Adaptive A/B Test on Networks with Cluster Structures

Yang Liu GWU
Yifan Zhou Prizer Inc.
Ping Li Baidu Research
Feifang Hu GWU

Abstract

Units in online A/B tests are often involved in social networks. Thus, their outcomes may depend on the treatment of their neighbors. Many of such networks exhibit certain cluster structures allowing the use of these features in the design to reduce the bias from network interference. When the average treatment effect (ATE) is considered from the individual perspective, conditions for the valid estimation restrict the use of these features in the design. We show that such restrictions can be alleviated if the ATE from the cluster perspective is considered. Using an illustrative example, we further show that the weights employed by the Horvitz-Thompson estimator may not appropriately accommodate the network structure, and purely relying on graph-cluster randomization may generate very unbalanced cluster-treated structures across the treatment arms. The measures of such structures for one cluster may depend on the treatment of other clusters and pose a great challenge for the design of A/B tests. To address these issues, we propose a rerandomized-adaptive randomization to balance the clusters and a cluster-adjusted estimator to alleviate the problem of the weights. Numerical studies are conducted to demonstrate the usage of the proposed procedure.

1 Introduction

A/B test, also called the randomized controlled study, has been widely used by large IT firms to compare a new version of a product to its older counterpart (Kohavi et al., 2013). In both online and offline A/B tests, the units may interact with each other. Thus, the outcome of one unit may depends on the treatment assignments of the others. This phenomenon often occurs when the units are involved in a network. For example, in the applications of social networks, the treatment may affect not only the behavior of the treated users, but also the behavior of their friends. This kind of network interference often complicates the evaluation of the average treatment effect (ATE) (Gui et al., 2015).

Many networks involved in A/B tests have certain cluster structures (Gui et al., 2015; Holtz et al., 2020). Recent works showed that using graph-cluster randomization, i.e., randomization at cluster level, can reduce the bias for estimation (Ugander et al., 2013; Eckles et al., 2016; Ugander and Yin, 2020). To ensure the bounded variance of the estimator for the ATE, Ugander et al. (2013) suggested that the cluster sizes have to be bounded. However, the clusters generated by community detection algorithms or those observed in practice may not satisfy such conditions, because the networks found in real applications often have both large and small clusters (Leskovec et al., 2008). Thus, the use of such clusters in graph-cluster randomization may be limited. However, the clusers found by the community detection algorithms may have useful interpretations and approximate the ground truth network structures (Clauset et al., 2004). If such clusters can be appropriately used in the design of A/B test, the evaluated effect may have practical meanings.

One possible reason for the restricted use of the cluster structures in randomization is that the estimator usually considers the ATE from the individual perspective but not the cluster perspective. To demonstrate this difference, consider comparing two treatments. Suppose we observe a network G with K clusters and $N = \sum_{k=1}^{K} n_k$ nodes, where n_k is the size of cluster k. Let $Z_{k,i} \in \{0, 1\}$ be the treatment assignment of node (unit) i of cluster k for $1 \le k \le K$. Let $Z_k = (Z_{k,1}, \cdots, Z_{k,n_k})'$ and $\boldsymbol{Z} = (Z_1', \cdots, Z_K')'$ represent the n_k-treatment assignments of cluster k and N-treatment assignment of the N units. Using the Neyman-Rubin's language (Rubin, 1974; Splawa-Neyman et al., 1990), let $Y_{k,i}(\boldsymbol{z})$ represent the potential outcome of node i of cluster k given $\boldsymbol{Z} = \boldsymbol{z}$. In many scenarios of A/B tests, the goal of the experiment is to

Proceedings of the 25[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

decide whether all of the units should be assigned with the new treatment. The estimand of interest is thus the "all versus nothing" ATE that compares $Y_{k,i}(\mathbf{1})$ with $Y_{k,i}(\mathbf{0})$, where $\mathbf{1}$ and $\mathbf{0}$ are N-vectors of ones and zeros, respectively. The ATE from the individual and cluster perspectives are

$$\tau_{IN}(\mathbf{1},\mathbf{0}) = N^{-1}\sum_{k=1}^{n}\sum_{i=1}^{n_k}\{Y_{k,i}(\mathbf{1}) - Y_{k,i}(\mathbf{0})\},$$

$$\tau_{CL}(\mathbf{1},\mathbf{0}) = K^{-1}\sum_{k=1}^{K} n_k^{-1}\sum_{i=1}^{n_k}\{Y_{k,i}(\mathbf{1}) - Y_{k,i}(\mathbf{0})\},$$

respectively. The difference between the two estimands is that $\tau_{CL}(\mathbf{1},\mathbf{0})$ puts equal weights on the clusters, whereas $\tau_{IN}(\mathbf{1},\mathbf{0})$ puts more weights on the large cluster. In some applications, $\tau_{CL}(\mathbf{1},\mathbf{0})$ can be meaningful because the cluster has certain interpretations. For instance, if the cluster represents the school, household, etc, then $\tau_{CL}(\mathbf{1},\mathbf{0})$ represents the ATE on these higher levels. Moreover, when the units within the same cluster are more frequently interfered, $\tau_{IN}(\mathbf{1},\mathbf{0})$ may be more sensitive to the peer effect evaluated from the large clusters than $\tau_{CL}(\mathbf{1},\mathbf{0})$. Therefore, $\tau_{CL}(\mathbf{1},\mathbf{0})$ can also be useful in describing the treatment effect.

The estimation of $\tau_{IN}(\mathbf{1},\mathbf{0})$ with graph-cluster randomization is investigated by Ugander et al. (2013) and Ugander and Yin (2020), but the estimation for $\tau_{CL}(\mathbf{1},\mathbf{0})$ with graph-cluster randomization has been scarcely considered. To compare the conditions under which the ATEs can be appropriately estimated, we construct the Horvitz-Thompson estimator (HTE) for the two ATEs by using the neighborhood interference assumption (Ugander et al., 2013; Eckles et al., 2016; Forastiere et al., 2021). We show that the order of the variance of the HTE for $\tau_{CL}(\mathbf{1},\mathbf{0})$ does not depend on the cluster sizes. Therefore, the restriction on the cluster sizes can be alleviated for estimating $\tau_{CL}(\mathbf{1},\mathbf{0})$.

Besides the choice of the estimands, other problems related to the design and the estimation may also restrict the use of graph-cluster randomization. A network with four clusters is used to demonstrate these issues. This example shows that balance needs to be achieved across treatment arms of clusters with respect to the cluster-treated structures. Unlike the problem of covariate balance without network interference, the value of such structures for one cluster may be affected by the treatment assignments of its connected clusters. Therefore, these measures are more difficult to balance by their nature. Furthermore, the HTEs are often quite sensitive to their weights (Aronow et al., 2017; Ugander and Yin, 2020). For instance, the HTEs can have large variance when several nodes have very small probabilities to be included in the estimation. If these problems are not considered in the A/B test, the evaluated results may not be reliable.

In this article, we propose an adaptive A/B test procedure consisting of a rerandomized-adaptive randomization (ReAR) and a cluster-adjusted estimator (CAE) to tackle these challenges. The ReAR uses the pairwise-sequential randomization (Qin et al., 2016) to balance the cluster-treated structures of a cluster that solely depend on its own treatment, and rerandomizes the pairwise-sequential randomization for the rest of the stuctures. Therefore, even the structures relying on the assignments of other clusters can be appropriately balanced with ReAR. In addition, the CAE adjusts the weights used by the HTEs according to the cluster structure. Numerical studies are conducted to demonstrate the significant improvement made by using our proposed procedure for estimating the ATEs from both of the two perspectives.

The outline of this paper is as follows. The framework is presented in Section 2. The HTEs for the two ATEs are studied in Section 3. The illustrating example is presented in Section 4. We propose our new procedure in Section 5. Numerical studies based on hypothetical and real world networks are presented in Section 6. The concluding remarks are given in Section 7. The proofs of the theoretical results, the details of the simulation settings, and one extra simulation study are presented in the supplementary material.

2 Notations and the Framework

Let C_k denote the set of the nodes that belong to cluster k, for $1 \le k \le K$. In some of the applications, the labels of the clusters are directly observed. For example, current education institutions of the students may correspond to clusters in the application comparing the treatments used to improve the student performance. In other scenarios, the information of the network, e.g., the adjacency matrix, can be used to generate the label of the clusters via community detection algorithms (Clauset et al., 2004). As the labels of the clusters can be obtained in either one of these two cases, we assume that both the network and the clusters are observed before the experiment.

In the design stage, graph-cluster randomization uses the clusters information by treating the cluster as the unit for randomization. Let T_k denote treatment assignments for cluster k, so that if $T_k = z$ then $Z_{k,i} = z$ for $1 \le i \le n_k$ and $z \in \{0,1\}$. In addition, denote $\boldsymbol{T} = (T_1, \cdots, T_K)'$ as the K-vector treatment assignment of the clusters. Without loss of generality, we assume the randomization used in graph-cluster randomization satisfies that $\mathbb{P}(T_k = 1|\mathcal{G}) = 1/2$, where $\mathcal{G}$ represents the information of G. Once the randomization is performed, the treatment assignments

of the N units are observed. Let $Y_{k,i}$ denote the observed response of node i in cluster k for $1 \le i \le n_k$. We assume that

$$Y_{k,i} = Y_{k,i}(\boldsymbol{Z}) = \sum_{\boldsymbol{z} \in \mathcal{Z}} \mathbb{I}(\boldsymbol{Z} = \boldsymbol{z}) Y_{k,i}(\boldsymbol{z}),$$

where $\mathcal{Z}$ is the domain of $\boldsymbol{Z}$. Therefore, only one version of the potential outcome is observed.

The network interference poses great challenge for estimating $\tau_{IN}(\mathbf{1}, \mathbf{0})$ or $\tau_{CL}(\mathbf{1}, \mathbf{0})$, as neither $\mathbf{1}$ nor $\mathbf{0}$ will be used in practice. As pointed in Basse and Airoldi (2018), an unbiased estimator may even not exist under the assumption of arbitrary network interference. The following assumption is introduced to reduce the global interference to a form of local interference.

Assumption 1 (Neighborhood Interference)**.** *Let $\delta_{k,i}$ denote the set of the neighbors of node i in cluster k. Define a partition of $\boldsymbol{Z}$ as $(Z_{k,i}, Z'_{\delta_{k,i}}, Z'_{-\delta_{k,i}})'$, where $Z_{\delta_{k,i}}$ represents the treatment assignments of the nodes in $\delta_{k,i}$, and $Z_{-\delta_{k,i}}$ represents the treatment assignments of the nodes in $\delta^c_{k,i}$. For any $1 \le k \le K$, $1 \le i \le n_k$, and $\forall \boldsymbol{Z}, \boldsymbol{Z}^\star$, if $Z_{k,i} = Z^\star_{k,i}$ and $Z_{\delta_{k,i}} = Z^\star_{\delta_{k,i}}$, then $Y_{k,i}(\boldsymbol{Z}) = Y_{k,i}(\boldsymbol{Z}^\star)$.*

Assumption 1 restricts the dependence of $Y_{k,i}$ on $\boldsymbol{Z}$ to the dependence on $Z_{k,i}$ and $Z_{\delta_{k,i}}$. Note that other forms of neighborhood interference conditions can also be used (Eckles et al., 2016; Aronow et al., 2017), one can adjust our results and the proposed method if other form of neighborhood interference is considered.

3 Horvitz-Thompson Estimator and its Properties

To construct the HTEs, we introduce the definition of the effectively treated node as follows.

Definition 3.1. *Under Assumption 1, let $\xi_{k,i}(z) = \mathbb{I}(Z_{k,i} = z, Z_{\delta_{k,i}} = \boldsymbol{z}_{\delta_{k,i}})$ and $\pi_{k,i}(z) = \mathbb{P}(\xi_{k,i}(z) = 1|\mathcal{G})$, where $\boldsymbol{z}_{\delta_{k,i}}$ is $|\delta_{k,i}|$-vector of z for $z \in \{0,1\}$ and $|A|$ is the size of A. For $z \in \{0,1\}$, the node i in cluster k is effectively treated with $Z_{k,i} = z$ if $\xi_{k,i}(z) = 1$.*

Under Assumption 1, if $\xi_{k,i}(1) = 1$, then $Y_{k,i} = Y_{k,i}(\mathbf{1})$; and if $\xi_{k,i}(0) = 1$, then $Y_{k,i} = Y_{k,i}(\mathbf{0})$. Therefore, if the nodes within the same cluster are assigned with the same treatment, the number of effectively treated nodes may increase and thus the bias generated from the interference can be reduced (Eckles et al., 2016). The effectively treated nodes can be used to construct the HTEs for $\tau_{IN}(\mathbf{1}, \mathbf{0})$ and $\tau_{CL}(\mathbf{1}, \mathbf{0})$ as

$$\hat{\tau}_{HT,IN} = \sum_{k=1}^{K} N^{-1} n_k \left[\hat{Y}_k(1) - \hat{Y}_k(0)\right],$$

$$\hat{\tau}_{HT,CL} = \sum_{k=1}^{K} K^{-1} \left[\hat{Y}_k(1) - \hat{Y}_k(0)\right],$$

where $\hat{Y}_k(z) = n_k^{-1} \sum_{i=1}^{n_k} \{[\pi_{k,i}(z)]^{-1} \xi_{k,i}(z)\} Y_{k,i}$ for $z \in \{0,1\}$. The weights $\pi_{k,i}(z)$ can be considered as the general propensity scores resulting in unbiased estimators for $\tau_{IN}(\mathbf{1}, \mathbf{0})$ and $\tau_{CL}(\mathbf{1}, \mathbf{0})$. These values are important to validate the use of the HTEs. When K is large, it is impractical to explicitly calculate $\pi_{k,i}(z)$. Ugander et al. (2013) suggest to use a network exposure model which characterizes the interference for a given randomization scheme. However, this model may not correctly specify the network structure. As an alternative, one can approximate $\pi_{k,i}(z)$ by using simulation to replicate $\boldsymbol{T}$ based on the given randomization procedure. For more details of this approach, we refer to Aronow et al. (2017).

To introduce the assumption ensuring that the HTEs are well defined, denote $\pi_{(k_1,i_1),(k_2,i_2)}(z_1, z_2) = \mathbb{E}\left[\xi_{k_1,i_1}(z_1)\xi_{k_2,i_2}(z_2)|\mathcal{G}\right]$ as the probability that two nodes from the clusters k_1, k_2 are effectively treated, for $1 \le k_1, k_2 \le K$, $1 \le i_1 \le n_{k_1}$, $1 \le i_2 \le n_{k_2}$, and $z_1, z_2 \in \{0,1\}$, respectively.

Assumption 2. *There exists $M_1 > 0$, such that $|Y_{k,i}(\boldsymbol{z})| < M_1$; for $1 \le k \le K$ and $1 \le i \le n_k$, $\pi_{k,i}(z) > 0$; and for $1 \le k_1, k_2 \le K$, $1 \le i_1 \le n_{k_1}$ and $1 \le i_2 \le n_{k_2}$, $\pi_{(k_1,i_1),(k_2,i_2)}(z_1, z_2) > 0$.*

Next, we develop the properties of the HTEs from the finite-sample population perspective in the following theorem. We denote $\mathbb{E}_{fs}[\cdot] = \mathbb{E}[\cdot|\mathcal{G}, \mathcal{Y}]$ and $\mathbb{V}_{fs}[\cdot] = \mathbb{V}[\cdot|\mathcal{G}, \mathcal{Y}]$ as the expectation and variance from the finite-sample population perspective, where $\mathcal{Y}$ is the sigma algebra generated by the N observed responses.

Theorem 3.1. *Under Assumptions 1 and 2, both $\hat{\tau}_{HT,IN}$ and $\hat{\tau}_{HT,CL}$ are unbiased, that is,*

$$\mathbb{E}_{fs}\left[\hat{\tau}_{HT,IN}\right] = \tau_{IN}(\mathbf{1}, \mathbf{0}), \quad \mathbb{E}_{fs}[\hat{\tau}_{HT,CL}] = \tau_{CL}(\mathbf{1}, \mathbf{0}),$$

and the variances of $\hat{\tau}_{HT,IN}$ and $\hat{\tau}_{HT,CL}$ satisfy that

$$\mathbb{V}_{fs}[\hat{\tau}_{HT,IN}] \lesssim O\left(N^{-2}\left[\sum_{k=1}^{K} n_k^2 + \kappa_G M_G\right]\right),$$
$$\mathbb{V}_{fs}[\hat{\tau}_{HT,CL}] \lesssim O\left(K^{-1} + K^{-2}\kappa_G M_G\right),$$

where M_G is maximum number of edges that connects two different clusters in G, and κ_G is the number of pairs of the connected clusters.

Theorem 3.1 complements Ugander et al. (2013) by providing additional results for $\hat{\tau}_{HT,CL}$. Similarly shown in Ugander et al. (2013), the valid use of $\hat{\tau}_{HT,IN}$ requires that the there can be only $O(N)$ clusters with size $O(1)$, and the interference between the clusters cannot be too strong, i.e., $\kappa_G M_G$ is bounded in N (Ugander et al., 2013; Ugander and Yin, 2020). Therefore, using $\hat{\tau}_{HT,IN}$ is desirable in situations where the cluster sizes n_k are bounded. However, n_k cannot

be assumed as bounded in many applications, and thus, using graph-cluster randomization together with $\hat{\tau}_{HT,IN}$ may be not appropriate in such situations. Theorem 3.1 shows that the order of $\mathbb{V}_{fs}[\hat{\tau}_{HT,CL}]$ does not depend on the cluster size n_k and the value decreases as K increases. Therefore, using $\hat{\tau}_{HT,CL}$ can be useful in situation where $\hat{\tau}_{HT,IN}$ does not work well.

4 Issues Related to Randomization and Horvitz-Thompson Estimators

Although Theorem 3.1 describes the usage of graph-cluster randomization together with the HTEs, several issues related to the randomization and the estimation could affect the results evaluating the ATEs. In this section, an example of a simple network is used to demonstrate these problems. Consider four clusters generated from a modified stochastic block models as shown in Figure 1, where the first two clusters belong to one type and the last two clusters belong to another. For simplicity, the labels of the clusters are assumed known before the experiment. For the design stage, we use the randomization scheme that two of the clusters are randomly chosen to one treatment arm with probability 1/6. The details of the simulation are presented in Section B of the supplementary materials.

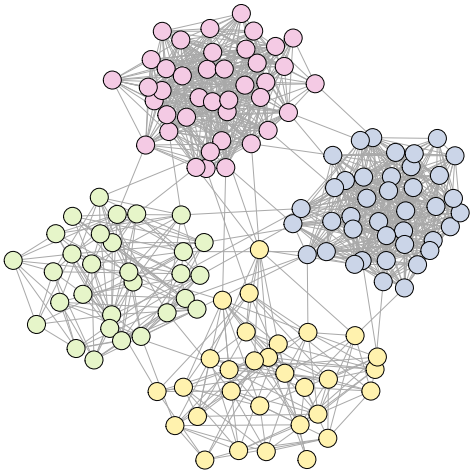

Figure 1: *An example of network consisting of 2 types of clusters. The probability that two nodes in different clusters are connected is* 180^{-1}.

Table 1: *Bias and standard deviation (SD) of the HTEs.*

	Bias	SD
$\hat{\tau}_{HT,IN}$	-0.060	1.005
$\hat{\tau}_{HT,CL}$	-0.058	0.589

We first evaluate the bias and the standard deviation (SD) of the HTEs in Table 1. Since a few nodes could have $\pi_{k,i}(z) = 0$, both of the two HTEs are slightly biased. In addition, the SD of $\hat{\tau}_{HT,CL}$ is much smaller than the SD of $\hat{\tau}_{HT,IN}$. Note that the values of the two ATEs are similar, i.e., $\tau_{IN}(\mathbf{1},\mathbf{0}) = 1$ and $\tau_{CL}(\mathbf{1},\mathbf{0}) = 0.979$. Thus, using $\hat{\tau}_{HT,CL}$ can be more efficient.

Other than the mean and SD, it is also desirable to evaluate the performance of the HTEs for a specific treatment assignment. We evaluate half of the six treatment assignments in Table 2. Since the first two clusters are of the same type, the most imbalanced treatment assignment $(0,0,1,1)'$ yields the values of the estimators quite different from $\tau_{IN}(\mathbf{1},\mathbf{0})$ and $\tau_{CL}(\mathbf{1},\mathbf{0})$. Although $(0,1,1,0)'$ and $(0,1,0,1)'$ are comparable with respect to cluster's type, the values of the HTEs under these two assignments are still quite different. This indicates that the choice of the assignment may greatly affect the evaluation of the ATE.

Table 2: *Evaluation of* $\hat{\tau}_{HT,IN}$ *and* $\hat{\tau}_{HT,CL}$ *and the difference-in-means (DIM) of the cluster-treated structures under three treatment assignments:* $X_{k,1,1}$ *is size of* In_k*;* $X_{k,1,2}$ *is the average degree of the nodes in* In_k*;* $X_{k,2,1}$ *is the number of effectively treated nodes in* Ot_k*; and* $X_{k,2,2}$ *is the average degree of the effectively treated nodes in* Ot_k*.*

$\boldsymbol{T}$	$(0,0,1,1)'$	$(0,1,0,1)'$	$(0,1,1,0)'$
$\hat{\tau}_{HT,IN}$	-0.730	0.711	0.919
$\hat{\tau}_{HT,CL}$	-0.001	0.622	0.929
$DIM(X_{k,1,1})$	-10.500	3.500	-0.500
$DIM(X_{k,1,2})$	-11.912	0.449	0.494
$DIM(X_{k,2,1})$	0.500	-2.500	-1.500
$DIM(X_{k,2,2})$	-9.225	-0.583	1.667

To better understand how the different treatment assignments affect the values of the HTEs, we introduce the concepts of inner nodes and outer nodes as follows.

Definition 4.1. *For* $1 \le k \le K$ *and* $1 \le i \le n_k$*, if* $Y_{k,i}(\boldsymbol{Z})$ *depends only on the treatment assignment of cluster* k*, i.e.,* T_k*, then node* i *is an inner node of cluster* k*; otherwise, node* i *is an outer node of cluster* k*.*

By Definition 4.1, we further denote In_k and Ot_k as the set of inner nodes and the set of outer nodes of cluster k, respectively. Assumption 1 implies that node i is an inner node of cluster k if $\delta_{k,i} \subset C_k$, and it is an outer node if $\delta_{k,i} \cap C_k^c \neq \emptyset$. This difference of nodes' types can affect their probability to be effectively treated. Under graph-cluster randomization, the weight of the HTEs satisfies that

$$\begin{aligned}\pi_{k,i}(z) &= \mathbb{P}(\xi_{k,i}(z) = 1|\mathcal{G}) \\ &= P(T_k = z|\mathcal{G})\mathbb{P}(\xi_{k,i}(z) = 1|T_k = z, \mathcal{G}), \quad (1)\end{aligned}$$

where $z \in \{0,1\}$. Therefore, Assumption 1 and the employed randomization scheme imply that $\pi_{k,i}(z) = 1/2$ if $i \in In_k$; $\pi_{k,i}(z) = 1/6$ if $i \in Ot_k$ and node i connects two clusters; and $\pi_{k,i}(z) = 0$ if $i \in Ot_k$ and node i connects more than two clusters. Therefore, the weights for the outer nodes are much larger than the weights for the inner nodes. As shown in Table 2,

the treatment assignment $(0,1,0,1)'$, resulting in more effectively treated outer nodes in the control arm, i.e., $DIM(X_{k,2,1}) = -2.5$, can impair the evaluation, because the weights amplify the outcomes of these nodes. Therefore, the comparability of the treatment arms also requires the balance of this kind of cluster-treated structures, such as the number of nodes in In_k, the average degree of the nodes in In_k, and the measures with respect to the effectively treated nodes in Ot_k. These structures are important because they may associate with the outcomes of different types of nodes used in estimation. We calculate the difference-in-means for four kinds of cluster-treated structures in Table 2. This table suggests that the most balanced treatment arms of the clusters with respect to the four measures, e.g., $(0,1,1,0)'$, has the best performance.

Besides the balance of cluster-treated structures, the weights of HTEs can also be problematic. This example also shows that the values of the weights are majorly determined by randomization, but are less determined by the cluster structures. In practice, the weight of an outer node can be extremely large not reflecting the network structure, i.e., when $\pi_{k,i}(z)$ is small due to the cluster structures, there may be only a few outer nodes in the network. Therefore, whether these nodes are effectively treated or not can greatly affect the evaluation of the ATE. This indicates the importance of the weights used for the construction of the estimator.

In practice, the network structure can be more complicated than the presented example. As such, purely counting on randomization to take care of cluster-treated structures can cause severe imbalanced treatment arms of clusters. Moreover, the probability of an outer node to be efficiently treated may not reflect the cluster structures. Consequently, using such weights as the HTEs may result in large variance. Therefore, the design and the estimation need to be adjusted according to network structure.

5 Adaptive A/B Test Procedure

In this section, we propose a rerandomized-adaptive randomization (ReAR) to balance cluster-treated structures and a cluster-adjusted estimator (CAE) to assign more appropriate weights for estimation. As such, the use of ReAR together with CAE can produce better results evaluating the ATE.

5.1 Rerandomized-Adaptive Randomization

Similar to the idea of the exposures to neighborhood treatments (Forastiere et al., 2021), when cluster is used as the unit for randomization, balance should take the nodes' type and the nodes' treatment exposures into account. The cluster-treated structures considered in Section 4 are the cluster level summary statistics demonstrating these concerns. Note that various measures of the cluster-treated structures can be used in different applications. As demonstrated in Section 4, using graph cluster randomization under Assumption 1 indicates that at least the measures associated with the inner nodes and those associated with the effectively treated outer nodes should be considered for balance.

According to the nature of these measures, the cluster-treated structures can be categorized as the following two types: (1) the measures that do not depend on the treatment assignment of other clusters, and (2) the measures whose values are determined only when the treatment assignment of other clusters are determined. For instance, the first type of measures may include the number of nodes in In_k, and the second type of measures may include the number of effectively treated nodes in Ot_k. Let $\boldsymbol{X}_{k,1}$ denote the m_1-covariates of cluster k whose values do not depend on the treatment assignments of other clusters, i.e., the first type of measures, and $\boldsymbol{X}_{k,2}$ denote the m_2-covariates of cluster k whose values depend on the treatment assignments of other clusters, i.e., the second type of measures. Therefore, $\boldsymbol{X}_k = (\boldsymbol{X}'_{k,1}, \boldsymbol{X}'_{k,2})'$ include both of the two types of cluster-treated structures.

The second type of the measures, i.e., $\boldsymbol{X}_{k,2}$, poses certain challenges for the design. Since $\boldsymbol{X}_{k,2}$ depends on the assignments of other clusters, traditional design approaches such as blocking, stratification, and the covariate-adaptive randomization (Pocock and Simon, 1975; Hu et al., 2012), that sequentially balance the covariates, are not applicable to $\boldsymbol{X}_{k,2}$. As an alternative, we can rerandomize the assignment $\boldsymbol{T}$ and choose one of such assignments that gives the desirable balanced treatment arms with respect to both $\boldsymbol{X}_{k,1}$ and $\boldsymbol{X}_{k,2}$ (Morgan et al., 2012; Morgan and Rubin, 2015). However, it may take a large number of rerandomizations to find one useful assignment (Qin et al., 2016). We thus consider to combine covariate-adaptive randomization and rerandomization to take the advantages of both of the two types of procedures.

Let B denote the number of rerandomization and $\boldsymbol{T}_b$ denote the bth treatment assignment generated from the pairwise-sequential randomization (Qin et al., 2016; Zhou et al., 2020) with $(\boldsymbol{X}_{1,1}, \cdots, \boldsymbol{X}_{K,1})'$ for $1 \leq b \leq B$. The pairwise-sequential randomization sequentially assigns a larger probability to the assignment leading to the smaller value of the Mahalanobis distance of $\boldsymbol{X}_{k,1}$ for a pair of clusters. Therefore, the imbalance with respect to $\boldsymbol{X}_{k,1}$ are minimized for large K. We present the pairwise-sequential randomization in Section C.1 of the supplementary material. Next, we rerandomize of the pairwise-sequential randomization to choose the treatment assignment which maintains the balance

with respect to both $\boldsymbol{X}_{k,1}$ and $\boldsymbol{X}_{k,2}$. Denote $\bar{\boldsymbol{X}}_b^1$ and $\bar{\boldsymbol{X}}_b^0$ as the sample means of $\boldsymbol{X}_k$ with respect to the treatment and control arms calculated with $\boldsymbol{T}_b$ for the bth rerandomization. In addition, let $\boldsymbol{D}_b = \bar{\boldsymbol{X}}_b^1 - \bar{\boldsymbol{X}}_b^0$ and denote $\boldsymbol{S}(\boldsymbol{D})$ as the sample covariance matrix calculated from $\boldsymbol{D}_1, \cdots, \boldsymbol{D}_B$. Consider the following modified Mahalanobis distance that measures the imbalance with respect to both $\boldsymbol{X}_{k,1}$ and $\boldsymbol{X}_{k,2}$ for the bth rerandomization,

$$Imb_b = \boldsymbol{D}_b' \boldsymbol{S}(\boldsymbol{D})^{-1} \boldsymbol{D}_b, \quad \text{for} \quad b = 1, \cdots, B.$$

Furthermore, denote $Imb_{(b)}$ as the bth ordered value of $Imb_1, \cdots, Imb_B$, then ReAR can be described as Algorithm 1. Note that an assignment is randomly selected from those with $Imb_{(1)}, \cdots, Imb_{(\lfloor \alpha B \rfloor)}$, the imbalance measure calculated with both $\boldsymbol{X}_{k,1}$ and $\boldsymbol{X}_{k,2}$ can thus be controlled by a given threshold α. Therefore, one can use a sufficiently large B and a small α to obtain the assignment with desirably balanced treatment arms. Furthermore, a relatively large α can increase the randomness for rerandomization and relieve the problem of confounding. For more detailed discussion for choosing B and α, please see Section C.2 of the supplementary material.

Algorithm 1 Rerandomized-Adaptive Randomization

1: **Input**: covariates $\boldsymbol{X}_1, \cdots, \boldsymbol{X}_K$; number of rerandomization B; threshold α;
2: **for** $b = 1$ **to** B **do**
3: Generate $\boldsymbol{T}_b$ from pairwise-squential randomization with $(\boldsymbol{X}_{1,1}, \cdots, \boldsymbol{X}_{K,1})'$;
4: **end for**
5: Calculate and order $Imb_1, \cdots, Imb_B$;
6: Select the b'th assignment with $Imb_{b'} \sim \text{Unif}(Imb_{(1)}, \cdots, Imb_{(\lfloor \alpha B \rfloor)})$;
7: **Output**: $\boldsymbol{T} = \boldsymbol{T}_{b'}$;

5.2 Cluster-Adjusted Estimator

To improve the estimation for ATE, it is still necessary to appropriately adjust the weight of the effectively treated nodes in estimation. Notice that a node of cluster k belongs to either In_k or Ot_k. This difference of nodes' type can affect the probability to be efficiently treated and consequently affects the cluster-treated structures. Ideally, the weight should reflect the nodes' type and should not over-amplify beyond the cluster size, e.g., if there is only one outer node in a cluster then the weight for this node should not be too large. We introduce the following weights that adaptively adjust the node according to the nodes' type

$$w_{k,i}(z) = \mathbb{P}(T_k = z|\mathcal{G}) \Big\{ \mathbb{I}(i \in In_k, T_k = z) + \mathbb{I}(i \in Ot_k) \frac{\sum_{j \in Ot_k} \xi_{k,j}(z)}{\sum_{j=1}^{n_k} \mathbb{I}(j \in Ot_k)} \Big\}.$$

The advantage of using $w_{k,i}(z)$ is that the weights for the outer nodes satisfy that $w_{k,i}(z) \leq \mathbb{P}(T_k = z|\mathcal{G}) \times |Ot_k|^{-1}$, whereas the weights for the inner nodes keep the same as the weights used by HTEs, i.e., $w_{k,i}(z) = \pi_{k,i}(z) = \mathbb{P}(T_k = z|\mathcal{G})$. Therefore, using $w_{k,i}$ can reduce the variance for estimating the ATEs. The CAEs for $\tau_{IN}(\mathbf{1}, \mathbf{0})$ and $\tau_{CL}(\mathbf{1}, \mathbf{0})$ can be defined as

$$\hat{\tau}_{CAE,IN} = \sum_{k=1}^{K} N^{-1} n_k \left[\tilde{Y}_k(1) - \tilde{Y}_k(0) \right],$$

$$\hat{\tau}_{CAE,CL} = \sum_{k=1}^{K} K^{-1} \left[\tilde{Y}_k(1) - \tilde{Y}_k(0) \right],$$

where $\tilde{Y}_k(z) = n_k^{-1} \sum_{i=1}^{n_k} [w_{k,i}(z)]^{-1} \ \xi_{k,i}(z) Y_{k,i}(z)$ for $z \in \{0, 1\}$. If none of the nodes in Ot_k are effectively treated, we replace n_k used in the CAEs with $|In_k|$.

Although the proposed CAEs can be biased, they still have several advantages compared to the HTEs. First, the CAEs and the HTEs are equivalent when the clusters are disjoint. Furthermore, CAEs can alleviate the problem of overweighting by adjusting weights of the nodes according the nodes' type. Therefore, the CAEs may have better performance when there are moderate amount of interference among the clusters.

6 Numerical Studies

In Section 6.1, hypothetical networks are used to show that the restrictions on cluster sizes can be alleviated when the estimators from the cluster perspective are used. In Section 6.2, the MIT phone call network listed in the Network Data Repository (Rossi and Ahmed, 2015) shows that the inappropriate weights used by HTEs may impair the evaluation of the ATEs, whereas our proposed adaptive A/B test procedure can still be useful. In Section D.3 of the supplementary material, the Facebook pages to pages network is used to demonstrate the performance of our proposed procedure, where the network contains only 64.1% of the inner nodes.

Throughout this section, the potential outcomes are generated according to the following model,

$$\begin{aligned} Y_{k,i}(\mathbf{0}) &= \mu_0 + \alpha_0 \cdot \bar{d}^{-1} d_{k,i} + \epsilon_{k,i}, \\ Y_{k,i}(\boldsymbol{Z}) &= Y_{k,i}(\mathbf{0}) + (\mu_1 - \mu_0) Z_{k,i} \\ &\quad + (\alpha_1 - \alpha_0) \cdot \bar{d}^{-1} \mathbf{1}_{\delta_{k,i}}' Z_{\delta_{k,i}}, \end{aligned} \tag{2}$$

where $d_{k,i}$ is the degree of node i in cluster k, $\bar{d}$ is the average degree of G, $\mathbf{1}_{\delta_{k,i}}$ is $|\delta_{k,i}|$-vector of ones, and $\epsilon_{k,i}$ are i.i.d. $\mathcal{N}(0, \sigma_\epsilon^2)$. The parameter setting $(\mu_1, \mu_0, \alpha_1, \alpha_0, \sigma_\epsilon) = (1.6, 1, 1.4, 1, 1)$ is used, resulting in the direct effect $\mu_1 - \mu_0 = 0.6$ and the spillover effect $\alpha_1 - \alpha_0 = 0.4$, respectively.

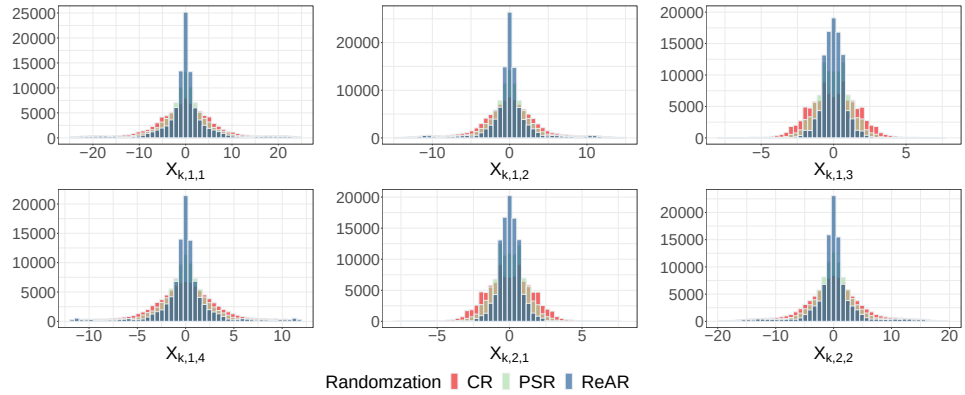

Figure 2: *Histograms of the difference-in-means of the six measures of the cluster-treated structures for* $K = 20$.

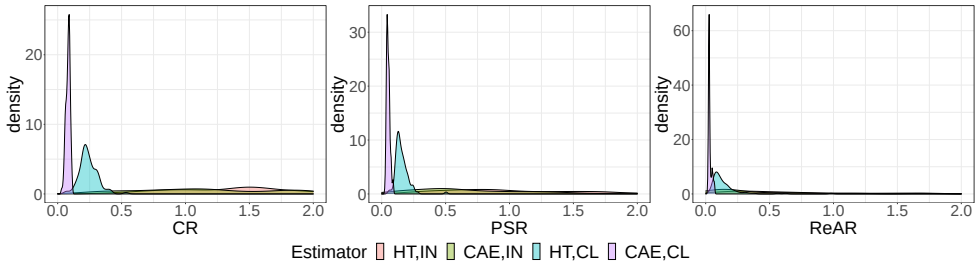

Figure 3: *The densities of the MSEs evaluated from the 100 hypothetical networks for* $K = 20$.

The settings of the design and estimation used for simulations are as follows. First, six measures of cluster-treated structures, i.e., $m_1 = 4$ and $m_2 = 2$, are considered, where $X_{k,1,1}$ is the size of In_k, $X_{k,1,2}$ is the average degree of the nodes in In_k, $X_{k,1,3}$ is the size of Ot_k, $X_{k,1,4}$ is the average degree of the nodes in Ot_k, $X_{k,2,1}$ is the number of the effectively treated nodes in Ot_k, and $X_{k,2,2}$ is the average degree of the effectively treated nodes in Ot_k. For the design stage, the following three randomization procedures with graph-cluster randomization are evaluated: complete randomization (CR), i.e., no neighborhood exposures are used; pairwise-sequential randomization (PSR) with $\boldsymbol{X}_{k,1}$; and ReAR using $B = 1000$ and $\alpha = 0.05$ with both $\boldsymbol{X}_{k,1}$ and $\boldsymbol{X}_{k,2}$. For the estimation stage, the HTEs from the two perspectives are compared with the corresponding CAEs for each of the randomization schemes. The weights for the HTEs are calculated with the simulation approach (Aronow et al., 2017).

6.1 Hypothetical Networks

The stochastic block model is modified to mimic the real-world network with clear cluster structures. According to the details presented in Section D.1 of the supplementary material, 100 networks are generated and 1,000 simulations are conducted for each of networks. The results are combined to generate the histograms of the difference-in-means (DIMs) of the six cluster-treated structures and are presented in Figure 2. The distributions of the MSEs for different estimators are presented in Figure 3. Note that the figures for $K = 50$ are similar to the Figure 2 and Figure 3 and are thus omitted. In Table 3, we compare the performance of different estimators under different randomization schemes with graph-cluster randomization.

The histograms presented in Figure 2 demonstrate the performance of the three designs on balancing the six measures of the cluster-treated structures. As $\boldsymbol{X}_{k,1}$ is used in pairwise-sequential randomization, the distributions of the difference-in-means are more concentrated around zero under pairwise-sequential randomization than the distributions under complete randomization. Because the imbalance score used by ReAR are calculated with both $\boldsymbol{X}_{k,1}$ and $\boldsymbol{X}_{k,2}$, the difference-in-means of all of the six measures are all more concentrated than their values under pairwise-sequential randomization. Therefore, the treatment arms generated by ReAR are of the most balanced with respect to the cluster-treated structures.

Comparing the different approaches for evaluating the ATEs, we have the following three folds of meaningful observations. First, the balance of the cluster-treated structures benefits the estimation of the ATEs. The average MSEs of the estimators following the design with better balance properties are smaller. In particular, the average MSEs of the estimators following ReAR are the smallest. Comparing the performance of the two kinds of estimators, the CAEs have slightly larger bias than the HTEs, but have smaller average SD. Therefore, the CAEs have better performance than the HTEs. Last but not least, the estimators from

Table 3: *The average bias, the average standard deviation (SD), and the average MSE of the HTEs and CAEs.*

K	Design	CR			PSR			ReAR		
	Estimators	Bias	SD	MSE	Bias	SD	MSE	Bias	SD	MSE
20	$\hat{\tau}_{HT,IN}$	0.004	1.784	3.774	-0.000	1.491	2.919	0.003	1.298	2.509
	$\hat{\tau}_{CAE,IN}$	-0.015	1.622	3.302	-0.018	1.328	2.532	-0.011	1.012	1.869
	$\hat{\tau}_{HT,CL}$	0.001	0.485	0.241	0.001	0.390	0.156	-0.000	0.373	0.171
	$\hat{\tau}_{CAE,CL}$	-0.002	0.282	0.081	-0.002	0.224	0.051	-0.002	0.177	0.032
50	$\hat{\tau}_{HT,IN}$	0.005	1.475	2.672	0.003	1.100	1.838	0.001	0.873	1.452
	$\hat{\tau}_{CAE,IN}$	-0.010	1.395	2.481	-0.012	1.042	1.733	-0.013	0.816	1.359
	$\hat{\tau}_{HT,CL}$	0.000	0.277	0.078	0.001	0.178	0.032	0.000	0.131	0.018
	$\hat{\tau}_{CAE,CL}$	-0.001	0.183	0.034	-0.001	0.124	0.016	-0.001	0.099	0.010

Table 4: *The bias, standard deviation (SD), and MSE of the estimators evaluated for the MIT phone call networks.*

Design	CR			PSR			ReAR		
Estimators	Bias	SD	MSE	Bias	SD	MSE	Bias	SD	MSE
$\hat{\tau}_{HT,IN}$	0.428	0.962	1.108	0.305	0.539	0.383	0.249	0.369	0.198
$\hat{\tau}_{CAE,IN}$	-0.028	0.342	0.117	-0.028	0.114	0.014	-0.029	0.069	0.006
$\hat{\tau}_{HT,CL}$	0.401	0.808	0.814	0.319	0.686	0.572	0.208	0.264	0.113
$\hat{\tau}_{CAE,CL}$	-0.001	0.142	0.020	0.002	0.138	0.019	0.001	0.135	0.018

the cluster perspective have better performance than the estimators from the individual level perspective. Note that the average SDs of the estimators from the individual perspective all above 0.8. On the other hand, the average SDs of the estimators from the cluster perspective are much smaller. In addition, the distributions of the MSEs for the estimators from the cluster perspective are more concentrated and more shifted to zero. Therefore, using the estimators from the cluster perspective may be useful in situations where the estimators from the individual perspective have large variances.

6.2 The MIT Phone Call network

This network consists of phone calls/voicemails between 6819 users at MIT, where nodes and edges represent users and calls/voicemails, respectively (Eagle and Pentland, 2006). In practice, an A/B test may be conducted by the landline provider to see the satisfactory improvement measured from the MIT faculties for a new service. To conduct such experiment, the network is partitioned to 82 clusters by the label propagation algorithm (Raghavan et al., 2007). Note that the portion of the inner nodes for this network is about 93%. Therefore, the outer nodes should not affect much on the evaluation of the ATEs. However, simulation results suggests that the weights of the outer nodes used by the HTEs can still affect the estimation.

Table 4 shows the advantage of using the CAEs for adjusting the inappropriate weight used by the HTEs. Note that $\tau_{IN}(\mathbf{1},\mathbf{0}) = 1$, and $\tau_{CL}(\mathbf{1},\mathbf{0})$ evaluated from this network graph is 0.99. The biases of the HTEs all above 0.2. Therefore, using the HTEs for the ATE may not be reliable. On the other hand, the biases of the CAEs are all less than 0.03. The SDs and MSEs of the CAEs are also much smaller than the values for the HTEs. Therefore, the CAEs can be more useful than the HTEs, when several nodes in the network have small probabilities to be efficiently treated.

In addition, Table 4 indicates the good performance of our proposed adaptive A/B test procedure. As the ReAR can generate more balanced treatment arms, the potential outcomes from the two treatment groups under ReAR are more feasible to be compared. Therefore, our proposed methods can greatly improve the performance of the estimators by reducing the SDs for estimating either $\tau_{IN}(\mathbf{1},\mathbf{0})$ or $\tau_{CL}(\mathbf{1},\mathbf{0})$.

7 Conclusion

In this article, we discuss several issues about using cluster information in the design and the estimation for network A/B tests. We show that the estimator from the cluster perspective can be useful in some of the network settings. Furthermore, we demonstrate the importance of balancing the cluster-treated structures and the appropriate adjustment of the weight used in the estimation according to the cluster structures.

Our work can be further extended in several ways. Note that the covariate balance promoted by the design often affects the distribution of the test statistics (Ma et al., 2015, 2020; Bugni et al., 2018), the valid inference with our procedure may require an appropriate estimator of the variance of the CAE following ReAR. This problem needs further investigation. Moreover, the equal allocation may not optimize the power for testing the ATE. Therefore, one should assign the units according to the ratio that can maximize the power for testing the ATE. It may be desirable to propose network adaptive A/B test procedure that has the same spirits as the response-adaptive randomization (Hu and Rosenberger, 2006; Zhang et al., 2007) to achieve this goal.

8 Response to the Reviewers

We would like to thank the meta-reviewer and the four anonymous reviewers for the supportive comments and the efforts to improve our manuscript. We have revised the manuscript to address the concerns raised by the reviewers. In particular, the overloaded acronyms are reduced, the abstract and the first three sections are shortened, and more details were added to Section 5. Other typos have also been checked throughout the manuscript. We describe our revision in more details as follows.

1. The paragraph following Theorem 3.1 is revised. The performances of $\hat{\tau}_{HT,IN}$ and $\hat{\tau}_{CL}$ are discussed and compared under different scenarios of the network structure. We also emphasize on our novelty compared to Ugander et al. (2013).

2. In the fourth paragraph of Section 4, we added explanations demonstrating the importance of the balance for certain types of cluster measures in the design. Furthermore, more descriptions about these measures are added in the first paragraph of Section 5.1. The term "cluster-treated structures" is introduced to describe these measures, because they depend on both the cluster structures and the treatment assignment of the clusters.

3. We added a brief description about the pairwise-sequential randomization. Furthermore, we referred to Section C.1 of the supplementary materials for more details about the pairwise-sequential randomization.

4. The first paragraph of Section 5.2 is also revised. We added more details about the explanations about our proposed cluster-adjusted estimator.

9 Acknowledgements

We thank the reviewers for the helpful comments, which led to a much improved version of this paper. This research project was supported partly by a National Science Foundation grant DMS-1712760. The work of Feifang Hu was conducted as a consulting researcher at Baidu Research, USA. He thanks Baidu Research, USA, for their generous support.

References

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

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

Bugni, F. A., Canay, I. A., and Shaikh, A. M. (2018). Inference under covariate-adaptive randomization. Journal of the American Statistical Association, 113(524):1784–1796.

Clauset, A., Newman, M. E., and Moore, C. (2004). Finding community structure in very large networks. Physical review E, 70(6):066111.

Eagle, N. and Pentland, A. (2006). Reality mining: sensing complex social systems. Personal and Ubiquitous Computing, 10(4):255–268.

Eckles, D., Karrer, B., and Ugander, J. (2016). Design and analysis of experiments in networks: Reducing bias from interference. Journal of Causal Inference, 5(1).

Forastiere, L., Airoldi, E. M., and Mealli, F. (2021). Identification and estimation of treatment and interference effects in observational studies on networks. Journal of the American Statistical Association, 116(534):901–918.

Gui, H., Xu, Y., Bhasin, A., and Han, J. (2015). Network a/b testing: From sampling to estimation. In Proceedings of the 24th International Conference on World Wide Web, pages 399–409.

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

Hu, F. and Rosenberger, W. F. (2006). The theory of response-adaptive randomization in clinical trials. John Wiley & Sons, Hoboken, N.J.

Hu, Y., Hu, F., et al. (2012). Asymptotic properties of covariate-adaptive randomization. The Annals of Statistics, 40(3):1794–1815.

Kohavi, R., Deng, A., Frasca, B., Walker, T., Xu, Y., and Pohlmann, N. (2013). Online controlled experiments at large scale. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 1168–1176.

Leskovec, J., Lang, K. J., Dasgupta, A., and Mahoney, M. W. (2008). Statistical properties of community structure in large social and information networks. In Proceedings of the 17th international conference on World Wide Web, pages 695–704.

Leskovec, J. and Sosič, R. (2016). Snap: A general-purpose network analysis and graph-mining library. ACM Transactions on Intelligent Systems and Technology (TIST), 8(1):1.

Ma, W., Hu, F., and Zhang, L. (2015). Testing hypotheses of covariate-adaptive randomized clinical trials. Journal of the American Statistical Association, 110(510):669–680.

Ma, W., Qin, Y., Li, Y., and Hu, F. (2020). Statistical inference for covariate-adaptive randomization procedures. Journal of the American Statistical Association, 115(531):1488–1497.

Morgan, K. L. and Rubin, D. B. (2015). Rerandomization to balance tiers of covariates. Journal of the American Statistical Association, 110(512):1412–1421.

Morgan, K. L., Rubin, D. B., et al. (2012). Rerandomization to improve covariate balance in experiments. The Annals of Statistics, 40(2):1263–1282.

Pocock, S. J. and Simon, R. (1975). Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. Biometrics, 31(1):103–115.

Qin, Y., Li, Y., Ma, W., and Hu, F. (2016). Pairwise sequential randomization and its properties. arXiv preprint arXiv:1611.02802.

Raghavan, U. N., Albert, R., and Kumara, S. (2007). Near linear time algorithm to detect community structures in large-scale networks. Physical review E, 76(3):036106.

Rosenberger, W. F. and Lachin, J. M. (2015). Randomization in clinical trials: theory and practice. John Wiley & Sons, Hoboken, N.J, 2nd edition.

Rossi, R. A. and Ahmed, N. K. (2015). The network data repository with interactive graph analytics and visualization. In AAAI.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of educational Psychology, 66(5):688.

Splawa-Neyman, J., Dabrowska, D. M., and Speed, T. (1990). On the application of probability theory to agricultural experiments. essay on principles. section 9. Statistical Science, pages 465–472.

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

Ugander, J. and Yin, H. (2020). Randomized graph cluster randomization. arXiv preprint arXiv:2009.02297.

Zhang, L.-X., Hu, F., Cheung, S. H., and Chan, W. S. (2007). Asymptotic properties of covariate-adjusted response-adaptive designs. The Annals of Statistics, 35(3):1166–1182.

Zhou, Y., Liu, Y., Li, P., and Hu, F. (2020). Cluster-adaptive network a/b testing: From randomization to estimation. arXiv preprint arXiv:2008.08648.

Supplementary Material: Adaptive A/B Test on Networks with Cluster Structures

A Proof of Theorem 3.1.

In this section, we provide the proof for Therorem 3.1 using a similar approach as presented in Ugander et al. (2013). To derive Theorem 3.1, we first prove the following result.

Proposition A.1. *For $z \in \{0,1\}$, let $\bar{Y}_k(z) = n_k^{-1} \sum_{i=1}^{n_k} Y_{k,i}(z \cdot \mathbf{1})$, and let $\mathrm{cov}_{fs}[\cdot]$ denote the covariance function from the finite sample perspective. Under Assumptions 1 and 2, the following statements hold.*

1. *For $1 \le k \le K$ and $z \in \{0,1\}$, $\mathbb{E}_{fs}[\hat{Y}_k(z)] = \bar{Y}_k(z)$.*
2. *For $1 \le k_1, k_2 \le K$, and $z_1, z_2 \in \{0,1\}$*

$$\begin{aligned}
&\mathrm{cov}_{fs}\left[\hat{Y}_{k_1}(z_1), \hat{Y}_{k_2}(z_2)\right] \\
&= \frac{1}{n_{k_1} n_{k_2}} \sum_{i_1=1}^{n_{k_1}} \sum_{i_2=1}^{n_{k_2}} \frac{\pi_{(k_1,i_1),(k_2,i_2)}(z_1, z_2) - \pi_{k_1,i_1}(z_1)\pi_{k_2,i_2}(z_2)}{\pi_{k_1,i_1}(z_1)\pi_{k_2,i_2}(z_2)} \\
&\times Y_{k_1,i_1}(z_1 \cdot \mathbf{1}) Y_{k_1,i_2}(z_2 \cdot \mathbf{1}).
\end{aligned}$$

   (a) *For $k_1 = k_2 = k$, $z_1 \neq z_2$,*

$$\mathrm{cov}_{fs}\left[\hat{Y}_k(z_1), \hat{Y}_k(z_2)\right] = -\bar{Y}_k(1)\bar{Y}_k(0) = O(n_k^2 n_k^{-2}) = O(1).$$

   (b) *For $k_1 = k_2 = k$, $z_1 = z_2 = z$,*

$$\begin{aligned}
&\mathrm{cov}_{fs}\left[\hat{Y}_k(z), \hat{Y}_k(z)\right] = \mathbb{V}_{fs}\left[\hat{Y}_k(z)\right] \\
&= \frac{1}{n_k^2} \sum_{i_1=1}^{n_k} \sum_{i_2=1}^{n_k} \frac{\pi_{(k,i_1),(k,i_2)}(z, z) - \pi_{k,i_1}(z)\pi_{k,i_2}(z)}{\pi_{k,i_1}(z)\pi_{k,i_2}(z)} \\
&\times Y_{k,i_1}(z \cdot \mathbf{1}) Y_{k,i_2}(z \cdot \mathbf{1}) \\
&= O(n_k^{-2} n_k^2) = O(1).
\end{aligned}$$

   (c) *For $k_1 \neq k_2$,*

$$\mathrm{cov}_{fs}\left[\hat{Y}_{k_1}(z_1), \hat{Y}_{k_2}(z_2)\right] = O\left(\{n_{k_1} n_{k_2}\}^{-1} M_{C_{k_1}, C_{k_2}}\right),$$

   where $M_{C_{k_1}, C_{k_2}}$ is the number of edges connecting clusters k_1 and k_2.

Proof of Proposition A.1. First, under Assumption 1, it follows from the definition of the effectively treated node that

$$\mathbb{E}_{fs}[\xi_{k,i}(z) Y_{k,i}(\boldsymbol{Z})] = \pi_{k,i}(z) Y_{k,i}(z \cdot \mathbf{1}),$$

for $z \in \{0,1\}$. Therefore, $\mathbb{E}_{fs}[\hat{Y}_k(z)] = \bar{Y}_k(z)$ and

$$\mathbb{E}_{fs}[\hat{Y}_{k_1}(z_1)\hat{Y}_{k_2}(z_2)] = \frac{1}{n_{k_1} n_{k_2}} \sum_{i_1=1}^{n_{k_1}} \sum_{i_2=1}^{n_{k_2}} \frac{\pi_{(k,i_1),(k,i_2)}(z_1, z_2)}{\pi_{k_1,i_1}(z_1)\pi_{k_2,i_2}(z_2)} Y_{k_1,i_1}(z_1 \cdot \mathbf{1}) Y_{k_2,i_2}(z_2 \cdot \mathbf{1}),$$

for $z_1, z_2 \in \{0,1\}$. We next prove for (a) - (c) of 2.

(a) For $k_1 = k_2 = k$, $z_1 \neq z_2$, first notice that $T_k(1 - T_k) = 0$, then $\xi_{k,i_1}(z_1)\xi_{k,i_2}(z_2) = 0$ and thus $\pi_{(k_1,i_1),(k_2,i_2)}(z_1, z_2) = 0$. Therefore, we have

$$\mathrm{cov}_{fs}\left[\hat{Y}_k(z_1), \hat{Y}_k(z_2)\right] = -\bar{Y}_k(1)\bar{Y}_k(0).$$

Note that Assumption 2 implies that $Y_{k,i}(z) = O(1)$. Then we have $\bar{Y}_k(z) = O(n_k^{-1}n_k)$ and thus (a) follows.

(b) For $k_1 = k_2 = k$, $z_1 = z_2 = z$, each term within the summation sign is $O(1)$. The sum goes for n_k^2 terms, so that the sum is $O(n_k^2)$ and (b) follows.

(c) For $k_1 \neq k_2$, consider the terms within the summation sign under the following three situations. (i) Under Assumption 1, for all $i_1 \in In_{k_1}$, $Y_{k_1,i_1}(\boldsymbol{Z})$ does not dependent on the treatment assignment of the nodes in other cluster. Therefore, $\xi_{k_1,i_1}(z_1)$ and $\xi_{k_2,i_2}(z_2)$ are conditional independent given G. Then the terms for this case ($i_1 \in In_{k_1}, i_2$) are zero as $\pi_{(k_1,i_1),(k_2,i_2)}(t_1, t_2) = \pi_{k_1,i_1}(z_1)\pi_{k_2,i_2}(z_2)$. (ii) Next, if $i_1 \in Ot_{k_1}$ and $i_2 \notin \delta_{k_1,i_1}$, the corresponding terms are also zero. (iii) If $i \in Ot_{k_1}$ and $i_2 \in \delta_{k_1,i_1}$, the terms are $O(1)$. Therefore, by (i), (ii) and (iii), the total number of terms within the summation sign is $M_{C_{k_1},C_{k_2}}$. Therefore, (c) follows. □

Proof of Theorem 3.1. The unbiasedness of $\hat{\tau}_{HT,IN}$ and $\hat{\tau}_{HT,CL}$ simply follows from Proposition A.1. We now derive the bounds for the finite-population variances for $\hat{\tau}_{HT,IN}$ and $\hat{\tau}_{HT,CL}$. It follows from Proposition A.1 that

$$\begin{aligned}
\mathbb{V}_{fs}\left[\hat{\tau}_{HT,IN}\right] &= \sum_{k=1}^{K} \frac{n_k^2}{N^2}\mathbb{V}_{fs}\left[\hat{Y}_k(1) - \hat{Y}_k(0)\right] + \sum_{k_1,k_2} \frac{n_{k_1}n_{k_2}}{N^2}\mathrm{cov}_{fs}\left[\hat{Y}_{k_1}(1) - \hat{Y}_{k_1}(0), \hat{Y}_{k_2}(1) - \hat{Y}_{k_2}(0)\right] \\
&\lesssim O\left(N^{-2}\sum_{k=1}^{K} n_k^2\right) + O\left(N^{-2}\sum_{k_1\neq k_2} M_{C_{k_1},C_{k_2}}\right) \\
&\lesssim O\left(N^{-2}\left\{\sum_{k=1}^{K} n_k^2 + \kappa_G M_G\right\}\right), \\
\mathbb{V}_{fs}\left[\hat{\tau}_{HT,CL}\right] &= \frac{1}{K^2}\left\{\sum_{k=1}^{K}\mathbb{V}\left[\hat{Y}_k(1) - \hat{Y}_k(0)\right] + \sum_{k_1\neq k_2}\mathrm{cov}_{fs}\left[\hat{Y}_{k_1}(1) - \hat{Y}_{k_1}(0), \hat{Y}_{k_2}(1) - \hat{Y}_{k_2}(0)\right]\right\} \\
&\lesssim O\left(K^{-2}K\right) + O(K^{-2}\kappa_G \sum_{k_1,k_2} M_{C_{k_1},C_{k_2}}) \\
&\lesssim O\left(K^{-1} + K^{-2}\kappa_G M_G\right),
\end{aligned}$$

where $\mathbb{V}_{fs}\left[\hat{Y}_k(1)\right]$, $\mathbb{V}_{fs}\left[\hat{Y}_k(0)\right]$, and $\mathrm{cov}_{fs}\left[\hat{Y}_k(1), \hat{Y}_k(0)\right]$ are $O(1)$; and $\mathrm{cov}_{fs}\left[\hat{Y}_{k_1}(1), \hat{Y}_{k_2}(0)\right]$, $\mathrm{cov}_{fs}\left[\hat{Y}_{k_1}(1), \hat{Y}_{k_2}(1)\right]$, and $\mathrm{cov}_{fs}\left[\hat{Y}_{k_1}(0), \hat{Y}_{k_2}(0)\right]$ are $O(n_{k_1}^{-1}n_{k_2}^{-1}M_{C_{k_1},C_{k_2}})$. This completes the proof of this theorem. □

B Example used in Section 4

In this section, we discuss the details of the network and the design used for the illustrative example presented in Section 4. The setting of the network is presented in Section B.1, and the employed randomization scheme is discussed in Section B.2.

B.1 Generation of the Network and the Responses

The network used in Section 4 is generated as follows. Let $G(n, p)$ be the Erdös-Rényi random graph model, where n is the number of nodes and p is the probability that two of the n nodes are connected. Suppose the first two clusters, C_1 and C_2, are generated from $G(40, 0.5)$, and the last two clusters, C_3 and C_4, are generated from $G(30, 0.3)$. Therefore, the first two clusters have larger number of nodes and larger values of the average within-cluster node degree. Furthermore, we assume that two nodes from different clusters are connected with probability 180^{-1}, so that the four clusters are sparsely connected.

The potential outcomes are generated from (2) according to the parameter setting $(\mu_1, \mu_0, \alpha_1, \alpha_0, \sigma_\epsilon) = (1.6, 1, 1.4, 1, 1)$. Then, the direct effect and the spill-over effect are $\mu_1 - \mu_0 = 0.6$ and $\alpha_1 - \alpha_0 = 0.4$, respectively.

B.2 Randomization Design

As the network consists of only four clusters, the randomization schemes that assign three or all four clusters to the same treatment arm may not be desirable. Therefore, consider the randomization scheme that assigns two of the clusters to the same treatment arm, that is,

$$\mathbb{P}(\boldsymbol{T} = \boldsymbol{t}_j) = 1/6,$$

for $j = 1, \cdots, 6$ and $\{\boldsymbol{t}_1, \cdots, \boldsymbol{t}_6\} = \{(0,0,1,1)', (0,1,0,1)', (0,1,1,0)', (1,1,0,0)', (1,0,1,0)', \ (1,0,0,1)'\}$. According to this randomization scheme, the probability of a node being effectively treated can be calculated as

$$\pi_{k,i}(z) = \begin{cases} \mathbb{P}(T_k = z) = \frac{1}{2} & \text{if node } i \text{ only connects with nodes in cluster } k \\ \mathbb{P}(T_k = T_s = z) = \frac{1}{6} & \text{if node } i \text{ connects to nodes in clusters } k \text{ and } s \\ 0 & \text{if node } i \text{ connects to more than 2 clusters} \end{cases} \tag{3}$$

for $z \in \{0, 1\}$. (3) shows that using $\pi_{k,i}(z)$ as the weight for the HTEs can be problematic. Note that the values of $\pi_{k,i}(z)$, i.e., $1/2$ and $1/6$, are determined by the randomization scheme, but not the cluster structures. Therefore, the value $1/6$ can be different from the fraction of the effectively treated outer nodes in a cluster. For instance, if such fraction is larger than $1/6$, i.e., the cluster contains only one outer nodes, then the evaluated effect from this cluster can be greatly affected due to this inappropriate weight.

C Further Information about the Rerandomized adaptive randomization

In this section, we first describe the pairwise sequential randomization (PSR) used in rerandomized adaptive randomization (ReAR) and briefly discuss how to choose the parameters for ReAR.

C.1 Pairwise-Sequential Randomization

Algorithm 2 Pairwise-sequential randomization.

Input: covariates $\boldsymbol{X}_{1,1}, \cdots, \boldsymbol{X}_{K,1}$; probability of the biased coin $1/2 < q < 1$;
Compute $\boldsymbol{S}_1$ based on $\boldsymbol{X}_{1,1}, \cdots, \boldsymbol{X}_{K,1}$;
Assign $T_1 \sim Bernoulli(1/2)$ and set $T_2 = 1 - T_1$;
for $k = 2$ **To** $\lceil K/2 \rceil$ **do**
  if $2k \leq K$ **then**
    ▷ Suppose $T_1, \cdots, T_{2(k-1)}$ are generated, next generate $(T_{2k-1}, T_{2k})'$ as follows;
    ▷ Let $Mah_{2k}^{(1)}$ and $Mah_{2k}^{(2)}$ be the pseudo imbalance scores computed by $\boldsymbol{X}_{1,1}, \cdots \boldsymbol{X}_{2k,1}$;
    Compute $Mah_{2k}^{(1)}$ from (4) by assuming $(T_{2k-1}, T_{2k})' = (0,1)'$;
    Compute $Mah_{2k}^{(2)}$ from (4) by assuming $(T_{2k-1}, T_{2k})' = (1,0)'$;
    if $Mah_{2k}^{(1)} < Mah_{2k}^{(2)}$ **then**
      Assign $T_{2k-1} \sim Bernoulli(1-q)$;
    end if
    if $Mah_{2k}^{(1)} > Mah_{2k}^{(2)}$ **then**
      Assign $T_{2k-1} \sim Bernoulli(q)$;
    else
      Assign $T_{2k-1} \sim Bernoulli(1/2)$;
    end if
    Set $T_{2k} = 1 - T_{2k-1}$;
  else
    Assign $T_{2k-1} \sim Bernoulli(1/2)$;
  end if
end for

For $0 < m < K$, let $m_1 = \sum_{k=1}^{m} T_k$ and $m_2 = \sum_{k=1}^{m}(1 - T_k)$. Denote $\bar{\boldsymbol{X}}_m^1 = m_1^{-1}\sum_{k=1}^{m} T_k \boldsymbol{X}_{k,1}$, $\bar{\boldsymbol{X}}_m^2 = m_2^{-1}\sum_{k=1}^{m}(1-T_k)\boldsymbol{X}_{k,1}$, and $\bar{\boldsymbol{X}}_1 = K^{-1}\sum_{k=1}^{K}\boldsymbol{X}_{k,1}$ as the sample averages of $\boldsymbol{X}_{k,1}$ with respect to the treated clusters, the controlled clusters, and all of the clusters, respectively. Furthermore, denote $\boldsymbol{\zeta}_k = \bar{\boldsymbol{X}}_k^1 - \bar{\boldsymbol{X}}_k^2$, then the Mahalanobis distance of $\boldsymbol{X}_{k,1}$ calculated for the first $2k$ clusters can be written as

$$Mah_{2k} = \boldsymbol{\zeta}_{2k}' \mathrm{cov}\left[\boldsymbol{\zeta}_{2k}\right]^{-1} \boldsymbol{\zeta}_{2k} \propto \frac{j}{2} \cdot \boldsymbol{\zeta}_{2k}' \boldsymbol{S}_1^{-1} \boldsymbol{\zeta}_{2k}, \tag{4}$$

where $\boldsymbol{S}_1 = K^{-1}\sum_{k=1}^{K}(\boldsymbol{X}_{k,1} - \bar{\boldsymbol{X}}_1)(\boldsymbol{X}_{k,1} - \bar{\boldsymbol{X}}_1)'$ is the sample covariance matrix of $\boldsymbol{X}_{k,1}$. The pairwise-sequential randomization is presented in Algorithm 2. For more details about the theoretical properties of the pairwise-sequential randomization, we refer to Qin et al. (2016); Zhou et al. (2020).

C.2 The Choice of the parameters

As the pairwise-sequential randomization is rerandomized in ReAR, the choice of B and α can ensure the balance with respect to $\boldsymbol{X}_{k,2}$ and further improve the balance with respect to $\boldsymbol{X}_{k,1}$. We can choose a large enough value of B, and a small enough value of α to guarantee the performance of ReAR. For instance, one can use $\alpha = B^{-1}$ and a large value of B, e.g., $B = 1000$, that is, using the treatment assignment that results in the smallest value of the imbalance measure. Note that if the users in the experiment have the information of the network and the design, they may use ReAR to predict the treatment assignment. Their prediction may affect their decision to join the experiment and affect the evaluation of the ATEs. In such case, when B is large, choosing $\alpha = B^{-1}$ may increase the predictability of the treatment assignment. To reduce such predictability and maintain the performance of ReAR, one can choose a relatively large value of α, e.g., $\alpha \in (B^{-1}, 0.05]$. For the discussion about the predictability of the treatment assignment, please see chapter 5 of Rosenberger and Lachin (2015). In our numerical studies, $(q, B, \alpha)' = (0.85, 1000, 0.05)'$ is used to ensure the balance of the two treatment arms with respect to the cluster-treated structures.

D Further Information about the Numerical studies

This section provides additional details about the numerical studies with the hypothetical networks and the MIT network. Furthermore, the Facebook page to page network is studied to show that our proposed method can still have good performance even when the number of the inner node is not as much as expected. The weights used for the HTEs are calculated from 1000 simulations by using the approach employed by Aronow et al. (2017).

D.1 The Hypothetical Network in Section 6.1

To mimic the real-world networks that may consist of a few large but more small clusters, the cluster sizes n_k are generated from the following power law distribution

$$\mathbb{P}(n_k = x) = \frac{x^{-a}}{\zeta(a, x_{\min})}$$

where $\zeta(a, x_{\min}) = \sum_{n=0}^{\infty}(n + x_{\min})^{-a}$ is the Hurwitz zeta function. The parameters $x_{\min} = 10$ and $a = 2.8$ are used. To generate the clusters, let p_k be i.i.d. $U(0.3, 0.5)$ representing the probability that two nodes in cluster k are connected. Then, the cluster k is generated from $G(n_k, p_k)$, so that different clusters have different values of the average within-cluster node degree. Moreover, the nodes from cluster k_1 and cluster k_2 are connected with probability q_{k_1,k_2}, where q_{k_1,k_2} are i.i.d. $U(0, (K \max_k n_k)^{-1})$ for $1 \le k_1 \ne k_2 \le K$. Therefore, the clusters in the hypothetical networks are sparsely connected. Therefore, the probabilities connecting two clusters are restricted and thus the sizes of outer nodes are not relatively large.

For simplicity, the label of the clusters are assumed known and are directly used in design. Consider $K \in \{20, 50\}$ and 100 networks are generated according to the same parameter setting for each value of K. For each of these networks, 1000 simulations are conducted.

D.2 The MIT phone call network in Section 6.2

The MIT phone call network has 6819 nodes and 82 clusters, which can be found from the Network Data Repository (Rossi and Ahmed, 2015). To see that the weights used for the HTEs may affect the evaluation of the

ATEs, we evaluate the number of the nodes within different ranges of $\pi_{k,i}(z)$ in Table 5. Although the majority of the nodes in this network are the inner nodes, i.e., $\pi_{k,i}(z) = 0.5$, the values of $\pi_{k,i}(z)$ for the outer nodes all less than 0.3 and thus the weights of the HTEs may affect the estimation of the ATEs.

Table 5: Number of nodes within intervals of $\pi_{k,i}(z)$.

$\pi_{k,i}(z)$	CR	PSR	ReAR
$[0, 0.1)$	80	84	90
$[0.1, 0.2)$	92	88	84
$[0.2, 0.3]$	325	325	323
$\{0.5\}$	6322	6322	6322

D.3 Numerical Study for the Facebook Page to Page Network

As suggested by Theorem 3.1, the valid estimation of the ATEs requires that the network should not have too many outer nodes. In practice, the clusters in real-world networks may not be sparsely connected, and thus have several outer nodes. The Facebook page to page network is used to show that even when the clusters are not sparsely connected, our proposed method can still be useful. This network is listed in the stanford large network dataset collection (Leskovec and Sosič, 2016), which can be found at `https://snap.stanford.edu`.

This web-graph is a page to page graph of verified Facebook sites. Nodes represent the official Facebook pages while the links are mutual likes between sites. The company may consider to test whether a new feature might increase the user engagement on the webpages, i.e., the number of clicks or the number of views. Because such pages may belong to different categories and are built for different purposes, the improvement of the user engagement on different pages might be different. Therefore, an A/B test might be conducted to detect the difference of the average improvement on these pages. As such, the page can be considered as the unit used in the experiment and the response may correspond to the number of the clicks on the pages. Note that this network consists of 22,472 nodes and 170,824 edges. The fast greedy optimization of modularity algorithm Clauset et al. (2004) is used to generate the labels of the clusters. Some of the clusters are combined so that the percentage of inner nodes of the clusters all above 20%. The network thus has 184 clusters in total. Note that the fraction of the inner nodes for this network is only 64.1% and 100 of the 184 clusters have the fraction of the inner nodes less than 60%. Therefore, this network structure poses certain difficulties for evaluating the ATE.

Table 6: The bias, standard deviation (SD), and MSE of the estimators evaluated for the Facebook pages to pages networks.

Design	CR			PSR			ReAR		
Estimators	Bias	SD	MSE	Bias	SD	MSE	Bias	SD	MSE
$\hat{\tau}_{HT,IN}$	1.026	5.250	28.620	4.188	7.199	69.369	0.985	1.695	3.845
$\hat{\tau}_{CAE,IN}$	-0.086	1.971	3.892	-0.067	1.028	1.062	-0.071	0.528	0.283
$\hat{\tau}_{HT,CL}$	0.335	1.298	1.796	3.537	7.428	67.689	0.247	1.158	1.401
$\hat{\tau}_{CAE,CL}$	-0.235	0.118	0.069	-0.236	0.097	0.065	-0.233	0.091	0.063

As shown in Table 6, the HTEs fails to work well but the CAEs can provide better performance in terms of bias and standard deviation. Note that the ATEs for this network are $\tau_{IN}(\mathbf{1}, \mathbf{0}) = 1$ and $\tau_{CL}(\mathbf{1}, \mathbf{0}) = 0.93$, the HTEs under the three randomization procedures are all biased. Furthermore, the standard deviations of the HTEs are also quite large resulting in large values of the MSE. On the other hand, the performance of the CAEs are less affected by the number outer nodes, as the weights used by the CAEs can be adjusted according to the cluster structures. As such, the standard deviations of the CAEs are smaller than the corresponding values of the HTEs. Note that $\hat{\tau}_{CAE,CL}$ still has relatively large bias due to the fractions of the outer nodes in the clusters, where as $\hat{\tau}_{CAE,IN}$ is much less biased.

In addition, Table 6 indicates that our proposed adaptive A/B test procedure can be a useful tool for evaluating the ATE. The standard deviation of $\hat{\tau}_{CAE,IN}$ under ReAR is 0.528, which is much smaller than the values under

complete randomization and the pairwise-sequential randomization. The bias of $\hat{\tau}_{CAE,IN}$ under ReAR is also moderate. Therefore, the MSE of $\hat{\tau}_{CAE,IN}$ under ReAR is the smallest and has the best performance.

E Code

The R codes used for our study is available at `https://github.com/LouisLiu-STAT/Cluster_Adaptive_AB_test`.