estimator performance.

Acknowledgements: We thank Guillaume Basse, Alex Chin, Dean Eckles, and Aaron Sidford for valuable discussions, as well as participants at the Conference on Digital Experimentation (CODE) and at the MIT Initiative on the Digital Economy Seminar.

Funding information: This work was supported in part by NSF grant IIS-1657104.

Author contributions: All authors have accepted responsibility for the entire content of this manuscript and approved its submission.

Conflict of interest: The authors state no conflict of interest.

Data availability statement: Replication code available at https://github.com/hyin-stanford/RGCR-code.

References

[1] Hudgens MG, Halloran ME. Toward causal inference with interference. J Amer Stat Assoc. 2008;103(482):832–42.
[2] Fienberg SE. A brief history of statistical models for network analysis and open challenges. J Comput Graph Stat. 2012;21(4):825–39.
[3] Aronow PM, Samii C. Estimating average causal effects under general interference, with application to a social network experiment. Ann Appl Stat. 2017;11(4):1912–47.
[4] Sävje F, Aronow P, Hudgens M. Average treatment effects in the presence of unknown interference. Ann Stat. 2021;49(2):673.
[5] Chin A. Central limit theorems via stein's method for randomized experiments under interference. 2018. arXiv: http://arXiv.org/abs/arXiv:1804.03105.
[6] Jagadeesan R, Pillai NS, Volfovsky A. Designs for estimating the treatment effect in networks with interference. Ann Stat. 2020;48(2):679–712.
[7] Eckles D, Karrer B, Ugander J. Design and analysis of experiments in networks: reducing bias from interference. J Causal Infer. 2017;5(1):20150021.
[8] Saint-Jacques G, Varshney M, Simpson J, Xu Y. Using ego-clusters to measure network effects at linkedin. 2019. arXiv: http://arXiv.org/abs/arXiv:1903.08755.
[9] Chin A. Regression adjustments for estimating the global treatment effect in experiments with interference. J Causal Infer. 2019;7(2).
[10] Saveski M, Pouget-Abadie J, Saint-Jacques G, Duan W, Ghosh S, Xu Y, et al. Detecting network effects: Randomizing over randomized experiments. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2017. p. 1027–35.
[11] Pouget-Abadie J, Mirrokni V, Parkes DC, Airoldi EM. Optimizing cluster-based randomized experiments under monotonicity. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2018. p. 2090–9.
[12] Pouget-Abadie J, Saint-Jacques G, Saveski M, Duan W, Ghosh S, Xu Y, et al. Testing for arbitrary interference on experimentation platforms. Biometrika. 2019;106(4):929–40.
[13] Blake T, Coey D. 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; 2014. p. 567–82.
[14] Fradkin A. A simulation approach to designing digital matching platforms. Boston University Questrom School of Business Research Paper Forthcoming; 2019. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3320080.
[15] Ha-Thuc V, Dutta A, Mao R, Wood M, Liu Y. A counterfactual framework for seller-side a/b testing on marketplaces. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, NY, USA; 2020. p. 2288–96.
[16] Holtz D, Aral S. Limiting bias from test-control interference in online marketplace experiments. 2020. arXiv: http://arXiv.org/abs/arXiv:2004.12162.

[17] Johari R, Li H, Liskovich I, Weintraub GY. Experimental design in two-sided platforms: an analysis of bias. Management Sci. 2022;68(10):7069–89.
[18] Ogburn EL, VanderWeele TJ. Causal diagrams for interference. Stat Sci. 2014;29(4):559–78.
[19] Ogburn EL, Shpitser I, Lee Y. Causal inference, social networks and chain graphs. J R Stat Soc Ser A (Stat Soc) 2020;183(4):1659–76.
[20] Cox DR. Planning of Experiments. New York: Wiley; 1958.
[21] Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. J Educat Psychol. 1974;66(5):688.
[22] Sobel ME. What do randomized studies of housing mobility demonstrate? causal inference in the face of interference. J Amer Stat Assoc. 2006;101(476):1398–407.
[23] Manski CF. Identification of treatment response with social interactions. Econom J. 2013;16(1):S1–23.
[24] Choi D. Estimation of monotone treatment effects in network experiments. J Amer Stat Assoc. 2017;112(519):1147–55.
[25] Sussman DL, Airoldi EM. Elements of estimation theory for causal effects in the presence of network interference. 2017. arXiv: http://arXiv.org/abs/arXiv:1702.03578.
[26] Forastiere L, Airoldi EM, Mealli F. Identification and estimation of treatment and interference effects in observational studies on networks. J Amer Stat Assoc. 2021;116(534)901–18.
[27] Ugander J, Karrer B, Backstrom L, Kleinberg J. Graph cluster randomization: network exposure to multiple universes. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM; 2013. p. 329–37.
[28] Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. J Amer Stat Assoc. 1952;47(260):663–85.
[29] VanderWeele TJ. Concerning the consistency assumption in causal inference. Epidemiology. 2009;20(6):880–3.
[30] Yu CL, Airoldi EM, Borgs C, Chayes JT. Estimating the total treatment effect in randomized experiments with unknown network structure. Proc National Acad Sci. 2022;119(44):e2208975119.
[31] Basu D. An essay on the logical foundations of survey sampling, Part I. In: Godambe V, Sprott D, editors. Foundations of statistical inferences, Toronto, Canada: Holt, Rinehart and Winston; 1971.
[32] Kojevnikov D, Marmer V, Song K. Limit theorems for network dependent random variables. J Econom. 2021;222(2):882–908.
[33] Leung MP. Causal inference under approximate neighborhood interference. Econometrica. 2022;90(1):267–93.
[34] Leskovec J, Horvitz E. Planetary-scale views on a large instant-messaging network. In: Proceedings of the 17th international conference on World Wide Web; 2008. p. 915–24.
[35] Backstrom L, Boldi P, Rosa M, Ugander J, Vigna S. Four degrees of separation. In: Proceedings of the 4th Annual ACM Web Science Conference; 2012. p. 33–42.
[36] Ugander J, Karrer B, Backstrom L, Marlow C. The anatomy of the Facebook social graph. 2011. arXiv: http://arXiv.org/abs/arXiv:1111.4503.
[37] Travers J, Milgram S. An experimental study of the small world problem. Sociometry. 1969;32(4):425–43.
[38] Radaelli L, Sapiezynski P, Houssiau F, Shmueli E, de Montjoye Y-A. Quantifying Surveillance in the Networked Age: Node-based Intrusions and Group Privacy. 2018. arXiv: http://arXiv.org/abs/arXiv:1803.09007.
[39] Su J, Sharma A, Goel S. The effect of recommendations on network structure. In: Proceedings of the 25th International Conference on World Wide Web; 2016. p. 1157–67.
[40] Spielman DA, Teng S-H. Spectral partitioning works: Planar graphs and finite element meshes. In: Proceedings of 37th Conference on Foundations of Computer Science; 1996. p. 96–105.
[41] Shi J, Malik J. Normalized cuts and image segmentation. IEEE Trans Pattern Anal Machine Intell. 2000;22(8):888–905.
[42] Lee JR, Gharan SO, Trevisan L. Multiway spectral partitioning and higher-order cheeger inequalities. JACM. 2014;61(6):1–30.
[43] Ugander J, Backstrom L. Balanced label propagation for partitioning massive graphs. In: Proceedings of the Sixth ACM International Conference on Web Search and Data Mining; 2013. p. 507–16.
[44] Nishimura J, Ugander J. Restreaming graph partitioning: simple versatile algorithms for advanced balancing. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2013. p. 1106–14.
[45] McPherson M, Smith-Lovin L, Cook JM. Birds of a feather: homophily in social networks. Ann Rev Sociol. 2001;27(1):415–44.
[46] Karger DR, Ruhl M. Finding nearest neighbors in growth-restricted metrics. In: Proceedings of the Thiry-fourth Aannual ACM Smposium on Theory of Computing; 2002. p. 741–50.
[47] Leskovec J, Kleinberg J, Faloutsos C. Graph evolution: densification and shrinking diameters. ACM Trans Knowledge Discovery Data (TKDD). 2007;1(1):2–es.
[48] Linial N, Saks M. Low diameter graph decompositions. Combinatorica. 1993;13(4):441–54.
[49] Alon N, Babai L, Itai A. A fast and simple randomized parallel algorithm for the maximal independent set problem. J Algorithms. 1986;7(4):567–83.

[50] Miller GL, Peng R, Chen Xu S. Parallel graph decompositions using random shifts. In: Proceedings of the Twenty-fifth Annual ACM Symposium on Parallelism in Algorithms and Architectures; 2013. p. 196–203.

[51] Calinescu G, Karloff H, Rabani Y. Approximation algorithms for the 0-extension problem. SIAM J Comput. 2005;34(2):358–72.

[52] Karzanov AV. Minimum 0-extensions of graph metrics. Europ J Combinatorics. 1998;19(1):71–101.

[53] Dahlhaus E, Johnson DS, Papadimitriou CH, Seymour PD, Yannakakis M. The complexity of multiway cuts. In: Proceedings of the Twenty-fourth Annual ACM Symposium on Theory of Computing; 1992. p. 241–51.

[54] Fakcharoenphol J, Rao S, Talwar K. A tight bound on approximating arbitrary metrics by tree metrics. J Comput Syst Sci. 2004;69(3):485–97.

[55] Gupta A, Krauthgamer R, Lee JR. Bounded geometries, fractals, and low-distortion embeddings. In: Proceedings of the 44th Annual IEEE Symposium on Foundations of Computer Science, 2003. IEEE; 2003. p. 534–43.

[56] Gleich DF, Seshadhri C. Vertex neighborhoods, low conductance cuts, and good seeds for local community methods. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2012. p. 597–605.

[57] Yin H, Benson AR, Leskovec J. The local closure coefficient: a new perspective on network clustering. In: Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining; 2019. p. 303–11.

[58] Blelloch GE, Fineman JT, Shun J. Greedy sequential maximal independent set and matching are parallel on average. In: Proceedings of the Twenty-Fourth Annual ACM Symposium on Parallelism in Algorithms and Architectures. ACM; 2012. p. 308–17.

[59] Blelloch GE. Programming parallel algorithms. Commun ACM. 1996;39(3):85–97.

[60] Trotter HF, Tukey JW. Conditional monte carlo for normal samples. In: Symposium on Monte Carlo Methods. 1956. p. 64.

[61] Swaminathan A, Joachims T. The self-normalized estimator for counterfactual learning. In: Advances in neural information processing systems; 2015. p. 3231–39.

[62] Shalizi CR, Thomas AC. Homophily and contagion are generically confounded in observational social network studies. Sociol Meth Res. 2011;40(2):211–39.

[63] Kleinberg J. The small-world phenomenon: an algorithmic perspective. In: Proceedings of the Thirty-second Annual ACM Symposium on Theory of Computing; 2000. p. 163–70.

[64] Watts DJ, Strogatz SH. Collective dynamics of ‘small-world’ networks. Nature. 1998;393(6684):440.

[65] Milgram S. The small world problem. Psychol Today. 1967;2(1):60–7.

[66] Clauset A, Shalizi CR, Newman MEJ. Power-law distributions in empirical data. SIAM Review. 2009;51(4):661–703.

[67] Traud AL, Mucha PJ, Porter MA. Social structure of facebook networks. Phys A Stat Mech Appl. 2012;391(16):4165–80.

[68] Basse GW, Airoldi EM. Model-assisted design of experiments in the presence of network-correlated outcomes. Biometrika. 2018;105(4):849–58.

[69] VonLuxburg U. A tutorial on spectral clustering. Stat Comput. 2007;17(4):395–416.

[70] Traud AL, Kelsic ED, Mucha PJ, Porter MA. Comparing community structure to characteristics in online collegiate social networks. SIAM Review. 2011;53(3):526–43.

[71] Jacobs AZ, Way SF, Ugander J, Clauset A. Assembling thefacebook: Using heterogeneity to understand online social network assembly. In: Proceedings of the ACM Web Science Conference; 2015. p. 1–10.

[72] Eto H, Guo F, Miyano E. Distance-*d* independent set problems for bipartite and chordal graphs. J Comb Optim. 2014;27(1):88–99.

Appendix

A Empirical study of social network growth rates

As described in Section 1, the theoretical analyses in this work are developed under either a bound on the maximum degree $d_{\max}$ of the interference graph or a stronger assumption of bounded geometry, assuming that the interference graph satisfy a restricted growth condition with coefficient κ. Given the central role that bounded geometry plays in our theoretical analysis, in this appendix, we present an empirical study on social network growth statistics.

We use the Facebook100 datasets [67,70,71], a collection of complete Facebook friendship networks at 100 American institutions collected and released in September 2005. The networks are quite diverse, most basically varying in size from 762 to $>30,000$ nodes, which allows us to also understand how the growth statistics can vary with network size and other properties. In Table A1, we present growth statistics from a random subset of 25 networks from the collection, ordered by size n.

The average growth geometry of the full population of 100 networks in the FB100 collection is illustrated in Figure A1. The more fine-grained growth of the Small World and FB-Stanford networks are illustrated in Figure A2.

We here give a concise summary of specific observations from Table A1 and these figures. First, per Table A1, diameter appears to be independent of network size. This is not surprising, as diameter is a fragile metric known to be sensitive to whiskers in the network. Second, the maximum degree, restricted growth coefficient, and network size all appear to be positively correlated. Third, an observation that impacts how we interpret our theoretical results, the restrictive growth coefficients κ are large and all above 100. They are typically 25–50% of the max degree $d_{\max}$.

Looking closer at the results across netowrks in both Table A1 and Figure A1, regarding the average ball-size at each radius r, we see that for $1 \le r \le 3$, the normalized average ball size ratio decreases with n, while for $r \ge 4$, the normalized average ball size saturates near 1. For the maximum ball-size at each radius r, for $r = 1$, $\max|B_1|$ increases with n, while the ratio $\max|B_1|/n$ decreases with n. For $r = 2$, $\max|B_2|$ is usually more than 90% of the nodes, and always at least 70%. Finally, for $r \ge 3$, $\max|B_r|$ is always more than 98% of the nodes.

B Weighted algorithms

B.1 Algorithms

Recall that, in the 1-hop-max clustering algorithm (Algorithm 2), we first independently generate a random number from the uniform distribution and construct a clustering based on these generated random numbers: nodes with higher numbers dominate their neighbors and are more likely to be in the center of a cluster. Since the numbers are generated from a uniform distribution, the probability that a given node generates a larger number than any other is always 1/2, making higher-degree nodes less likely to dominate all their neighbors.

Algorithm 3: Weighted 1-hop-max clustering

```
Input: Graph $G = (V, E)$, node weights $\boldsymbol{w} \in \mathbb{R}_+^n$
Output: Graph clustering $\boldsymbol{c} \in \mathbb{R}^n$
for $i \in V$ do
    $X_i \leftarrow \beta(w_i, 1)$
for $i \in V$ do
    $c_i \leftarrow \max([X_j \text{ for } j \in B_1(i)])$
return $\boldsymbol{c}$
```

Table A1: Network statistics of 25 randomly selected networks from the Facebook-100 collection, listing the number of nodes n, number of edges m, diameter D, average degree $\bar{d} = 2m/n$ maximum degree $d_{\max}$, and the restrictive growth coefficient κ. In addition, we present the average ball size of each radius until $r = 4$ and the maximum ball size until $r = 3$, both normalized by the number of nodes in the network

University	n	m	D	$\bar{d}$	$d_{\max}$	κ	$\text{avg}_{i\in V}\lvert B_r(i)\rvert/n$				$\max_{i\in V}\lvert B_r(i)\rvert/n$		
							$r = 1$	$r = 2$	$r = 3$	$r = 4$	$r = 1$	$r = 2$	$r = 3$
Caltech36	762	16,651	6	43.70	248	102	0.0587	0.6448	0.9618	0.9986	0.3268	0.9357	0.9987
Swarthmore42	1,657	61,049	6	73.69	577	149	0.0451	0.6594	0.9790	0.9991	0.3488	0.9523	0.9994
Trinity100	2,613	111,996	6	85.72	404	148	0.0332	0.5735	0.9684	0.9980	0.1550	0.9529	0.9981
Wellesley22	2,970	94,899	8	63.91	746	160	0.0219	0.4781	0.9221	0.9938	0.2515	0.9232	0.9963
Pepperdine86	3,440	152,003	9	88.37	674	301	0.0260	0.5400	0.9420	0.9936	0.1962	0.9340	0.9968
Mich67	3,745	8,1901	7	43.74	419	157	0.0119	0.2945	0.8644	0.9906	0.1121	0.8179	0.9939
Rice31	4,083	184,826	6	90.53	581	256	0.0224	0.5434	0.9675	0.9994	0.1425	0.9224	0.9995
Wake73	5,366	279,186	9	104.06	1341	671	0.0196	0.5141	0.9561	0.9973	0.2501	0.9702	0.9983
UChicago30	6,561	208,088	10	63.43	1624	813	0.0098	0.3375	0.8661	0.9832	0.2477	0.9218	0.9938
UC64	6,810	155,320	8	45.62	660	282	0.0068	0.2103	0.7921	0.9777	0.0971	0.8026	0.9872
WashU32	7,730	367,526	8	95.09	1794	898	0.0124	0.4244	0.9328	0.9957	0.2322	0.9578	0.9988
Yale4	8,561	405,440	9	94.72	2517	1259	0.0112	0.4132	0.9082	0.9909	0.2941	0.9429	0.9961
Georgetown15	9,388	425,619	11	90.67	1235	618	0.0098	0.3546	0.8946	0.9867	0.1317	0.8838	0.9923
Northwestern25	10,537	488,318	9	92.69	2105	1053	0.0089	0.3624	0.9126	0.9936	0.1999	0.9503	0.9974
Stanford3	11,586	568,309	9	98.10	1172	587	0.0086	0.3375	0.8529	0.9841	0.1012	0.8443	0.9906
USF51	13,367	321,209	8	48.06	897	319	0.0037	0.1433	0.7428	0.9785	0.0672	0.7441	0.9915
Northeastern19	13,868	381,919	9	55.08	968	393	0.0040	0.1721	0.8036	0.9850	0.0699	0.8017	0.9930
UCSD34	14,936	443,215	9	59.35	2165	1083	0.0040	0.1877	0.8158	0.9868	0.1450	0.9129	0.9970
UMass92	16,502	519,376	8	62.95	3684	1843	0.0039	0.2068	0.8621	0.9939	0.2233	0.9575	0.9994
UConn91	17,206	604,867	8	70.31	1709	855	0.0041	0.2082	0.8754	0.9946	0.0994	0.9167	0.9980
Auburn71	18,448	973,918	7	105.59	5160	2581	0.0058	0.3672	0.9528	0.9989	0.2798	0.9795	0.9998
Maryland58	20,829	744,832	7	71.52	3784	1893	0.0035	0.2031	0.8631	0.9937	0.1817	0.9474	0.9989
Wisconsin87	23,831	835,946	9	70.16	3484	1620	0.0030	0.1888	0.8607	0.9929	0.1462	0.9361	0.9985
Indiana69	29,732	1,305,757	8	87.84	1358	479	0.0030	0.1794	0.8624	0.9941	0.0457	0.8102	0.9960
MSU24	32,361	1,118,767	8	69.14	5267	2634	0.0022	0.1413	0.8222	0.9913	0.1628	0.9478	0.9989

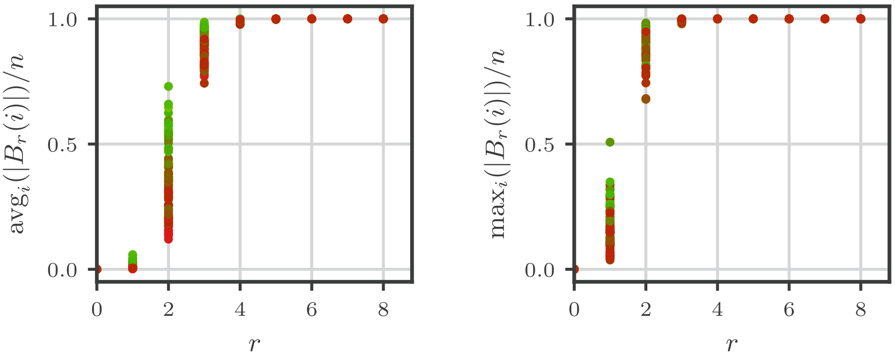


Figure A1: Mean and max ball size at each radius r, for each of the 100 networks in the FB100 collection. The points are colored based on the size of the networks, n, with smaller networks green and larger networks red.

Our proposed fix to this problem is to change the first step of the algorithm, generating numbers X_i from a different nonuniform distribution at each node. Let each node i be associated with a weight $w_i > 0$ and then generate its number from a Beta distribution, $X_i \sim \beta(w_i, 1)$. The full algorithm of weighed 1-hop-max is given in Algorithm 3.

To understand the intuition behind the weighted scheme, we first note the following basic and well-known properties of the beta distribution, proven for completeness in Appendix C.

Theorem B.1. *For independent random variables* $X_i \sim \beta(w_i, 1)$, $X_j \sim \beta(w_j, 1)$, *we have*
(*a*) $\mathbb{P}[X_i > X_j] = \frac{w_i}{w_i + w_j}$,
(*b*) $\max\{X_i, X_j\} \sim \beta(w_i + w_j, 1)$.

According to part (a) of Theorem B.1, a node with a larger weight is more likely to generate a larger number. Thus, by adopting larger weights at high degree nodes, we can make the large degree nodes more likely to dominate their neighbors, correcting their disadvantage in the unweighted scheme.

Algorithm 4: Weighted 3-net clustering

```
Input: Graph $G = (V, E)$, node weights $\boldsymbol{w} \in \mathbb{R}^n_+$.
Output: Graph clustering $\boldsymbol{c} \in \mathbb{R}^n$
for $i \in V$ do
    $X_i \leftarrow \beta(w_i, 1)$
$\pi \leftarrow \text{arg sort}([X_i]_{i \in V}, \textit{descend})$
$S \leftarrow \varnothing$, unmark all nodes
for $i \in \pi$ do
    if $i$ is unmarked then
        $S \leftarrow S \cup \{i\}$
        for $j \in B_2(i)$ do
            mark node $j$ if it is unmarked yet
for $i \in V$ do
    $c_i \leftarrow \text{arg min}\{j \in S, j \rightarrow \textbf{dist}(i, j)\}$, i.e., the id of the node in $S$ with shortest graph distance to $i$ (arbitrary tie breaking)
return $\boldsymbol{c}$
```

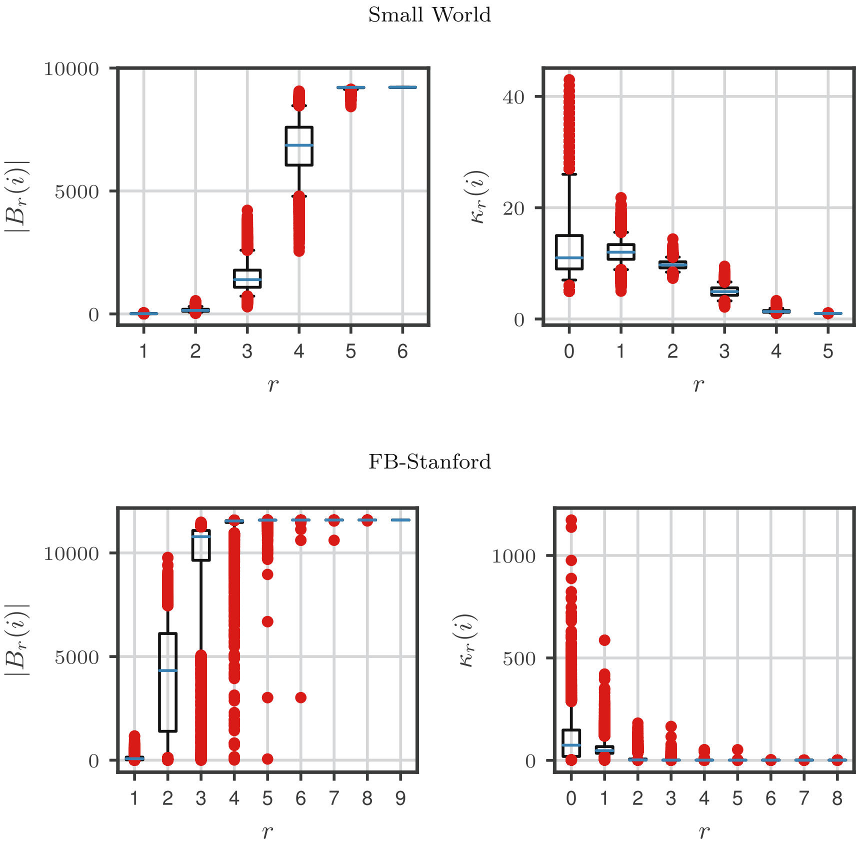


Figure A2: The ball sizes $|B_r(i)|$ and local growth coefficient $\kappa_i = |B_{r+1}(i)| / |B_r(i)|$ of each node i in the synthetic Small World and real-world FB-Stanford network used for Section 6.

This idea of node weighting can also be applied to 3-net clustering. In the unweighted version, we first generate a uniform random ordering of all nodes, which is used to form a seed set and partition the network. In a uniform random ordering where each node has an equal probability of ranking first, large degree nodes are at disadvantage of being selected into the seed set and being the center of a cluster and thus less likely to be network exposed. To compensate for this disadvantage, we can generate a nonuniform random ordering where large degree nodes are more likely to rank high. A nonuniform random ordering can be generated by a combination of Beta-distributed samples and sorting. Specifically, if each node i is associated with a weight w_i, then we can first generate $X_i \sim \beta(w_i, 1)$ and sort the samples in decreasing order. In this way, nodes associated with a larger weight are more likely to rank higher after sorting. Formally this weighted 3-net clustering algorithm is given in Algorithm 4.

We note two connections between the weighted 3-net and 1-hop-max clustering algorithms and their original unweighted versions. First, the weighted version can be considered an extension of the unweighted algorithms: when all nodes have the same weight, the weighted 3-net and 1-hop-max algorithm are equivalent to the original algorithm. Second, for either 3-net or 1-hop-max clustering, the distribution of the random clustering returned from the unweighted and weighted algorithms have the same support, i.e., for clusterings that has nonzero probability of being generated from the unweighted version, the probability of being generated from the weighted version is also nonzero, and *vice versa*. While the support of possible clusterings is the same, certain clusterings are more or less likely to be generated in the weighted versions than the unweighted versions. Consequently, conditioning on the generated clustering and using it in a GCR scheme, there is no difference between which version is used to generate the clustering. However, in RGCR, which is based on a distribution of clusterings, the weighted version might have superior properties due to its node-level adjustments.

B.2 Properties with arbitrary node weights

In this section, we discuss properties of the weighted 3-net and 1-hop-max algorithms with an arbitrary set of node weights. The result motivates our discussion on good choices of node weights in section that follows.

First, we have the following lower bound on exposure probabilities at each node. Similar to Theorem 4.2, the result is based on analyzing the probability that a node is ranked first in its 2-hop-neighborhood. The proof is given in Appendix C.

Theorem B.2. *With the weighted 3-net or 1-hop-max random clustering algorithm, using either independent or complete randomization at the cluster level, the full-neighborhood exposure probabilities for any node i satisfy*

$$\mathbb{P}[E_i^{\mathbf{1}}|\mathcal{P}] \geq \frac{w_i}{\sum_{j\in B_2(i)} w_j} \cdot p, \quad \mathbb{P}[E_i^{\mathbf{0}}|\mathcal{P}] \geq \frac{w_i}{\sum_{j\in B_2(i)} w_j} \cdot (1-p).$$

When all nodes have equal weights, then the weighted 3-net and 1-hop-max algorithm degenerates to the original version, making Theorem B.2 a generalization of Theorem 4.2.

Computing the exposure probability of each node might now be challenging, but we again show that Monte Carlo estimation, as in equation (10), can efficiently achieve low relative error.

Theorem B.3. *By using either weighted 3-net or weighted 1-hop-max random clustering algorithm, and with K Monte-Carlo trials, for any node i, the relative error of the Monte-Carlo estimator is upper bounded in MSE as follows*:

$$\mathbb{E}\left[\left.\frac{\hat{\mathbb{P}}[E_i^{\mathbf{1}}|\mathcal{P}] - \mathbb{P}[E_i^{\mathbf{1}}|\mathcal{P}]}{\mathbb{P}[E_i^{\mathbf{1}}|\mathcal{P}]}\right|\mathcal{P}\right]^2 \leq \frac{1}{Kp} \cdot \frac{\sum_{j\in B_2(i)} w_j}{w_i}.$$

As stated earlier, stratified sampling can also be adapted for the weighted clustering methods. Similar to the procedure in Section 4.2.2, we generate Kn clustering samples $\{\mathbf{c}^{(k,i)}\}$, where in clusterings $\mathbf{c}^{(k,i)}$, $k \in \{1, 2, \dots, n\}$, node i is "favored" and deterministically placed first. Note that the likelihood of node i naturally generating the largest draw is proportional to w_i, per Theorem B.1. The sample $\mathbf{c}^{(k,j)}$ should be weighted accordingly. The estimated exposure probabilities should then be

$$\hat{\mathbb{P}}[E_i^{\mathbf{z}}|\mathcal{P}] = \frac{\sum_{k=1}^{K}\sum_{j=1}^{n} w_j \mathbb{P}[E_i^{\mathbf{z}}|\mathbf{c}^{(k,j)}]}{K\sum_{j=1}^{n} w_j}.$$

Again this stratified method is preferred over Monte Carlo estimation with independent samples since it guarantees positivity in the estimated exposure probabilities and reduces variance in the probability estimation.

B.3 Choice of node weights

With the node-level flexibility in the weighted 3-net and 1-hop-max clustering, a natural subsequent question is to find a good choice of node weights. In this section, we discuss two heuristics that lead to different sets of node weights. The first heuristic suggests node weights based on the eigenvector of an eigenvalue problem associated with the network's squared adjacency matrix. The second heuristic suggests uniform weights, i.e., the unweighted versions of the algorithms.

B.3.1 Maximizing the minimal exposure probability lower bound

As is discussed in the previous sections, high-degree nodes are less likely than low-degree nodes to be network exposed using the unweighted 3-net or 1-hop-max clustering. To correct this disadvantage, it might be ideal if all nodes have the same exposure probability, or at least the same lower bound.

Given a graph $G = (V, E)$, let $G_2 = (V, E_2)$ denote the "squared" graph, i.e., with the same node set V, and an edge $(i, j) \in E_2$ if $i \in B_2(j)$ in the original network. The adjacency matrix of G_2 is an irreducible nonnegative matrix, and according to the Perron-Frobenius theorem, its spectral radius, denoted as λ^*, is also its largest positive eigenvalue. Moreover, for the eigenvector $\boldsymbol{w}^*$ associated with this eigenvalue, i.e.,

$$\sum_{j \in B_2(i)} w_j^* = \lambda^* w_i^*, \tag{A1}$$

all the elements w_i^* are positive. Therefore, $\boldsymbol{w}^*$ provides a valid set of node weights, which we call the *spectral weights*.

By using these spectral weights in the weighted 3-net or 1-hop-max scheme, we show that as a corollary of Theorem B.2 and equation (A1) (the proof logic is identical), all nodes now have the same exposure probability lower bound.

Theorem B.4. *With the spectral-weighted 3-net or 1-hop-max random clustering algorithm, using either independent or complete randomization at the cluster level, the full-neighborhood exposure probabilities for any node i satisfy*

$$\mathbb{P}[E_i^{\mathbf{1}} | \mathcal{P}] \geq \frac{p}{\lambda^*}, \quad \mathbb{P}[E_i^{\mathbf{0}} | \mathcal{P}] \geq \frac{1-p}{\lambda^*},$$

a uniform lower bound on the full neighborhood exposure probability of all nodes.

We then have the following corollary (of Theorem 4.7) upper bound on the variance of HT GATE estimators using RGCR with spectral-weighted 1-hop-max random clustering.

Theorem B.5. *Using RCGR with spectral-weighted 1-hop-max clustering, if every node's response is within* $[0, \bar{Y}]$, *then*

$$\mathbf{Var}[\hat{\mu}_{\mathcal{P}}(\mathbf{1})] \leq \frac{1}{n} \cdot \bar{Y}^2 \lambda^* (d_{\max} + 1) \kappa^3 p^{-1},$$

for both independent and complete cluster-level randomization.

Proof. We first note that, with an identical proof, one can verify that Theorem 4.6 also hold with the weighed 1-hop-max clustering with any weights $\boldsymbol{w}$. Now similar to the proof of Theorem 4.7, we have

$$\mathbf{Var}[\hat{\mu}_{\mathcal{P}}(\mathbf{1})] \leq \frac{\bar{Y}^2}{n^2} \sum_{i=1}^{n} \left[\frac{\lambda^*}{p} \cdot |B_4(i)| \right] \leq \frac{1}{n} \cdot \bar{Y}^2 \lambda^* (1 + d_{\max}) \kappa^3 p^{-1},$$

where the first inequality is due to the exposure probability lower bound in Theorem B.4. □

As a final corollary, we have the following upper bound on the variance of the HT GATE estimator, by a proof identical to that of Theorem 4.8.

Theorem B.6. *Using RCGR with spectral-weighted 1-hop-max clustering, if every node's response is within* $[0, \bar{Y}]$, *then*

$$\mathbf{Var}[\hat{\tau}_{\mathcal{P}}] \leq \frac{2}{n} \cdot \bar{Y}^2 \lambda^* (d_{\max} + 1) \kappa^3 (p^{-1} + (1-p)^{-1}),$$

for both independent and complete cluster-level randomization.

Of note, according to the Perron-Frobenius theorem, we also have

$$\lambda^* \leq \max_i (|B_2(i)|) \leq (d_{\max} + 1)\kappa.$$

As a result, this variance upper bound using spectral-weighted 1-hop-max clustering can be used to furnish the variance upper bound for the unweighted 1-hop-max clustering (Theorem 4.8) as well. These final inequalities are not necessarily strict improvements – they become equalities for a regular graph – but in practical settings, they can lead to sizable improvements over unweighted clustering methods.

Having the same exposure probabilities at each node is ideal, whereas we note that our spectral weights do not exactly achieve that. They merely maximize a uniform lower bound, the lower bound given in Theorem B.2. The tightness of this lower bound might not be equal at each node, since it only captures the scenario when the node is at the interior of a cluster. If a node is not in the interior and thus adjacent to multiple clusters, then a lower-degree node is likely to be adjacent to fewer clusters and thus still has higher exposure probability. Therefore, in reality, one might use a weight where high-degree nodes are even more aggressively favored than under spectral weighting. In Section 6, besides uniform weight and spectral weight, we also consider weighting each node by their degree directly. Simulation results show that this aggressive degree weight strategy usually yields lower variance than both uniform weights and spectral weights.

B.3.2 Minimizing a variance proxy

The aforementioned heuristic is intended to reduce the estimator variance, but a more direct approach would be to find the optimal weights that minimize the actual estimator variance.

That said, optimizing the variance, as formulated in equations (4) to (6), is challenging because (i) it consists of cross-terms associated with the joint exposure probability of node pairs that are hard to analyze, and (ii) the nodes' response is unknown before the experiment, but can play a significant role in determining the variance. One compromise is to use a proxy objective function that resembles the variance formula. We consider the following function

$$H(\boldsymbol{w}) = \sum_{i=1}^{n} \frac{1}{\mathbb{P}[E_i^{\mathbf{1}} | \mathcal{P}, \boldsymbol{w}]}, \tag{A2}$$

which overlooks the cross-terms and assumes a uniform response from all nodes.

Note that this proxy function is also intractable since one cannot efficiently compute the exposure probability of each node given the weights. However, one can obtain an upper bound of $H(\boldsymbol{w})$ using the exposure probability lower bound in Theorem B.2, i.e.,

$$\bar{H}(\boldsymbol{w}) = \frac{1}{p} \sum_{i=1}^{n} \frac{\sum_{j \in B_2(i)} w_j}{w_i}, \tag{A3}$$

and attempt to minimize this variance surrogate. We have the following result.

Theorem B.7. *The minimum of $\bar{H}(\boldsymbol{w})$ of is achieved with uniform weighting, i.e.,*

$$\bar{H}(\mathbf{1}) \le \bar{H}(\boldsymbol{w})$$

for any $\boldsymbol{w} \in \mathbb{R}_+^n$.

The first heuristic increased the exposure probability of high degree nodes, but came at the cost of decreasing the exposure probabilities of low degree nodes. Thus, it is not certain whether this heuristic would actually reduces variance. It is therefore interesting that under this second heuristic, if trusting $\bar{H}(\boldsymbol{w})$ as a surrogate, according to Theorem B.7, the optimal weights are the uniform weights, corresponding to the unweighted 3-net or 1-hop-max clustering algorithms.

The construction of the surrogate variance, $\bar{H}(\boldsymbol{w})$, is based on the lower bound exposure probability in Theorem B.2, whose tightness varies between high and low degree nodes. Specifically, for a low-degree node i, its exposure probability trivially satisfies $\mathbb{P}[E_i^{\mathbf{1}}] \ge p^{d_i}$, a bound that could potentially be much higher than the lower bound $pw_i / (\sum_{j \in B_2(i)} w_j)$. Consequently, assigning i a low weight would not significantly increase its inverse exposure probability as penalized in $\bar{H}(\boldsymbol{w})$. Therefore, for a real-world network with a wide range of node degrees, it can certainly still be a good idea to use a weighted clustering algorithm with high weights for high degree nodes. Our simulations presented in Section 6 further demonstrate this intuition.

C Proofs

C.1 Proof of Theorem 4.1

Proof. In the first step of the algorithm (line 1–2), every node independently generates a random number which can be executed in parallel. Therefore, the depth is $O(1)$ and the work is $O(n)$. In the second step (line 3–4), each node i computes the maximum of $|B_1(i)| = O(d_i)$ numbers, and parallel implementation of this max procedure requires $O(\log(d_i))$ depth and $O(d_i)$ work [59]. Moreover, note that the maximization task at different nodes can also be executed in parallel, and thus in the second step, the total work is $\sum_i O(d_i) = O(m)$ and the total depth is $\max_i O(\log(d_i)) = O(\log(d_{\max}))$. Combining both steps, the total work is $O(m)$ and total depth is $O(\log(d_{\max}))$. □

C.2 Proof of Theorem 4.2

Proof. We first show that the probability that node i is in the interior of a cluster is lower bounded by $1/|B_2(i)|$, i.e.,

$$\mathbb{P}[\forall j \in B_1(i), C_j = C_i | \mathcal{G}] \geq 1/|B_2(i)|.$$

To this end, we consider a sufficient condition of this event, for the 3-net clustering and 1-hop-max clustering separately. With 3-net clustering, in the first step when we generate a random ordering of all nodes, if node i ranks first among $B_2(i)$, then it must be outside the 2-hop neighborhood of every node ahead of itself in the ordering. Therefore, it is left unmarked and will be added as a seed, and thus, its neighbors must belong to the same cluster as i. The probability of this situation, i.e., node i is ranked first amongst $B_2(i)$, is $1/|B_2(i)|$ since we are generating the orderings uniformly. When the 1-hop-max clustering algorithm is used, in the first step where every node independently generates a random number, if node i generates the largest number amongst $B_2(i)$, then we have $C_j = X_i$ for every $j \in B_1(i)$. This scenario happens with probability $1/|B_2(i)|$ since the random numbers generated at each node is *i.i.d.*

Now we derive the results in the theorem. Conditioning on the event that node i is in the interior of a cluster, it is full-neighborhood exposed to the treatment condition if this cluster is assigned into the treatment group, which happens with probability p. Therefore, by combining the result in the previous paragraph, we have

$$\mathbb{P}[E_i^{\mathbf{1}} | \mathcal{G}] \geq \mathbb{P}[\forall j \in B_1(i), C_j = C_i | \mathcal{G}] \cdot p \geq \frac{p}{|B_2(i)|} \geq \frac{p}{(1 + d_{\max})\kappa}.$$

With the same reasoning, it can be easily verified that $\mathbb{P}[E_i^{\mathbf{0}} | \mathcal{P}] \geq \frac{1-p}{|B_2(i)|} \geq \frac{1-p}{(1+d_{\max})\kappa}$, and this proof applies to both independent and complete randomization scenario. □

C.3 Proof of Theorem 4.3

Proof. By symmetry, we only need to consider treatment (control is analogous). Suppose node i generates the kth largest value in $B_2(i)$ for some $1 \leq k \leq |B_2(i)|$. If $k \leq d_i$, then with the generated clustering $\boldsymbol{c}$, we have

$\mathbb{P}[E_i^{\mathbf{1}}|\boldsymbol{c}] \geq p^k$ since node i is adjacent to at most k clusters. If $k \geq 1 + d_i$, then we have $\mathbb{P}[E_i^{\mathbf{1}}|\boldsymbol{c}] \geq p^{1+d_i}$, the trivial lower bound when every node in $B_1(i)$ is assigned to a different cluster. By combining the two scenarios, we have

$$\mathbb{P}[E_i^{\mathbf{1}}|\mathcal{P}] = \mathbb{E}[\mathbb{P}[E_i^{\mathbf{1}}|\boldsymbol{C}]|\mathcal{P}] \geq \sum_{k=1}^{d_i} \frac{p^k}{|B_2(i)|} + \sum_{k=1+d_i}^{|B_2(i)|} \frac{p^{1+d_i}}{|B_2(i)|} = \frac{p}{|B_2(i)|} \cdot \left[\frac{1-p^{d_i}}{1-p} + p^{d_i}(|B_2(i)|-d_i)\right],$$

where we use the fact that the probability of node i generating the kth largest number in $B_2(i)$ in the unweighted 1-hop-max clustering algorithm is $1/|B_2(i)|$. As a final step, as $|B_2(i)| - d_i \geq 1/(1-p)$ we conclude $\mathbb{P}[E_i^{\mathbf{1}}|\mathcal{P}] \geq \frac{p}{|B_2(i)|} \cdot \frac{1}{1-p}$. □

C.4 Proof of Theorem 4.4

Proof. Here, we provide a proof for the specific case $r = 3$, corresponding to the 3-net clustering algorithm used in the original analysis of the graph cluster randomization scheme. The proof for other r can be constructed analogously.

We present a polynomial-time reduction to network exposure probability computation from the minimum maximal distance-3 independent set (MD3IS) problem, which is known to be NP-complete [72]. This problem is as follows.

- Minimum Maximal Distance-3 Independent Set problem (decision version): Given a graph $G = (V, E)$ and an integer $K \leq |V|$, determine whether there is a maximal distance-3 independent set of size no greater than K, i.e., a subset of nodes $V_s \subseteq V$ such that
  (a) for any pair of nodes $u, v \in V_s$, their graph distance $\textbf{dist}(u, v) \geq 3$ (i.e., a distance-3 independent set);
  (b) V_s is not a subset of any other distance-3 independent set (i.e., maximal);
  (c) $|V_s| \leq K$.

For any instance of the MD3IS problem with input $G = (V, E)$ and K, we construct the following instance of network exposure probability computation problem on $\tilde{G} = (\tilde{V}, \tilde{E})$ around a node i_0, and $p = 1/(2|V| + 1)!$. We construct $\tilde{G}$ as follows, where $\sqcup$ is indicates a multi-set union. First, we make two copes if V and add i_0 as an additional node. Let E_1 be edges connecting the corresponding nodes in V and the copy V'. Let V have edges E from the original graph, while let V' be a clique. Finally, connect every node in V' to i_0. More formally:

- $\tilde{V} = V \sqcup V' \sqcup \{i_0\}$, where there is a bijection ϕ between V and V' (and consequently $|V| = |V'|$).
- $\tilde{E} = E \sqcup E_1 \sqcup E_2 \sqcup E_3$, where
- $E_1 = \{(u, \phi(u))|u \in V\}$, i.e., connecting the every pair of corresponding nodes in V and V'.
- $E_2 = \{(u', v')|u', v' \in V', u' \neq v'\}$, i.e., connecting every pair of nodes in V'.
- $E_3 = \{(u', i_0)\}$, i.e., connecting node i_0 with every node in V'.

Before connecting this exposure probability computation problem with the original MD3IS instance, we first present several properties of the maximal distance-3 independent sets of $\tilde{G}$. The proofs are found at the end of this section.

Lemma C.1. *For any node subset in the original graph $V_s \subset V$, it is a maximal distance-3 independent set of $\tilde{G}$ if and only if it is a maximal distance-3 independent set of G.*

Lemma C.2. *Any maximal distance-3 independent set of $\tilde{G}$, unless it is also a maximal distance-3 independent set of G, contains only a single-node $u' \in V' \sqcup \{i_0\}$.*

Lemma C.2 illustrates the two types of maximal distance-3 independent set of $\tilde{G}$. With each of the types as the seed set in 3-net clustering, the following lemma states the conditional exposure probability of node i_0.

Lemma C.3. *For a random sample of 3-net clustering c on $\tilde{G}$, let V_s be the seed set, then*

- *with probability $\frac{|V|+1}{2|V|+1}$, $V_s = \{u'\}$ for some $u' \in V' \sqcup \{i_0\}$, and $\mathbb{P}[E_{i_0}^{\mathbf{1}}|c] = p$;*
- *with probability $\frac{|V|}{2|V|+1}$, V_s is a maximal distance-3 independent set of G, and $\mathbb{P}[E_{i_0}^{\mathbf{1}}|\mathbf{c}] = p^{|V_s|}$.*

Now we have the following key result connecting the exposure probability value to the MD3IS problem.

Lemma C.4. *The exposure probability value of node* i$_0$ *under randomized 3-net clustering corresponds to the MD3IS problem instance as the following*:

(1) *If there exists a maximal distance-3 independent set of size ≤K, then $\mathbb{P}[E_{i_0}^{\mathbf{1}}] \geq \frac{n+1}{2n+1} \cdot p + p^{K+1}$.*

(2) *If every maximal distance-3 independent set is of size ≥$K + 1$, then $\mathbb{P}[E_{i_0}^{\mathbf{1}}] < \frac{n+1}{2n+1} \cdot p + \frac{1}{2} \cdot p^{K+1}$.*

Combining the aforementioned results, we show that exact computation of the exposure probability solves the MD3IS instance. Suppose there is a polynomial algorithm such that, for any graph $\tilde{G}$, node i_0, treatment probability p, and any precision $\varepsilon > 0$, it outputs the treatment exposure probability of precision ε in time poly(sizeof($\tilde{G}$), sizeof(p), $\log_2(1/\varepsilon)$). Choosing $\varepsilon = p^{K+1}/4$, we compare the output probability P^{out} with $P^* = \frac{n+1}{2n+1} \cdot p + \frac{3}{4} p^{K+1}$. If $P^{\text{out}} \geq P^*$, due to the precision ε, we have

$$\mathbb{P}[E_{i_0}^{\mathbf{1}}] \geq P^* - \varepsilon = \frac{n+1}{2n+1} \cdot p + \frac{1}{2} \cdot p^{K+1}.$$

Now according to scenario (2) in Lemma C.4, the MD3IS instance must have a maximal distance-3 independent set of size ≤K. If $P^{\text{out}} < P^*$, again due to the precision ε, we have

$$\mathbb{P}[E_{i_0}^{\mathbf{1}}] < P^* + \varepsilon = \frac{n+1}{2n+1} \cdot p + p^{K+1},$$

and thus, the MD3IS instance cannot have a maximal distance-3 independent set of size ≤K according to scenario (1) in Lemma C.4.

Given this reduction, we must show that the reduction from the MD3IS is polynomial. The size of the MD3IS problem is sizeof(G) = $|V| + |E|$. For the constructed graph $\tilde{G}$, we have $|\tilde{V}| = 2|V| + 1$, and $|\tilde{E}| = |E| + |V|^2 + 2|V|$, and thus constructing $\tilde{G}$ takes polynomial time and space. In addition, we have

$$\text{sizeof}(p) = \log_2(1/p) = \log_2((2|V| + 1)!) = O(|V|\log(|V|)),$$

and the log value of required precision $\log(1/\varepsilon) = (K + 1)\log_2(1/p) = O(K|V|\log(|V|))$ is also polynomial in the size of the MD3IS problem. In summary, the reduction is a polynomial reduction. □

Before presenting the proof of Lemma C.1, we first give an auxiliary result.

Lemma C.5. *For any distinct nodes $u, v \in V$, we have*

$$\mathbf{dist}_{\tilde{G}}(u, v) \geq 3 \quad \Leftrightarrow \quad \mathbf{dist}_G(u, v) \geq 3. \tag{A4}$$

Proof. "⇒". Since all the edges in G are preserved in $\tilde{G}$, we have $\mathbf{dist}_{\tilde{G}}(u, v) \leq \mathbf{dist}_G(u, v)$. Consequently, if $\mathbf{dist}_{\tilde{G}}(u, v) \geq 3$, we must have $\mathbf{dist}_G(u, v) \geq 3$.

"⇐=". Note that node $u \in V$ is not directly connected to any node in $V' \sqcup \{i_0\}$ other than $\phi(u)$, and thus for any path connecting nodes u and v through nodes in $V' \sqcup \{i_0\}$, the path length is at least 3. Therefore, if $\mathbf{dist}_{\tilde{G}}(u, v) < 3$, the shortest path can only consists of nodes and edges in the original graph G, and thus, $\mathbf{dist}_G(u, v) < 3$. □

Proof (Lemma C.1). Note that a corollary of Lemma C.5 is the following:

- V_s is a distance-3 independent set of $\tilde{G}$ if and only if it is also a distance-3 independent set of G.

This is the lemma result regarding only the independent set condition.

We first show the sufficiency in Lemma C.1. If V_s is a maximal distance-3 independent set of G, then according to the argument presented earlier, V_s is also a distance-3 independent set of $\tilde{G}$, and we now show it is maximal, i.e., introducing any other node w into V_s would break the distance-3 independent set condition. There are two scenarios: $w \in V$ and $w \in V' \sqcup \{i_0\}$. First, for any node $w \in V$, $w \notin V_s$, since V_s is a maximal distance-3 independent set in G, there exists node $u \in V_s$ such that $\mathbf{dist}_G(u, w) \leq 2$, and thus, $\mathbf{dist}_{\tilde{G}}(u, w) \leq 2$ according to Lemma C.5. Therefore, introducing node v into V_s makes V_s no longer a distance-3 independent set of $\tilde{G}$. Second, note that V_s must be nonempty and denote u as an arbitrary node therein. For any node $w \in V' \sqcup \{i_0\}$, due to the length-2 path $(w, \phi(u), u)$, we have $\mathbf{dist}_{\tilde{G}}(u, w) \leq 2$, and thus, one cannot include node v to V_s while maintaining that it is a distance-3 independent set.

Next we show the necessity in Lemma C.1. If $V_s \subset V$ is a maximal distance-3 independent set of $\tilde{G}$, again V_s is a distance-3 independent set of G and we only need to show its maximality. Suppose it is not maximal, and there is another node $w \in V$, $w \notin V_s$ such that $V_s \sqcup \{w\}$ is a distance-3 independent set of $\tilde{G}$, then according to the argument presented earlier, $V_s \sqcup \{w\}$ is also a distance-3 independent set of $\tilde{G}$. This means that V_s is not maximal in $\tilde{G}$, which creates a contradiction. □

Proof (Lemma C.2). Suppose V_s is a maximal distance-3 independent set of $\tilde{G}$ but not of G. Note that V_s must contain a node $u' \in V' \sqcup \{i_0\}$ because otherwise with $V_s \subset V$, according to Lemma C.1, V_s must also be a maximal distance-3 independent set of G. Now due to the fact that $V' \sqcup \{i_0\}$ induces a complete graph, and thus any other node in $V' \sqcup \{i_0\}$ cannot be included in V_s. Moreover, for any node $u \in V$, we have $\mathbf{dist}_{\tilde{G}}(u, u') \leq 2$ due to the path $(w, \phi(u), u)$, and thus, V_s cannot contain any node in V. In summary, u' is the only node in V_s. □

Proof (Lemma C.3). In 3-net clustering, we use the randomized greedy algorithm to construct a maximal distance-3 independent set as the seed set. With probability $\frac{|V|+1}{2|V|+1}$, the first randomly selected node is from $V' \sqcup \{i_0\}$, and the seed set is a single-node set according to Lemma C.2, and the conditional network exposure probability is $\mathbb{P}[E_{i_0}^{\mathbf{1}}|c] = p^1$.

Similarly, with probability $\frac{|V|}{2|V|+1}$, the first randomly selected node is from V, and the seed set is a maximal distance-3 independent set of G. Now we show the conditional network exposure probability. Note that in 3-net clustering, for any seed node $u \in V_s$, node $\phi(u) \in V'$ belong to the same cluster as u's due to it being a directed neighbor of node u. Since $B_1(i_0) = V' \sqcup \{i_0\}$, $B_1(i_0)$ has nonempty intersection with all the $|V_s|$ clusters in this 3-net clustering, and thus, $\mathbb{P}[E_{i_0}^{\mathbf{1}}|c] = p^{|V_s|}$. □

Proof (Lemma C.4). For the first result, if the MD3IS problem has a maximal distance-3 independent set of size $\leq K$, then this set will be selected as the seed set in 3-net clustering with probability no less than $\frac{1}{(2n+1)!}$. Recall that $p = 1/(2|V| + 1)!$ by construction. Thus,

$$\mathbb{P}[E_{i_0}^{\mathbf{1}}] \geq \frac{n+1}{2n+1} \cdot p + \frac{1}{(2n+1)!} \cdot p^K = \frac{n+1}{2n+1} \cdot p + p^{K+1}$$

where the first term comes from the case when a single-node maximal distance-3 independent set is used as seeds.

For the second result, if every maximal distance-3 independent set of size $\geq K + 1$ in the MD3IS problem, then any class-2 maximal independent set of $\tilde{G}$ is of size $\geq K + 1$, and thus,

$$\mathbb{P}[E_{i_0}^{\mathbf{1}}] \leq \frac{n+1}{2n+1} \cdot p + \frac{n}{2n+1} \cdot p^{K+1} < \frac{n+1}{2n+1} \cdot p + \frac{1}{2} \cdot p^{K+1}.$$ □

C.5 Proof of Theorem 4.5

Proof. Unless otherwise stated, all the expectation and variance in this proof are taken conditioned on the random clustering distribution $\mathcal{P}$ in use; thus, to simplify the notation, we sometimes discard the conditional notation in the expectation and variance symbol.

Let $\boldsymbol{C}$ be a random clustering generated from 3-net or 1-hop-max algorithm, we first show that

$$\mathbf{Var}\boldsymbol{C} \sim \mathcal{P}\left[\frac{\mathbb{P}[E_i^{\mathbf{1}}|\boldsymbol{C}] - \mathbb{P}[E_i^{\mathbf{1}}|\mathcal{P}]}{\mathbb{P}[E_i^{\mathbf{1}}|\mathcal{P}]}\right] \leq \frac{|B_2(i)|}{p}.$$

Consider the Bernoulli random variable $\mathbb{1}[E_i^{\mathbf{1}}]$, and we have

$$\mathbf{Var}[\mathbb{1}[E_i^{\mathbf{1}}]] = \mathbb{P}[E_i^{\mathbf{1}}|\mathcal{P}]\cdot(1 - \mathbb{P}[E_i^{\mathbf{1}}|\mathcal{P}]) \leq \mathbb{P}[E_i^{\mathbf{1}}|\mathcal{P}].$$

Moreover, note that $\mathbb{E}[\mathbb{1}[E_i^{\mathbf{1}}]|\boldsymbol{C}] = \mathbb{P}[E_i^{\mathbf{1}}|\boldsymbol{C}]$, and consequently, due to the law of total variance, we have

$$\mathbf{Var}\boldsymbol{C} \sim \mathcal{P}[\mathbb{P}[E_i^{\mathbf{1}}|\boldsymbol{C}]] = \mathbf{Var}\boldsymbol{C} \sim \mathcal{P}[\mathbb{E}[\mathbb{1}[E_i^{\mathbf{1}}]|\boldsymbol{C}]] \leq \mathbf{Var}[\mathbb{1}[E_i^{\mathbf{1}}]] \leq \mathbb{P}[E_i^{\mathbf{1}}|\mathcal{P}],$$

and thus,

$$\mathbf{Var}\boldsymbol{C} \sim \mathcal{P}\left[\frac{\mathbb{P}[E_i^{\mathbf{1}}|\boldsymbol{C}] - \mathbb{P}[E_i^{\mathbf{1}}|\mathcal{P}]}{\mathbb{P}[E_i^{\mathbf{1}}|\mathcal{P}]}\right] \leq \frac{\mathbb{P}[E_i^{\mathbf{1}}|\mathcal{P}]}{\mathbb{P}[E_i^{\mathbf{1}}|\mathcal{P}]^2} = \frac{1}{\mathbb{P}[E_i^{\mathbf{1}}|\mathcal{P}]} \leq \frac{|B_2(i)|}{p}, \tag{A5}$$

where the second inequality is due to Theorem 4.2.

Now to prove the inequality in Theorem 4.5, we note that

$$\frac{\hat{\mathbb{P}}[E_i^{\mathbf{1}}|\mathcal{P}] - \mathbb{P}[E_i^{\mathbf{1}}|\mathcal{P}]}{\mathbb{P}[E_i^{\mathbf{1}}|\mathcal{P}]} = \frac{\left(\frac{1}{K}\sum_{k=1}^{K}\mathbb{P}[E_i^{\mathbf{1}}|\boldsymbol{c}^{(k)}]\right) - \mathbb{P}[E_i^{\mathbf{1}}|\mathcal{P}]}{\mathbb{P}[E_i^{\mathbf{1}}|\mathcal{P}]} = \frac{1}{K}\sum_{k=1}^{K}\left[\frac{\mathbb{P}[E_i^{\mathbf{1}}|\boldsymbol{c}^{(k)}] - \mathbb{P}[E_i^{\mathbf{1}}|\mathcal{P}]}{\mathbb{P}[E_i^{\mathbf{1}}|\mathcal{P}]}\right],$$

and thus,

$$\mathbf{Var}\left[\frac{\hat{\mathbb{P}}[E_i^{\mathbf{1}}|\mathcal{P}] - \mathbb{P}[E_i^{\mathbf{1}}|\mathcal{P}]}{\mathbb{P}[E_i^{\mathbf{1}}|\mathcal{P}]}\middle|\mathcal{P}\right] = \frac{1}{K}\mathbf{Var}\boldsymbol{C} \sim \mathcal{P}\left[\frac{\mathbb{P}[E_i^{\mathbf{1}}|\boldsymbol{C}] - \mathbb{P}[E_i^{\mathbf{1}}|\mathcal{P}]}{\mathbb{P}[E_i^{\mathbf{1}}|\mathcal{P}]}\right] \leq \frac{|B_2(i)|}{Kp}.$$

Now note that the expected value of $\frac{\hat{\mathbb{P}}[E_i^{\mathbf{1}} \mid \mathcal{P}] - \mathbb{P}[E_i^{\mathbf{1}} \mid \mathcal{P}]}{\mathbb{P}[E_i^{\mathbf{1}} \mid \mathcal{P}]}$ is zero due to the fact that $\hat{\mathbb{P}}[E_i^{\mathbf{1}}|\mathcal{P}]$ is an unbiased estimator of $\mathbb{P}[E_i^{\mathbf{1}}|\mathcal{P}]$. Thus, the squared error is the same as the variance. □

C.6 Proof of Theorem 4.6

We first present the following result on the joint exposure probability of a pair of nodes.

Lemma C.6. *For the 1-hop-max random clustering algorithm, if* $\mathbf{dist}(i, j) > 4$ *for a pair of nodes i and j, then for* $\boldsymbol{z} = \mathbf{1}$ *or* $\boldsymbol{z} = \mathbf{0}$ *and*

- *independent randomization, we have* $\mathbb{P}[E_i^{\boldsymbol{z}} \cap E_j^{\boldsymbol{z}}|\mathcal{P}] = \mathbb{P}[E_i^{\boldsymbol{z}}|\mathcal{P}]\cdot \mathbb{P}[E_j^{\boldsymbol{z}}|\mathcal{P}]$;
- *complete randomization, we have* $\mathbb{P}[E_i^{\boldsymbol{z}} \cap E_j^{\boldsymbol{z}}|\mathcal{P}] \leq \mathbb{P}[E_i^{\boldsymbol{z}}|\mathcal{P}]\cdot \mathbb{P}[E_j^{\boldsymbol{z}}|\mathcal{P}]$.

Proof. We first show that nodes i and j satisfying the aforementioned requirements can not be adjacent to the same cluster in any clustering generated from 1-hop-max. If otherwise, then there exists nodes $i' \in B_1(i)$ and $j' \in B_1(j)$ such that $C_{i'} = C_{j'}$, and a node k such that $i', j' \in B_1(k)$ with $C_{i'} = C_{j'} = X_k$ (recall that in the 1-hop-max algorithm, $X_1, \cdots, X_n$ are the $\mathcal{U}(0, 1)$ samples and also the signifiers of the clusters), we have $\mathbf{dist}(i, j) \leq 4$ due to the path $[i, i', k, j', j]$, contradictory to our assumption that $\mathbf{dist}(i, j) > 4$.

Now we prove the results in the lemma. Due to symmetry, it suffices to just prove for the case of $\boldsymbol{z} = \mathbf{1}$, and we first analyze the independent randomization scenario. For any clustering $\boldsymbol{C}$ generated from 1-hop-max, since i and j are not adjacent to a same cluster in $\boldsymbol{C}$, we have $\mathbb{P}[E_i^{\mathbf{1}} \cap E_j^{\mathbf{1}}|\boldsymbol{C}] = \mathbb{P}[E_i^{\mathbf{1}}|\boldsymbol{C}]\cdot \mathbb{P}[E_i^{\mathbf{1}}|\boldsymbol{C}]$. Moreover, since $\mathbf{dist}(i, j) > 4$, we have $B_2(i) \cap B_2(j) = \varnothing$, and according to Lemma 4.1, we have $\boldsymbol{C}_{B_1(i)}$ and $\boldsymbol{C}_{B_1(j)}$ independent. Combining both results, and note the fact that $\mathbb{P}[E_i^{\mathbf{1}}|\boldsymbol{C}] = \mathbb{P}[E_i^{\mathbf{1}}|\boldsymbol{C}_{B_1(i)}]$ and $\mathbb{P}[E_j^{\mathbf{1}}|\boldsymbol{C}] = \mathbb{P}[E_j^{\mathbf{1}}|\boldsymbol{C}_{B_1(j)}]$, we have

$$\begin{aligned}\mathbb{P}[E_i^{\mathbf{1}} \cap E_j^{\mathbf{1}}|\mathcal{P}] &= \mathbb{E}[\mathbb{P}[E_i^{\mathbf{1}} \cap E_j^{\mathbf{1}}|\boldsymbol{C}]]\\ &= \mathbb{E}[\mathbb{P}[E_i^{\mathbf{1}}|\boldsymbol{C}]\cdot \mathbb{P}[E_i^{\mathbf{1}}|\boldsymbol{C}]]\\ &= \mathbb{E}[\mathbb{P}[E_i^{\mathbf{1}}|\boldsymbol{C}_{B_1(i)}]\cdot \mathbb{P}[E_i^{\mathbf{1}}|\boldsymbol{C}_{B_1(j)}]]\\ &= \mathbb{E}[\mathbb{P}[E_i^{\mathbf{1}}|\boldsymbol{C}_{B_1(i)}]]\cdot \mathbb{E}[\mathbb{P}[E_i^{\mathbf{1}}|\boldsymbol{C}_{B_1(j)}]]\\ &= \mathbb{P}[E_i^{\mathbf{1}}|\mathcal{P}]\cdot \mathbb{P}[E_j^{\mathbf{1}}|\mathcal{P}],\end{aligned}$$

where the last but one equality is due to $\boldsymbol{C}_{B_1(i)}$ and $\boldsymbol{C}_{B_1(j)}$ being independent.

We now prove the result for the complete randomization scenario. The proof is almost identical to that of independent randomization, except for the fact that $\mathbb{P}[E_i^{\mathbf{1}} \cap E_j^{\mathbf{1}}|\boldsymbol{C}] \le \mathbb{P}[E_i^{\mathbf{1}}|\boldsymbol{C}]\cdot \mathbb{P}[E_i^{\mathbf{1}}|\boldsymbol{C}]$ which is an equality in the independent randomization. Here, the joint probability might be smaller due to the scenario if one of the clusters adjacent to node i is paired with one adjacent to node j, and thus, $\mathbb{P}[E_i^{\mathbf{1}} \cap E_j^{\mathbf{1}}|\boldsymbol{C}] = 0$. Formally, we have

$$\mathbb{P}[E_i^{\mathbf{1}} \cap E_j^{\mathbf{1}}|\mathcal{P}] = \mathbb{E}[\mathbb{P}[E_i^{\mathbf{1}} \cap E_j^{\mathbf{1}}|\boldsymbol{C}]] \le \mathbb{E}[\mathbb{P}[E_i^{\mathbf{1}}|\boldsymbol{C}]\cdot \mathbb{P}[E_i^{\mathbf{1}}|\boldsymbol{C}]] = \mathbb{P}[E_i^{\mathbf{1}}|\mathcal{P}]\cdot \mathbb{P}[E_j^{\mathbf{1}}|\mathcal{P}],$$

where the last equality is verified in the proof for independent randomization. □

With this lemma in hand, we now prove Theorem 4.6.

Proof (Theorem 4.6). According to Lemma C.6, for any pair of nodes i and j such that $j \notin B_4(i)$, we have $\frac{\mathbb{P}[E_i^{\mathbf{1}} \cap E_j^{\mathbf{1}}]}{\mathbb{P}[E_i^{\mathbf{1}}]\mathbb{P}[E_j^{\mathbf{1}}]} - 1 \le 0$ for both independent and complete randomization. Now the variance of the mean outcome estimator, as formulated in equation (4), satisfies

$$\begin{aligned}\mathbf{Var}[\hat{\mu}_{\mathcal{P}}(\mathbf{1})] &= \frac{1}{n^2}\sum_{i=1}^{n}\left[\left(\frac{1}{\mathbb{P}[E_i^{\mathbf{1}}|\mathcal{P}]} - 1\right)Y_i(\mathbf{1})^2 + \sum_{j\neq i}\left(\frac{\mathbb{P}[E_i^{\mathbf{1}} \cap E_j^{\mathbf{1}}|\mathcal{P}]}{\mathbb{P}[E_i^{\mathbf{1}}|\mathcal{P}]\mathbb{P}[E_j^{\mathbf{1}}|\mathcal{P}]} - 1\right)Y_i(\mathbf{1})Y_j(\mathbf{1})\right]\\ &\le \frac{1}{n^2}\sum_{i=1}^{n}\left[\left(\frac{1}{\mathbb{P}[E_i^{\mathbf{1}}|\mathcal{P}]} - 1\right)Y_i(\mathbf{1})^2 + \sum_{j\in B_4(i), j\neq i}\left(\frac{\mathbb{P}[E_i^{\mathbf{1}} \cap E_j^{\mathbf{1}}|\mathcal{P}]}{\mathbb{P}[E_i^{\mathbf{1}}|\mathcal{P}]\mathbb{P}[E_j^{\mathbf{1}}|\mathcal{P}]} - 1\right)Y_i(\mathbf{1})Y_j(\mathbf{1})\right]\\ &\le \frac{1}{n^2}\sum_{i=1}^{n}\left[\left(\frac{1}{\mathbb{P}[E_i^{\mathbf{1}}|\mathcal{P}]} - 1\right)Y_i(\mathbf{1})^2 + \sum_{j\in B_4(i), j\neq i}\left(\frac{1}{\mathbb{P}[E_i^{\mathbf{1}}|\mathcal{P}]} - 1\right)Y_i(\mathbf{1})Y_j(\mathbf{1})\right]\\ &\le \frac{\bar{Y}^2}{n^2}\sum_{i=1}^{n}\left[\left(\frac{1}{\mathbb{P}[E_i^{\mathbf{1}}|\mathcal{P}]} - 1\right)\cdot|B_4(i)|\right]\\ &\le \frac{\bar{Y}^2}{n^2}\sum_{i=1}^{n}\frac{|B_4(i)|}{\mathbb{P}[E_i^{\mathbf{1}}|\mathcal{P}]},\end{aligned}$$

where the second inequality is due to $\mathbb{P}[E_i^{\mathbf{1}} \cap E_j^{\mathbf{1}}|\mathcal{P}] \le \mathbb{P}[E_j^{\mathbf{1}}|\mathcal{P}]$. □

C.7 Proof of Theorem 4.8

Proof. Analogous to Theorem 4.7, it can be verified that $\mathbf{Var}[\hat{\mu}_{\mathcal{P}}(\mathbf{0})] \leq \frac{1}{n} \cdot \bar{Y}^2(d_{\max} + 1)^2 \kappa^4 (1 - p)^{-1}$. Now according to the variance formula in equation (5), we have

$$\begin{aligned}\mathbf{Var}[\hat{\tau}_{\mathcal{P}}] &= \mathbf{Var}[\hat{\mu}_{\mathcal{P}}(\mathbf{1})] + \mathbf{Var}[\hat{\mu}_{\mathcal{P}}(\mathbf{0})] - 2 \cdot \mathbf{Cov}[\hat{\mu}_{\mathcal{P}}(\mathbf{1}), \hat{\mu}_{\mathcal{P}}(\mathbf{0})] \\ &\leq \mathbf{Var}[\hat{\mu}_{\mathcal{P}}(\mathbf{1})] + \mathbf{Var}[\hat{\mu}_{\mathcal{P}}(\mathbf{0})] + 2 \cdot \sqrt{\mathbf{Var}[\hat{\mu}_{\mathcal{P}}(\mathbf{1})] \cdot \mathbf{Var}[\hat{\mu}_{\mathcal{P}}(\mathbf{0})]} \\ &\leq \mathbf{Var}[\hat{\mu}_{\mathcal{P}}(\mathbf{1})] + \mathbf{Var}[\hat{\mu}_{\mathcal{P}}(\mathbf{0})] + \mathbf{Var}[\hat{\mu}_{\mathcal{P}}(\mathbf{1})] + \mathbf{Var}[\hat{\mu}_{\mathcal{P}}(\mathbf{0})] \\ &\leq \frac{2}{n} \cdot \bar{Y}^2(d_{\max} + 1)^2 \kappa^4 (p^{-1} + (1 - p)^{-1}).\end{aligned}$$

where the second inequality is due to the mean inequality, and the last inequality is due to the variance upper bound of $\hat{\mu}_{\mathcal{P}}(\mathbf{1})$ and $\hat{\mu}_{\mathcal{P}}(\mathbf{0})$, respectively. □

C.8 Proof of Theorem 4.10

Proof. With a constant treatment effect τ, we have

$$\tilde{\mu}(\mathbf{1}) = \frac{\sum_{i=1}^n \mathbb{1}[E_i^{\mathbf{1}}] / \mathbb{P}[E_i^{\mathbf{1}}] \cdot (Y_i(\mathbf{0}) + \tau)}{\sum_{i=1}^n \mathbb{1}[E_i^{\mathbf{1}}] / \mathbb{P}[E_i^{\mathbf{1}}]} = \tau + \frac{\sum_{i=1}^n \mathbb{1}[E_i^{\mathbf{1}}] / \mathbb{P}[E_i^{\mathbf{1}}] \cdot Y_i(\mathbf{0})}{\sum_{i=1}^n \mathbb{1}[E_i^{\mathbf{1}}] / \mathbb{P}[E_i^{\mathbf{1}}]},$$

and consequently,

$$\mathbb{E}[\tilde{\mu}(\mathbf{1})] - \tau = \mathbb{E}\left[\frac{\sum_{i=1}^n \mathbb{1}[E_i^{\mathbf{1}}] / \mathbb{P}[E_i^{\mathbf{1}}] \cdot Y_i(\mathbf{0})}{\sum_{i=1}^n \mathbb{1}[E_i^{\mathbf{1}}] / \mathbb{P}[E_i^{\mathbf{1}}]}\right] = \mathbb{E}\left[\frac{\sum_{i=1}^n \mathbb{1}[E_i^{\mathbf{0}}] / \mathbb{P}[E_i^{\mathbf{0}}] \cdot Y_i(\mathbf{0})}{\sum_{i=1}^n \mathbb{1}[E_i^{\mathbf{0}}] / \mathbb{P}[E_i^{\mathbf{0}}]}\right] = \mathbb{E}[\tilde{\mu}(\mathbf{0})],$$

where the second equality is due to the symmetry of network exposure to treatment and control, more specifically, the joint distribution of $\{\mathbb{1}[E_i^{\mathbf{1}}]\}_{i=1}^n$ is the same as that of $\{\mathbb{1}[E_i^{\mathbf{0}}]\}_{i=1}^n$.

Since $\mathbb{E}[\tilde{\mu}(\mathbf{1})] = \mathbb{E}[\tilde{\mu}(\mathbf{0})] + \tau$, then $\mathbb{E}[\tilde{\tau}] = \mathbb{E}[\tilde{\mu}(\mathbf{1})] - \mathbb{E}[\tilde{\mu}(\mathbf{0})] = \tau$. □

C.9 Proof of Theorem 5.1

Proof. Note that in any oracle k-partition of a cycle, each cluster contains two nodes on the boundary. Therefore, for any node i, the probability of being on the boundary of a random such partition is $2k/n = o(1)$, and thus, as $n \to \infty$, it is almost surely between nodes in the same cluster and

$$\mathbb{P}[E_i^{\mathbf{1}} | \mathcal{P}] = \mathbb{P}[E_i^{\mathbf{0}} | \mathcal{P}] \to \frac{1}{2}. \tag{A6}$$

For any node pair i and j, define their *angle distance* as

$$\delta(\alpha_i, \alpha_j) \triangleq \min\{|\alpha_i - \alpha_j|, 2\pi - |\alpha_i - \alpha_j|\},$$

a quantity in $[0, \pi]$ which is zero if and only if $i = j$. Note that if $C_i = C_j$, i.e., if i and j belongs to a same cluster in a oracle k-partition, we must have $\delta(\alpha_i, \alpha_j) < \frac{2\pi}{k}$. On the contrary, if $\delta(\alpha_i, \alpha_j) < \frac{2\pi}{k}$, the probability of them belonging to a same cluster in a random oracle k-partition is $1 - \delta(\alpha_i, \alpha_j)/\frac{2\pi}{k}$. Therefore, for any node pair i and j, we have

$$\mathbb{P}[C_i = C_j | \mathcal{P}] = \mathbb{1}[\delta(\alpha_i, \alpha_j) < 2\pi/k] \cdot \left[1 - \frac{k\delta(\alpha_i, \alpha_j)}{2\pi}\right]. \tag{A7}$$

□

C.9.1 Variance under independent randomization

We start with computing the joint exposure probabilities $\mathbb{P}[E_i^1 \cap E_j^1|\mathcal{P}]$ and $\mathbb{P}[E_i^1 \cap E_j^0|\mathcal{P}]$. Note that in the limit of $n \to \infty$, the probability of either node being on a boundary vanishes, and thus, we have

$$\mathbb{P}[E_i^1 \cap E_j^1|\boldsymbol{C}] \to \begin{cases} 1/2 & \text{if } C_i = C_j \\ 1/4 & \text{if } C_i \neq C_j \end{cases}, \quad \mathbb{P}[E_i^1 \cap E_j^0|\boldsymbol{C}] \to \begin{cases} 0 & \text{if } C_i = C_j \\ 1/4 & \text{if } C_i \neq C_j \end{cases}.$$

By combining with equation (A7), we have

$$\mathbb{P}[E_i^1 \cap E_j^1|\mathcal{P}] \to \frac{1}{2} \cdot \mathbb{P}[C_i = C_j|\mathcal{P}] + \frac{1}{4} \cdot \mathbb{P}[C_i \neq C_j|\mathcal{P}] = \begin{cases} \frac{1}{2} - \frac{k\delta(\alpha_i, \alpha_j)}{8\pi} & \text{if } \delta(\alpha_i, \alpha_j) < \frac{2\pi}{k} \\ \frac{1}{4} & \text{if } \delta(\alpha_i, \alpha_j) \geq \frac{2\pi}{k}, \end{cases} \tag{A8}$$

and

$$\mathbb{P}[E_i^1 \cap E_j^0|\mathcal{P}] \to \frac{1}{4} \cdot \mathbb{P}[C_i \neq C_j|\mathcal{P}] = \begin{cases} \frac{k\delta(\alpha_i, \alpha_j)}{8\pi} & \text{if } \delta(\alpha_i, \alpha_j) < \frac{2\pi}{k} \\ \frac{1}{4} & \text{if } \delta(\alpha_i, \alpha_j) \geq \frac{2\pi}{k}. \end{cases} \tag{A9}$$

Now we compute the variance of the mean outcome HT estimators given these probabilities. Note that the variance, as given in equation (4), is equivalent to

$$\mathbf{Var}[\hat{\mu}(\boldsymbol{z})] = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left(\frac{\mathbb{P}[E_i^{\boldsymbol{z}} \cap E_j^{\boldsymbol{z}}]}{\mathbb{P}[E_i^{\boldsymbol{z}}]\mathbb{P}[E_j^{\boldsymbol{z}}]} - 1 \right) Y_i(\boldsymbol{z}) Y_j(\boldsymbol{z})$$

due to the fact that $\mathbb{P}[E_i^{\boldsymbol{z}} \cap E_i^{\boldsymbol{z}}] = \mathbb{P}[E_i^{\boldsymbol{z}}]$. Therefore,

$$\begin{aligned} \mathbf{Var}[\hat{\mu}(\mathbf{1})] &= \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left(\frac{\mathbb{P}[E_i^1 \cap E_j^1]}{\mathbb{P}[E_i^1]\mathbb{P}[E_j^1]} - 1 \right) Y_i(E_i^1) Y_j(E_j^1) \\ &= \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j:\delta(\alpha_i,\alpha_j)<\frac{2\pi}{k}} \left(1 - \frac{k\delta(\alpha_i, \alpha_j)}{2\pi} + o(1) \right) (a + b\sin\alpha_i + \tau)(a + b\sin\alpha_j + \tau) \\ &\quad + \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j:\delta(\alpha_i,\alpha_j)\geq\frac{2\pi}{k}} \left(4 \cdot \frac{1}{4} - 1 + o(1) \right) (a + b\sin\alpha_i + \tau)(a + b\sin\alpha_j + \tau) \\ &= \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j:\delta(\alpha_i,\alpha_j)<\frac{2\pi}{k}} \left(1 - \frac{k\delta(\alpha_i, \alpha_j)}{2\pi} \right) (a + b\sin\alpha_i + \tau)(a + b\sin\alpha_j + \tau) \\ &\quad + \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j:\delta(\alpha_i,\alpha_j)\geq\frac{2\pi}{k}} \left(4 \cdot \frac{1}{4} - 1 \right) (a + b\sin\alpha_i + \tau)(a + b\sin\alpha_j + \tau) + o(1), \end{aligned}$$

where the third equality is due to the average of $n^2 o(1)$ terms being $o(1)$. We then take a limit corresponding to Riemann integration and obtain:

$$\begin{aligned}
\mathbf{Var}[\hat{\mu}(\mathbf{1})] &\to \frac{1}{4\pi^2}\int_0^{2\pi}(a+b\sin\alpha_i+\tau)\cdot\left(\int_{\alpha_i-\frac{2\pi}{k}}^{\alpha_i+\frac{2\pi}{k}}\left(1-\frac{k|\alpha_i-\alpha_j|}{2\pi}\right)(a+b\sin\alpha_j+\tau)\mathrm{d}\alpha_j\right)\mathrm{d}\alpha_i\\
&= \frac{1}{4\pi^2}\int_0^{2\pi}(a+b\sin\alpha_i+\tau)\cdot\left(\int_0^{\frac{2\pi}{k}}\left(1-\frac{k\delta}{2\pi}\right)(2a+2\tau+b\sin(\alpha_i+\delta)+b\sin(\alpha_i-\delta))\mathrm{d}\delta\right)\mathrm{d}\alpha_i\\
&= \frac{1}{4\pi^2}\int_0^{2\pi}(a+b\sin\alpha_i+\tau)\cdot\left(\int_0^{\frac{2\pi}{k}}\left(1-\frac{k\delta}{2\pi}\right)(2a+2\tau+2b\sin\alpha_i\cos\delta)\mathrm{d}\delta\right)\mathrm{d}\alpha_i\\
&= \frac{1}{4\pi^2}\int_0^{2\pi}(a+\tau+b\sin\alpha_i)\cdot\left((a+\tau)\cdot\frac{2\pi}{k}+\frac{bk(1-\cos(2\pi/k))}{\pi}\sin\alpha_i\right)\mathrm{d}\alpha_i\\
&= \frac{(a+\tau)^2}{k}+\frac{b^2k(1-\cos(2\pi/k))}{4\pi^2}.
\end{aligned}$$

Analogously, it can be shown that

$$\begin{aligned}
\mathbf{Var}[\hat{\mu}(\mathbf{0})] &\to \frac{a^2}{k}+\frac{b^2k(1-\cos(2\pi/k))}{4\pi^2},\\
\mathbf{Cov}[\hat{\mu}(\mathbf{1}),\hat{\mu}(\mathbf{0})] &\to \frac{a(a+\tau)}{k}+\frac{b^2k(1-\cos(2\pi/k))}{4\pi^2}.
\end{aligned}$$

Consequently,

$$\mathbf{Var}[\hat{\tau}] = \mathbf{Var}[\hat{\mu}(\mathbf{1})]+\mathbf{Var}[\hat{\mu}(\mathbf{0})]-2\mathbf{Cov}[\hat{\mu}(\mathbf{1}),\hat{\mu}(\mathbf{0})] \to \frac{(2a+\tau)^2}{k}+\frac{b^2k}{\pi^2}(1-\cos(2\pi/k)).$$

C.9.2 Variance under complete randomization

We also first compute the joint exposure probabilities $\mathbb{P}[E_i^{\mathbf{1}}\cap E_j^{\mathbf{1}}]$. If $C_i=C_j$ we have $\mathbb{P}[E_i^{\mathbf{1}}\cap E_j^{\mathbf{1}}|C]=1/2$ and $\mathbb{P}[E_i^{\mathbf{1}}\cap E_j^{\mathbf{0}}|C]=0$. When $C_i\neq C_j$, i.e., nodes i and j belongs to different clusters, there are two scenarios: the two clusters are assigned together and oppositely into treatment and control, or the two clusters are assigned independently. Under the first scenario, which happens with probability $\frac{1}{k-1}$ conditional on $C_i\neq C_j$, $E_i^{\mathbf{1}}\cap E_j^{\mathbf{1}}$ is not possible; under the second scenario which happens with probability $\frac{k-2}{k-1}$ conditional on $C_i\neq C_j$, $E_i^{\mathbf{1}}\cap E_j^{\mathbf{1}}$ happens when both clusters are assigned into treatment and thus the conditional probability is 1/4 as $n\to\infty$. Therefore, we have

$$\mathbb{P}[E_i^{\mathbf{1}}\cap E_j^{\mathbf{1}}|C_i\neq C_j]\to\frac{1}{k-1}\cdot 0+\frac{k-2}{k-1}\cdot\frac{1}{4},$$

and thus,

$$\mathbb{P}[E_i^{\mathbf{1}}\cap E_j^{\mathbf{1}}]\to\frac{1}{2}\cdot\mathbb{P}[C_i=C_j]+\frac{k-2}{4(k-1)}\cdot\mathbb{P}[C_i\neq C_j]=\begin{cases}\dfrac{1}{2}-\dfrac{k\delta(\alpha_i,\alpha_j)}{8\pi}\cdot\dfrac{k}{k-1} & \text{if } \delta(\alpha_i,\alpha_j)<\dfrac{2\pi}{k}\\[2ex] \dfrac{1}{4}\cdot\dfrac{k-2}{k-1} & \text{if } \delta(\alpha_i,\alpha_j)\geq\dfrac{2\pi}{k}.\end{cases}$$

Similarly, one can show that

$$\mathbb{P}[E_i^{\mathbf{1}} \cap E_j^{\mathbf{0}}] \to \begin{cases} \dfrac{k\delta(\alpha_i, \alpha_j)}{8\pi} \cdot \dfrac{k}{k-1} & \text{if } \delta(\alpha_i, \alpha_j) < \dfrac{2\pi}{k} \\ \dfrac{1}{4} \cdot \dfrac{k}{k-1} & \text{if } \delta(\alpha_i, \alpha_j) \geq \dfrac{2\pi}{k}. \end{cases}$$

With these exposure probabilities, we have

$$\begin{aligned}
\mathbf{V\ ar}[\hat{\mu}(\mathbf{1})] &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\frac{\mathbb{P}[E_i^{\mathbf{1}} \cap E_j^{\mathbf{1}}]}{\mathbb{P}[E_i^{\mathbf{1}}]\mathbb{P}[E_j^{\mathbf{1}}]} - 1\right)Y_i(E_i^{\mathbf{1}})Y_j(E_j^{\mathbf{1}}) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\left(\sum_{j:\delta(\alpha_i,\alpha_j)<\frac{2\pi}{k}}\left(1 - \frac{k\delta(\alpha_i, \alpha_j)}{2\pi} \cdot \frac{k}{k-1}\right)(a + b\sin\alpha_i + \tau)(a + b\sin\alpha_j + \tau)\right. \\
&\quad \left. + \sum_{j:\delta(\alpha_i,\alpha_j)\geq\frac{2\pi}{k}}\left(\frac{k-2}{k-1} - 1\right)(a + b\sin\alpha_i + \tau)(a + b\sin\alpha_j + \tau)\right) + o(1) \\
&\to \frac{1}{4\pi^2}\int_0^{2\pi}(a + b\sin\alpha_i + \tau)\cdot\left[\int_0^{\frac{2\pi}{k}}\left(1 - \frac{k\delta}{2\pi}\cdot\frac{k}{k-1}\right)(2a + 2\tau + b\sin(\alpha_i + \delta) + b\sin(\alpha_i - \delta))\mathrm{d}\delta\right. \\
&\quad \left. + \int_{\frac{2\pi}{k}}^{\pi}\left(-\frac{1}{k-1}\right)(2a + 2\tau + b\sin(\alpha_i + \delta) + b\sin(\alpha_i - \delta))\mathrm{d}\delta\right]\mathrm{d}\alpha_i \\
&= \frac{1}{4\pi^2}\int_0^{2\pi}(a + b\sin\alpha_i + \tau)\cdot\left[\int_0^{\frac{2\pi}{k}}\left(1 - \frac{k\delta}{2\pi}\cdot\frac{k}{k-1}\right)(2a + 2\tau + 2b\sin\alpha_i\cos\delta)\mathrm{d}\delta\right. \\
&\quad \left. + \int_{\frac{2\pi}{k}}^{\pi}\left(-\frac{1}{k-1}\right)(2a + 2\tau + 2b\sin\alpha_i\cos\delta)\mathrm{d}\delta\right]\mathrm{d}\alpha_i \\
&= \frac{1}{4\pi^2}\int_0^{2\pi}(a + \tau + b\sin\alpha_i)\cdot\left[(a + \tau)\frac{2\pi(k-2)}{k(k-1)} - \frac{2b}{k-1}\sin(2\pi/k)\sin\alpha_i\right. \\
&\quad + \frac{bk(1 - \cos(2\pi/k))}{\pi}\cdot\frac{k}{k-1}\sin\alpha_i \\
&\quad \left. - (a + \tau)\frac{2\pi(k-2)}{k(k-1)} + \frac{2b}{k-1}\sin(2\pi/k)\sin\alpha_i\right]\mathrm{d}\alpha_i \\
&= \frac{1}{4\pi^2}\int_0^{2\pi}(a + \tau + b\sin\alpha_i)\cdot\frac{bk(1 - \cos(2\pi/k))}{\pi}\cdot\frac{k}{k-1}\sin\alpha_i\mathrm{d}\alpha_i \\
&= \frac{b^2k^2(1 - \cos(2\pi/k))}{4\pi^2(k-1)}.
\end{aligned}$$

Analogously, it can be shown that

$$\begin{aligned}
\mathbf{Var}[\hat{\mu}(\mathbf{0})] &\to \frac{b^2k^2(1 - \cos(2\pi/k))}{4\pi^2(k-1)}, \\
\mathbf{Cov}[\hat{\mu}(\mathbf{1}), \hat{\mu}(\mathbf{0})] &\to \frac{b^2k^2(1 - \cos(2\pi/k))}{2\pi^2(k-1)}.
\end{aligned}$$

Consequently,

$$\mathbf{Var}[\hat{\tau}] = \mathbf{Var}[\hat{\mu}(\mathbf{1})] + \mathbf{Var}[\hat{\mu}(\mathbf{0})] - 2\mathbf{Cov}[\hat{\mu}(\mathbf{1}), \hat{\mu}(\mathbf{0})] \to \frac{b^2k^2(1-\cos(2\pi/k))}{\pi^2(k-1)}.$$

C.9.3 Variance with increasing number of clusters

In the end, we show the variance when the number of clusters k increases. Due the fact that $1-\cos(x) \sim x^2/2$ as $x \to 0$, under the independent randomization scheme, we have

$$\mathbf{Var}[\hat{\tau}] \to \frac{(2a+\tau)^2}{k} + \frac{b^2k}{\pi^2}(1-\cos(2\pi/k)) \sim \frac{(2a+\tau)^2}{k} + \frac{b^2k}{\pi^2}\cdot\frac{2\pi^2}{k^2} = [(2a+\tau)^2 + 2b^2]\cdot\Theta(1/k).$$

Under the complete randomization scheme, we have

$$\mathbf{Var}[\hat{\tau}] \to \frac{b^2k^2}{\pi^2(k-1)}(1-\cos(2\pi/k)) \sim \frac{b^2k^2}{\pi^2(k-1)}\cdot\frac{2\pi^2}{k^2} = 2b^2\cdot\Theta(1/k).$$

C.10 Proof of Theorem B.1

Proof. For result (a), Since the PDF of $\beta(w,1)$ distribution is $f(x) = wx^{w-1}$, we have

$$\mathbb{P}[X_j < X_i] = \int_0^1 f(x_i)\mathbb{P}[X_j \le x_i]\mathrm{d}x_i = \int_0^1 w_i x^{w_i-1}\cdot x_i^{w_j}\mathrm{d}x_i = \frac{w_i}{w_i+w_j}.$$

For result (b), we have

$$\mathbb{P}[\max\{X_i, X_j\} \le x] = \mathbb{P}[X_i \le x]\mathbb{P}[X_j \le x] = x^{w_i+w_j},$$

which is the CDF of the $\beta(w_i + w_j, 1)$ distribution. □

C.11 Proof of Theorem B.2

We first prove the following useful lemma.

Lemma C.7. *If every node generates $X_i \sim \beta(w_i, 1)$ independent, then we have $\mathbb{P}[X_i \ge X_j,\ \forall j \in B_2(i)] = w_i/\left(\sum_{j\in B_2(i)} w_j\right)$.*

Proof. According to the second result in Theorem B.1, the distribution of $\bar{X} \triangleq \max([X_j]_{j\in B_2(i), j\ne i})$ is $\beta(S_i, 1)$ where $S_i = \sum_{j\in B_2(i), j\ne i} w_j$. Moreover, we note that $\bar{X}$ and X_i are independent, and thus, $\mathbb{P}[X_i \ge \bar{X}] = w_i/(S_i + w_i)$ according to the first result in Theorem B.1. This result is equivalent to the lemma statement. □

Proof (Theorem B.2). Analogous to the proof of Theorem 4.2, one only need to show that the probability of node i being in the interior of a cluster is lower bounded by $w_i/\left(\sum_{j\in B_2(i)} w_j\right)$.

For 1-hop-max clustering, similarly a sufficient condition for node i being in the interior of a clustering is if node i generates the largest number in $B_2(i)$, i.e., $X_i \ge X_j$ for any $j \in B_2(i)$. By Lemma C.7, we have its probability being $w_i / \left(\sum_{j\in B_2(i)} w_j\right)$, and which is a lower bound on the probability of node i being in the interior of a cluster.

For 3-net clustering, again we consider the scenario when node i is ranked first among nodes in $B_2(i)$, a sufficient condition of node i being a seed node and thus in the interior of a cluster. According to the procedure in Algorithm 4, this is equivalent to node i generating the largest number in $B_2(i)$. According to Lemma C.7, that probability is $w_i / \left(\sum_{j\in B_2(i)} w_j\right)$. □

C.12 Proof of Theorem B.3

The proof is almost the same as that of Theorem 4.5, except that in equation (A5), we apply the lower bound result of Theorem B.2.

C.13 Proof of Theorem B.7

Proof. Note that $j \in B_2(i)$ is equivalent as $\mathbf{dist}(i, j) \leq 2$, and thus,

$$\begin{aligned}\bar{H}(\boldsymbol{w}) &= \frac{1}{p}\sum_{i=1}^{n}\frac{\sum_{j\in B_2(i)} w_j}{w_i} = \frac{1}{p}\sum_{(i,j):\mathbf{dist}(i,j)\leq 2}\left(\frac{w_j}{w_i}+\frac{w_i}{w_j}\right)\\ &\geq \frac{1}{p}\sum_{(i,j):\mathbf{dist}(i,j)\leq 2} 2\cdot\sqrt{\frac{w_j}{w_i}\cdot\frac{w_i}{w_j}} = \frac{1}{p}\sum_i |B_2(i)|,\end{aligned}$$

where the inequality applies the AM–GM inequality, which holds as equality if all nodes have the same weight. □