Locally Optimal Design for A/B Tests in the Presence of Covariates and Network Dependence

[1], Qiong Zhang*[1], and Lulu Kang†[2]

[1]School of Mathematical and Statistical Sciences, Clemson University
[2]Department of Applied Mathematics, Illinois Institute of Technology

Abstract

A/B test, a simple type of controlled experiment, refers to the statistical procedure of conducting an experiment to compare two treatments applied to test subjects. For example, many IT companies frequently conduct A/B tests on their users who are connected and form social networks. Often, the users' responses could be related to the network connection. In this paper, we assume that the users, or the test subjects of the experiments, are connected on an undirected network, and the responses of two connected users are correlated. We include the treatment assignment, covariate features, and network connection in a conditional autoregressive model. Based on this model, we propose a design criterion that measures the variance of the estimated treatment effect and allocate the treatment settings to the test subjects by minimizing the criterion. Since the design criterion depends on an unknown network correlation parameter, we adopt the locally optimal design method and develop a hybrid optimization approach to obtain the optimal design. Through synthetic and real social network examples, we demonstrate the value of including network dependence in designing A/B experiments and validate that the proposed locally optimal design is robust to the choices of parameters.

Keywords: A/B test; Conditional autoregressive model; Controlled experiments; Covariates; Optimal design.

1 Introduction

A/B or A/B/n test, a simple type of controlled experiment, refers to the procedure of comparing the outcomes of two or more treatment settings from a finite number of test subjects. In the literature, controlled experiments have been widely used in agricultural, clinical trials, engineering and science studies, marketing research, etc (Atkinson and Bailey, 2001). Due to the advent of Internet technologies, large-scale A/B test has been commonly used by technology companies such as Amazon, Facebook, LinkedIn, Netflix, etc., to compare different versions of algorithms, web designs, and other online products and services. For example, Nandy et al. (2020) showed a case study on the LinkedIn newsfeed, which is a content recommender system with hundreds of millions of users. A recommender system is referred to as the infrastructure that provides a personalized recommendation on products or services based on users' personal information or past behaviors (Kohavi et al., 2020). The accuracy of the recommender algorithm is crucial to the quality and/or

*qiongz@clemson.edu
†lkang2@iit.edu

profit of these companies. A practical problem is to decide if an innovative update should be made to the algorithm in use. Therefore, an A/B test (i.e., "A" refers to the updated algorithm and "B" the current one) is used to make a comparison of the two and make the decision. To make a robust comparison of the two algorithms, the experiment should last for a certain period to make sure that users can receive enough exposure to the updated recommender system. The outcome of each user can be the total time spent on the recommended products/service and the click-through rate to the recommended products/service.

In its simplest form, the experimenter wants to compare the outcomes of two different treatments, labeled by A and B. A completely randomized design is commonly used, in which the treatment setting is randomly assigned to different test subjects. The randomization leads to unbiased estimates of certain estimands, typically, the average treatment (or causal) effect (Rubin, 2005), under minimum assumptions. However, there is still room for improvement in the efficiency of the A/B test procedure when certain practical challenges are involved. Besides the treatment setting, many other variables can affect a user's outcome, including the covariates information and social network connection of the user. Covariates, such as users' demographic, educational, financial information, are usually available to the experimenter and can significantly contribute to the behaviors and opinions of users. In the aforementioned scenarios, the experimenter also possesses the network connections of the users. In Section 7, we simulate an A/B experiment for the music recommender system based on the real dataset collected from the music streaming service Deezer. The data contains the friendship network of users and their covariates information regarding preferences to different music genres, which should be highly influential to the music recommender system. Intuitively, the outcomes of two connected users might be correlated to some degree. This intuition is reflected in the model assumption of the outcome regarding the network structure, and referred to as the network-correlated outcomes in Basse and Airoldi (2018b). In Section 2 and 3, we explain in details the assumption on the effects of the network to a user's outcome.

The rest of the paper is arranged as follows. In Section 2 we highlight some relevant existing works and point out the differences between the proposed method and the existing ones. Section 3 introduces the regression model including both the covariates and the correlation between users due to network connections. Based on this model, in Section 4, we propose a locally optimal design method in which the network correlation parameter is set to be the mean of its prior distribution. In Section 5, a hybrid approach is proposed to solve the optimization problem to obtain the optimal design. Through numerical experiments in Section 6 and 7, we demonstrate the benefit of the proposed approach. We conclude the paper in Section 8 with some discussion of the limitation of the proposed method and some future research directions.

2 Previous Work and Our Contribution

2.1 Existing Literature

For A/B tests that only involve covariates but not networks, most existing works advocate the necessity of covariate balancing between the treatment groups (Morgan and Rubin, 2012; Rubin, 2005; Morgan and Rubin, 2015; Bertsimas et al., 2015; Kallus, 2018; Li et al., 2021). For controlled experiments on networks, both theoretical and methodological works have been developed. See Gui et al. (2015); Phan and Airoldi (2015); Eckles et al. (2016); Basse and Airoldi (2018a), etc. Among them, Gui et al. (2015) proposed an estimator of average treatment effect considering the interference between users on the network and a randomized balance graph partition to assign treatments to each of the subnetworks. Eckles et al. (2016) used a graph cluster randomization to reduce the bias of the average treatment effect estimate. Nandy et al. (2020) proposed the strategy

to first apply approximate randomized controlled experiments solved by optimization and then use importance sampling to correct bias. Although focusing on networks, these works do not consider covariates.

In causal inference literature, the potential outcome framework is usually used. The average treatment effect is the target parameter for estimation and inference (Imbens and Rubin, 2015). Under this setup, many causal inference works do not require any probabilistic model assumption on the response variable. Alternatively, some recent works on the design for A/B experiments have operated under specific parametric model assumptions of the response variable, and optimal design idea is used to propose new design methods. For example, Bhat et al. (2020) developed off-line and online mathematical programming approaches to solve this optimization problem, the objective function of which is exactly the D_s-optimal design criterion (Kiefer, 1961; Atkinson and Donev, 1992). In this case, the D_s-optimality criterion minimizes the variance of the treatment effect of a parametric linear model. Optimal design strategies have also been used under the assumption of the network-correlated outcome, such as Basse and Airoldi (2018b) and Pokhilko et al. (2019). Outside the A/B test literature, there have been papers considering the optimal design problem with dependence between test subjects. For example, Martin (1986) considered the restricted randomized design when the test subjects are spatially correlated. Parker et al. (2017) and Koutra (2017) considered the optimal design under linear network effects.

Similar to the aforementioned works, we also opt for the optimal design direction as indicated by the title of this paper. We argue that although the nonparametric potential outcome framework has an important theoretical basis, the reasonable model-assisted design approaches are not meritless. Even in the works based on the potential outcome framework, certain linear model assumptions are also used in both theoretical and numerical proofs to show the advantages and properties of the balancing criteria and the design approaches. For example, Morgan and Rubin (2012) assumed an additive linear model to show how much variance reduction can be obtained by rerandomization using Mahalanobis distance. Gui et al. (2015) used a linear additive model in terms of treatment effect, neighboring covariates, and neighboring responses as the rationale to create the sample estimator of average treatment effect, as well as to simulate data in numerical experiments.

2.2 Differences and New Contributions

In this paper, we develop an optimal design approach for A/B experiments in the presence of both covariates and network connections. The scope of the paper targets the social networks of users whose covariates information are influential to their reactions to the treatments. With a parametric conditional autoregressive (CAR) model that assumes the outcome is the sum of treatment effect, covariate effects, and correlated residuals for capturing network dependence, we focus on the estimation of the treatment effect parameter. Based on this model, we develop an optimal design criterion such that the variance of the estimated treatment effect is minimized. By design, we mean the assignment of treatment settings to each test subject in the context of this paper. We focus on the simplest case where the experiment only involves two treatments, A and B. But the proposed modeling and design method can be extended to the case of multiple treatment settings, as discussed in Section 8. The design of the treatment settings for multiple experimental factors is not the focus of this paper.

The resulting design criterion in Section 3 depends on the network structure, the covariates, and an unknown network correlation parameter, and it can not be simply expressed as a sparse quadratic function of the design variables, which is different from Pokhilko et al. (2019). Therefore, the mathematical formulation developed by Pokhilko et al. (2019) is infeasible to solve this new optimal design problem.

We also assume the common Stable Unit Treatment Value Assumption (SUTVA) (Rubin, 1974), which states that the outcome of a test subject is unaffected by the treatment assignments of any other subjects. In other words, we do not think there is any direct interference from the neighbors' treatment settings to the focused test subject's outcome. This assumption is appropriate for many applications where users are unawarely participating in the experiments run by the online service providers. Users' outcomes can still be correlated due to their network connections and covariates information.

This non-interference assumption is different from the interference assumption in some existing works, such as Parker et al. (2017). In Parker et al. (2017), the proposed model includes the treatment assignments of connected subjects as linear predictors in their model. The experimental outcome of a subject under this model is affected by the treatment assignments of connected subjects. Different from this assumption, we assume that the experimental outcomes are correlated due to the network connection between subjects, which is characterized by the error term of the CAR model. However, the experimental outcomes are unaffected by the treatment assignments of connected subjects. Therefore, the proposed model of this paper is not comparable with the one in Parker et al. (2017) due to the different assumptions. Both can be useful under suitable scenarios and assumptions.

3 Optimal Design with Network Connection

Consider n test subjects participating in the experiment. For the i-th subject, let $x_i \in \{-1, 1\}$ represent the experimental allocation of A or B treatment, $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{ip})^\top$ be the p-dimensional covariates, and y_i be the experimental outcome. Assume that the outcome y_i is a continuous random variable. Bhat et al. (2020) models the relationship between y_i and the effects of the treatment and covariates as

$$y_i = x_i\theta + \boldsymbol{f}_i^\top \boldsymbol{\beta} + \delta_i \quad \text{for} \quad i = 1, \ldots, n, \tag{1}$$

Here $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$ is the vector of the linear coefficients for $\boldsymbol{f}_i = (1, \boldsymbol{z}_i^\top)^\top$. We name θ as the treatment effect. Note that it is different from the notion of average treatment effect which is the usual estimand in the potential outcome framework. The model in (1) does not involve the network and the error terms δ_i's are assumed to be independent and identically distributed (iid) normal random variables with mean zero and a constant variance σ^2. The purpose of design allocation is to reduce the variance of the least square estimator, which is unbiased if assumption (1) stands. According to Bhat et al. (2020), the variance of the least squares estimator $\hat{\theta}$ from (1) is $\text{var}(\hat{\theta}) = \sigma^2[\boldsymbol{x}^\top(\boldsymbol{I}_n - \boldsymbol{F}(\boldsymbol{F}^\top\boldsymbol{F})^{-1}\boldsymbol{F}^\top)\boldsymbol{x}]^{-1}$, where $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$, $\boldsymbol{F}^\top = (\boldsymbol{f}_1, \ldots, \boldsymbol{f}_n)$ and $\boldsymbol{I}_n$ is the identity matrix of size n. Therefore, the optimal design is obtained by minimizing $\text{var}(\hat{\theta})$, which is equivalent to

$$\begin{aligned} &\min\ \boldsymbol{x}^\top\boldsymbol{F}(\boldsymbol{F}^\top\boldsymbol{F})^{-1}\boldsymbol{F}^\top\boldsymbol{x} \\ &\text{s.t.}\ \ -1 \le \sum_{i=1}^{n} x_i \le 1, \quad \boldsymbol{x} \in \{-1, 1\}^n. \end{aligned} \tag{2}$$

The constraint $-1 \le \sum_{i=1}^{n} x_i \le 1$ is imposed to make sure that the numbers of test subjects assigned to 1 and -1 are equal or within the difference of 1, which corresponds to the situation of even or odd sample size n.

Next, we extend the linear model (1) to the case with the network connection. We require that the network between subjects is known to the experimenter just like the covariate information. Assume this network form a simple undirected graph with nodes representing the test subjects. If

two test subjects are connected, there is a single edge between the two corresponding nodes. Such a network can be represented by an $n \times n$ adjacency matrix $\boldsymbol{W}$, or incidence matrix. Its diagonal entries are 0's, whereas off-diagonal entries ($i \neq j$) are

$$w_{ij} = \begin{cases} 1, & \text{if node } i \text{ and node } j \text{ are adjacent or connected} \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

Obviously, $\boldsymbol{W}$ is symmetric. We denote the number of adjacent neighbors, or degree, of the i-th node as $m_i = \sum_{j=1}^{n} w_{ij}$, and $m = \sum_{i=1}^{n} m_i$ is twice of the total number of edges in this graph.

To add the network's influence into the linear additive model (1), we propose the conditional autoregressive or CAR distribution (Cressie, 1993; Rue and Held, 2005; Banerjee et al., 2014) for δ_i's to represent the network dependence between the connected test subjects. According to the CAR model,

$$\delta_i | \delta_1, \ldots, \delta_{i-1}, \delta_{i+1}, \ldots, \delta_n \sim N\left(\rho \sum_{j \neq i} \frac{w_{ij}\delta_j}{m_i}, \frac{\sigma^2}{m_i}\right), \tag{4}$$

where σ^2 is the variance and $0 \leq \rho < 1$ is a correlation parameter characterizing the strength of network dependence. Equivalently, $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)^\top$ follows a multivariate normal distribution

$$\boldsymbol{\delta} \sim \mathcal{MVN}_n(0, \sigma^2 \boldsymbol{R}^{-1}(\rho, \boldsymbol{W})), \tag{5}$$

where $\boldsymbol{R}(\rho, \boldsymbol{W}) = (\boldsymbol{D} - \rho \boldsymbol{W})$ with $\boldsymbol{D} = \text{diag}\{m_1, \ldots, m_n\}$. The matrix $\boldsymbol{R}(\rho, \boldsymbol{W})$ is positive definite when $0 \leq \rho < 1$ and $m_i \geq 1$ for $i = 1, \ldots, n$ (Ver Hoef et al., 2018). The proof of the equivalence of (4) and (5) is given by Besag (1974) and illustrated by Pokhilko et al. (2019) under the framework of network A/B test.

With the CAR model assumption, the outcome y_i depends on the network connection for the i-th subject but does not depend on the treatment allocation of the connected subjects. In this way, the outcome of the subject is mainly decided by him/herself and the treatment he/she receives. Since most social networks are built on positive connections between users, we assume the influence from the network is synergistic to users and thus ρ is positive. When $\rho = 0$, the model assumption returns to the linear model (1) which does not involve network.

If the network correlation parameter ρ is known and $0 \leq \rho < 1$, the variance of the least squares estimator $\hat{\theta}$ can be expressed by

$$\text{var}(\hat{\theta}) = \sigma^2 \left[\boldsymbol{x}^\top \boldsymbol{K} \boldsymbol{x}\right]^{-1}, \tag{6}$$

where $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$ is the treatment assignments for all subjects, and $\boldsymbol{K}$ is an $n \times n$ matrix

$$\boldsymbol{K} = (\boldsymbol{D} - \rho \boldsymbol{W}) - (\boldsymbol{D} - \rho \boldsymbol{W})\boldsymbol{F}\left[\boldsymbol{F}^\top (\boldsymbol{D} - \rho \boldsymbol{W})\boldsymbol{F}\right]^{-1} \boldsymbol{F}^\top (\boldsymbol{D} - \rho \boldsymbol{W}), \tag{7}$$

with the covariates matrix $\boldsymbol{F}^\top = (\boldsymbol{f}_1, \ldots, \boldsymbol{f}_n)$. Here, to make $\boldsymbol{F}^\top (\boldsymbol{D} - \rho \boldsymbol{W})\boldsymbol{F}$ invertible, we require $\boldsymbol{F}$ to be full-rank, i.e., $\text{rank}(\boldsymbol{F}) = p + 1$. For the same reason, we do not consider the isolated node whose degree $m_i = 0$. In Section 7, we explain how to deal with the isolated nodes if they exist. The optimal design $\boldsymbol{x}$ minimizes the variance of estimated treatment effect in (6). This optimal design is also known as a D_s-optimal design in the optimal design literature (Kiefer, 1961; Atkinson and Donev, 1992). Equivalently, we express the optimal design as the solution of

$$\begin{aligned} &\max\ T(\boldsymbol{x}, \rho) := \boldsymbol{x}^\top \boldsymbol{K} \boldsymbol{x}, \\ &\text{s.t.}\ \ -1 \leq \sum_{i=1}^{n} x_i \leq 1, \ \text{and}\ \boldsymbol{x} \in \{-1, 1\}^n, \end{aligned} \tag{8}$$

which maximizes the precision (as the inverse of variance) of the estimated treatment effect. Since

$$\boldsymbol{x}^\top \boldsymbol{K} \boldsymbol{x} = \boldsymbol{x}^\top (\boldsymbol{D} - \rho \boldsymbol{W}) \boldsymbol{x} - \boldsymbol{x}^\top (\boldsymbol{D} - \rho \boldsymbol{W}) \boldsymbol{F} \left[\boldsymbol{F}^\top (\boldsymbol{D} - \rho \boldsymbol{W}) \boldsymbol{F} \right]^{-1} \boldsymbol{F}^\top (\boldsymbol{D} - \rho \boldsymbol{W}) \boldsymbol{x}, \tag{9}$$

and $\boldsymbol{x}^\top (\boldsymbol{D} - \rho \boldsymbol{W}) \boldsymbol{x} = m - \rho \boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{x}$, we have that

$$T(\boldsymbol{x}, \rho) = m - T_1(\boldsymbol{x}, \rho) - T_2(\boldsymbol{x}, \rho), \tag{10}$$

where

$$T_1(\boldsymbol{x}, \rho) = \rho \boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{x} = \rho \sum_{ij} w_{ij} x_i x_j,$$
$$T_2(\boldsymbol{x}, \rho) = \boldsymbol{x}^\top (\boldsymbol{D} - \rho \boldsymbol{W}) \boldsymbol{F} \left[\boldsymbol{F}^\top (\boldsymbol{D} - \rho \boldsymbol{W}) \boldsymbol{F} \right]^{-1} \boldsymbol{F}^\top (\boldsymbol{D} - \rho \boldsymbol{W}) \boldsymbol{x}.$$

Note that minimizing $T_1(\boldsymbol{x}, \rho)$ would push x_i and x_j to be assigned with different treatments whenever $w_{ij} = 1$. To facilitate the discussion, we name this condition "connection balance", meaning that the two connected subjects are assigned with different treatment settings. Intuitively, this is a meaningful condition since two connected test subjects are usually similar in many aspects of their background. Thus, the most likely factor contributing to their difference in outcomes is the treatment setting. This condition is consistent with the optimal design for the A/B test without covariates in Pokhilko et al. (2019). In the extremely simple and artificial case illustrated later in Figure 2 in Section 5, such perfect balance can be achieved. For real networks, the connection balance can only be achieved to a certain degree but rarely perfectly. Also, $T_2(\boldsymbol{x}, \rho)$ can be viewed as a network re-weighted Mahalanobis distance in Morgan and Rubin (2012), since it can be expressed by

$$T_2(\boldsymbol{x}, \rho) = \boldsymbol{x}^\top (\boldsymbol{D} - \rho \boldsymbol{W}) \boldsymbol{F} \boldsymbol{\Sigma}_n^{-1} \boldsymbol{F}^\top (\boldsymbol{D} - \rho \boldsymbol{W}) \boldsymbol{x},$$

with $\boldsymbol{\Sigma}_n = \boldsymbol{F}^\top (\boldsymbol{D} - \rho \boldsymbol{W}) \boldsymbol{F}$. Therefore, the objective in (10) contains $T_1(\boldsymbol{x}, \rho)$ to achieve connection balance, and $T_2(\boldsymbol{x}, \rho)$ to achieve covariate balance. One critical issue is that the optimality criterion depends on the value of ρ. In practice, ρ is an unknown parameter. Next, we are going to discuss the choice of ρ.

4 Locally Optimal Design

The optimal design criterion $T(\boldsymbol{x}, \rho)$ depends on the network correlation parameter ρ, which is usually unknown before experiments. We can use Bayesian optimal design to handle the uncertainty of the unknown parameters. Using its most common formulation, we should optimize the expectation of the design criterion, i.e., $\mathbb{E}_\rho[T(\boldsymbol{x}, \rho)]$, with respect to a user-specified prior distribution of the parameter ρ. But even with the simple uniform prior for ρ, the expectation does not have a tractable form. Many numerical methods, such as quadrature, Quasi-Monte Carlo, Markov Chain Monte Carlo, etc., have to be used to compute the integration. Please see Ryan et al. (2014), Ryan et al. (2016), and Drovandi and Tran (2018) for more comprehensive review on the advanced computational methods on Bayesian optimal designs.

To simplify the computation, we investigate the property of $T(\boldsymbol{x}, \rho)$ with respect to ρ to find an analytic surrogate of $\mathbb{E}[T(\boldsymbol{x}, \rho)]$. We first discover the concavity of $T(\boldsymbol{x}, \rho)$ with respect to ρ in Theorem 1. Based on Jenson's Inequality, the conclusion in Corollary 1 holds directly. The proof of Theorem 1 is provided in the Supplement. Based on the two results, we propose to use $T(\boldsymbol{x}, \rho_0)$, the upper bound of $\mathbb{E}[T(\boldsymbol{x}, \rho)]$, as the surrogate of the objective to obtain design allocation.

Theorem 1. *For $\rho \in (0,1)$, and any given design $\boldsymbol{x}$, the design criterion $T(\boldsymbol{x},\rho)$ is a concave function with respect to ρ.*

Corollary 1. *Given a prior distribution of ρ, $p(\rho)$, for $\rho \in (0,1)$, a tight upper bound for $\mathbb{E}\left[T(\boldsymbol{x},\rho)\right]$ is $\mathbb{E}\left[T(\boldsymbol{x},\rho)\right] \leq T(\boldsymbol{x},\rho_0)$, where $\rho_0 := \mathbb{E}(\rho)$ is the population mean of ρ based on $p(\rho)$.*

We define the locally optimal design by solving

$$\begin{aligned} &\max\ T(\boldsymbol{x},\rho_0) \\ &\text{s.t.}\ \ -1 \leq \sum_{i=1}^{n} x_i \leq 1,\ \text{and}\ \boldsymbol{x} \in \{-1,1\}^n, \end{aligned} \tag{11}$$

whose objective function is equivalent to the original objective in (10) with ρ specified as the mean of the prior distribution. Using a specific ρ_0 in the design criterion to obtain the optimal design is known as the locally optimal design (Chaloner and Verdinelli, 1995).

The quality of the design based on the surrogate problem in (11) can be investigated from two aspects. First, we provide the analytic gap between $T(\boldsymbol{x},\rho_0)$ and $\mathbb{E}[T(\boldsymbol{x},\rho)]$ and a simulation example to illustrate the typical range of the gap between the surrogate local design criterion $T(\boldsymbol{x},\rho_0)$ and the global criterion $\mathbb{E}[T(\boldsymbol{x},\rho)]$. Proposition S1 of the analytic gap and simulation results in Figure S1 are given in the Supplement.

Next, we investigate whether the surrogate design criterion $T(\boldsymbol{x},\rho_0)$ is robust to the choice of ρ_0. To do so, we check of the correlation between any $T(\boldsymbol{x},\rho_0)$ and $T(\boldsymbol{x},\rho)$ for a pair of fixed (ρ_0,ρ) for any randomly generated design $\boldsymbol{x}$. If the correlation between $T(\boldsymbol{x},\rho_0)$ and $T(\boldsymbol{x},\rho)$ is large and positive, it indicates that a design resulting in large $T(\boldsymbol{x},\rho_0)$ is also likely to lead to large $T(\boldsymbol{x},\rho)$. In Proposition S2, we have given the formula to calculate the exact correlation $\text{cor}_{\boldsymbol{x}}(T(\boldsymbol{x},\rho_0),T(\boldsymbol{x},\rho))$ for all the completely randomized design in which x_i's are i.i.d. random variables and $\Pr(x_i = 1) = \Pr(x_i = -1) = 0.5$. To visualize the correlation, we also provide a simulated example using a network with 50 nodes and five-dimensional covariates associated with each node. The edges of the network are generated as independent Bernoulli random variables with probability 0.08. The covariates are generated as independent random variables taking values from $\{-1,1\}$ with equal probabilities. The values of ρ_0 and ρ are set to be 0.1, 0.3, 0.5, 0.7, and 0.9 and omit the case when $\rho_0 = \rho$. For each pair of (ρ_0,ρ) values, we generate 1000 completely randomized designs and compute the corresponding $T(\boldsymbol{x},\rho_0)$ and $T(\boldsymbol{x},\rho)$ for each design. Figure 1 returns the scatter plot of $T(\boldsymbol{x},\rho_0)$ and $T(\boldsymbol{x},\rho)$ for the 1000 completely randomized design for different (ρ_0,ρ) values. It shows that $T(\boldsymbol{x},\rho_0)$ and $T(\boldsymbol{x},\rho)$ are strongly linearly correlated. Also, the *exact* correlation values based on Proposition S2 ranges from 0.75-0.99 for values of (ρ_0,ρ) in the simulation. From these results, it is safe to say that the locally optimal design is robust to the choice of ρ_0 value. Particularly, for $\rho_0 = 0.5$, the correlation values between $T(\boldsymbol{x},\rho_0)$ and $T(\boldsymbol{x},\rho)$ where $\rho = 0.1, 0.3, 0.7, 0.9$ are all above 0.9. Therefore, the optimal design obtained based on $\rho_0 = 0.5$ is the most robust for the model with the true value of $\rho \in (0.1, 0.9)$.

Although using the local design criterion $T(\boldsymbol{x},\rho_0)$ is a simple solution, the quality of the resulting design can be validated. Simulation examples in Section 6 can further demonstrate that the performance of the locally optimal design is equally good as the *true* optimal design in which parameter ρ is set to be its true known value.

5 A Hybrid Solution Approach to Obtain Optimal Design

Since the optimal design in (11) is the integer solution of the maximum of a quadratic form, obtaining the exact solution of such problem is challenging (Belotti et al., 2013; Bhat et al., 2020).

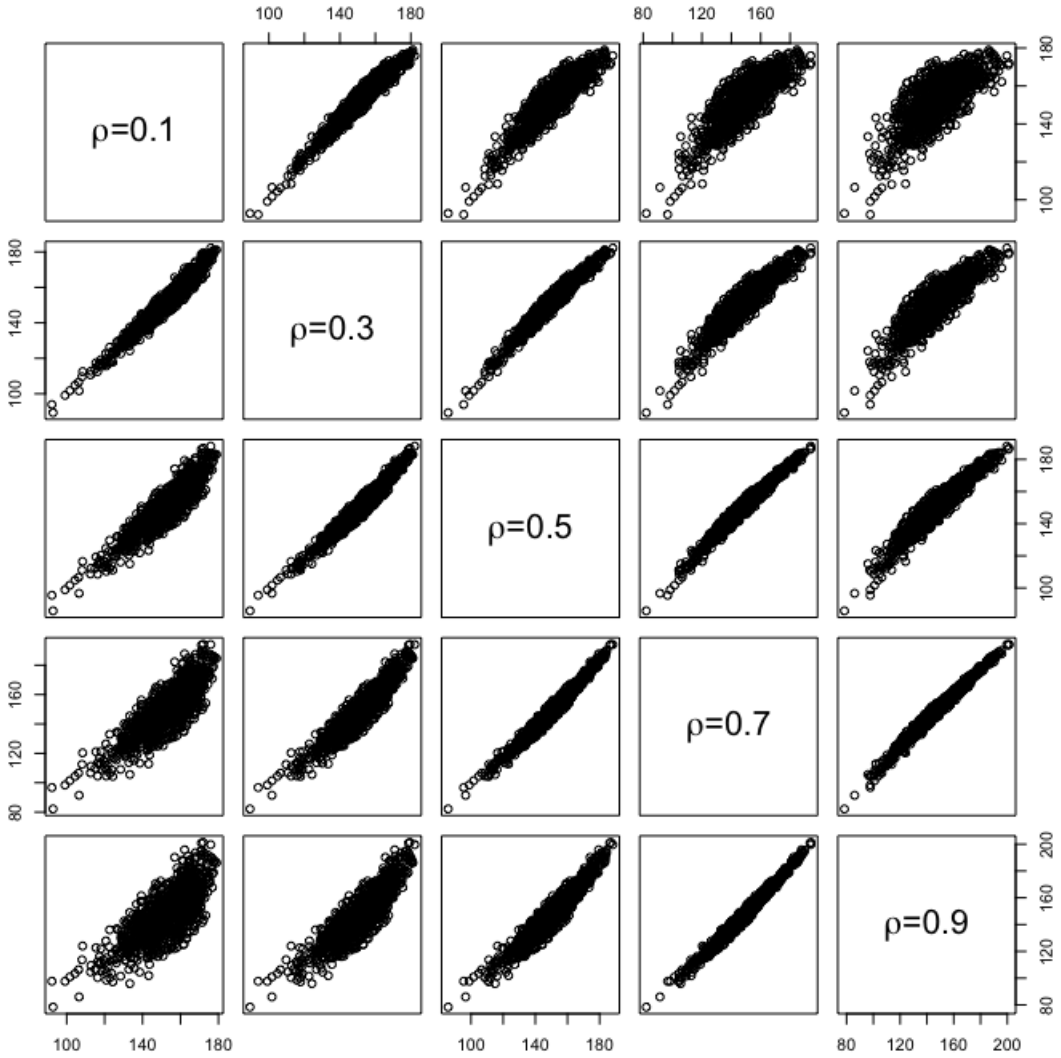

Figure 1: Scatter plot of $T(\boldsymbol{x}, \rho_0)$ and $T(\boldsymbol{x}, \rho)$ of each pair of (ρ_0, ρ) with 1000 randomly generated designs.

According to (10) and (11), the maximization can be converted to minimization of $T_1(\boldsymbol{x}, \rho_0) + T_2(\boldsymbol{x}, \rho_0)$, where

$$T_1(\boldsymbol{x}, \rho_0) = \rho_0 \boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{x}$$
$$T_2(\boldsymbol{x}, \rho_0) = \boldsymbol{x}^\top (\boldsymbol{D} - \rho_0 \boldsymbol{W}) \boldsymbol{F} \left[\boldsymbol{F}^\top (\boldsymbol{D} - \rho_0 \boldsymbol{W}) \boldsymbol{F} \right]^{-1} \boldsymbol{F}^\top (\boldsymbol{D} - \rho_0 \boldsymbol{W}) \boldsymbol{x},$$

and

$$T_1(\boldsymbol{x}, \rho_0) + T_2(\boldsymbol{x}, \rho_0) = \boldsymbol{x}^\top \left[\rho_0 \boldsymbol{W} + (\boldsymbol{D} - \rho_0 \boldsymbol{W}) \boldsymbol{F} \left[\boldsymbol{F}^\top (\boldsymbol{D} - \rho_0 \boldsymbol{W}) \boldsymbol{F} \right]^{-1} \boldsymbol{F}^\top (\boldsymbol{D} - \rho_0 \boldsymbol{W}) \right] \boldsymbol{x}.$$

The matrix

$$\rho_0 \boldsymbol{W} + (\boldsymbol{D} - \rho_0 \boldsymbol{W}) \boldsymbol{F} \left[\boldsymbol{F}^\top (\boldsymbol{D} - \rho_0 \boldsymbol{W}) \boldsymbol{F} \right]^{-1} \boldsymbol{F}^\top (\boldsymbol{D} - \rho_0 \boldsymbol{W})$$

is not necessarily a positive semi-definite matrix. As a result, the minimization problem of $T_1(\boldsymbol{x}, \rho_0) + T_2(\boldsymbol{x}, \rho_0)$ can not be solved directly as the problem in (2). We develop a hybrid solution approach to resolve this issue.

Notice that the matrix $(\boldsymbol{D} - \rho_0 \boldsymbol{W}) \boldsymbol{F} \left[\boldsymbol{F}^\top (\boldsymbol{D} - \rho_0 \boldsymbol{W}) \boldsymbol{F} \right]^{-1} \boldsymbol{F}^\top (\boldsymbol{D} - \rho_0 \boldsymbol{W})$ in $T_2(\boldsymbol{x}, \rho_0)$ is positive definite, and thus the minimization of $T_2(\boldsymbol{x}, \rho)$ can be solved by an outer-approximation based branch-and-cut algorithm as in (2). Hence, we reformulate the minimization of $T_1(\boldsymbol{x}, \rho_0) + T_2(\boldsymbol{x}, \rho_0)$ to be

$$\begin{aligned} &\min\ T_2(\boldsymbol{x}, \rho_0) \\ &\text{s.t. } T_1(\boldsymbol{x}, \rho_0) \leq q, \\ &-1 \leq \sum_{i=1}^{n} x_i \leq 1, \text{ and } \boldsymbol{x} \in \{-1, 1\}^n, \end{aligned} \tag{12}$$

which uses the constraint $T_1(\boldsymbol{x}, \rho_0) \leq q$ to control the value of $T_1(\boldsymbol{x}, \rho_0)$ to be small enough. Since ρ_0 is only a constant multiplier in $T_1(\boldsymbol{x}, \rho_0)$, this constraint can be reduced to $\boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{x} \leq q$, and it is critical to specify the value of q.

The value of $\boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{x}$ greatly depends on the number of subjects and the structure of the network. Therefore, the cap q should be related to a specific network. In Theorem 2, we investigate the asymptotic behaviors of $\boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{x}$ via random allocation with equal probability to decide a viable way to specify the value of q. Assume that the entire network contains unlimited users with a deterministic network structure, and the experiments are conducted on a subset of n users from the entire network. Therefore, the design vector $\boldsymbol{x}$ is the only random component that causes the stochastic behavior of the statistic $\boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{x}$. The proof of Theorem 2 is provided in the Supplement.

Theorem 2. *Consider that $x_1, \ldots, x_n$ in $\boldsymbol{x}$ are independent and identically distributed random variables from the discrete distribution with $\Pr(x_i = 1) = \Pr(x_i = -1) = 0.5$. As $n \to \infty$,*

$$\frac{\boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{x}}{\sqrt{m}} \xrightarrow{d} N(0, 1),$$

where $\xrightarrow{d}$ represents convergence in distribution and $m = \sum_{i,j} w_{ij}$.

Since $\boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{x} / \sqrt{m}$ asymptotically follows the standard normal distribution, we can specify the gap q according to the standard normal percentiles. Let z_α be the $100\alpha\%$ percentile of the

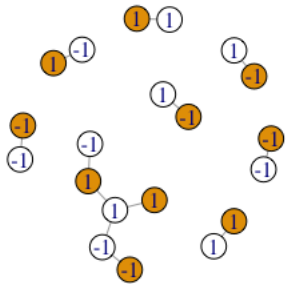

Figure 2: Visualization of the optimal design allocation. Two treatments are denoted by different colors. The covariate value 1 or -1 of each subject is labeled in each node.

standard normal distribution. If we specify a smaller value of α, the constraint is more restrictive. The optimization problem is reformulated as

$$\begin{aligned} \min \boldsymbol{x}^\top (\boldsymbol{D} - \rho_0 \boldsymbol{W}) \boldsymbol{F} \left[\boldsymbol{F}^\top (\boldsymbol{D} - \rho_0 \boldsymbol{W}) \boldsymbol{F} \right]^{-1} \boldsymbol{F}^\top (\boldsymbol{D} - \rho_0 \boldsymbol{W}) \boldsymbol{x} \qquad (13) \\ \text{s.t. } \boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{x} \leq \sqrt{m} z_\alpha, \\ -1 \leq \sum_{i=1}^{n} x_i \leq 1, \\ \boldsymbol{x} \in \{-1, 1\}^n. \end{aligned}$$

This minimization problem with a positive definite quadratic objective and two-level decision variables can be solved by off-the-shelf optimization solvers. Note that, the constraint $-1 \leq \sum_{i=1}^{n} x_i \leq 1$ is inserted to achieve a balanced allocation of two treatments. In terms of implementation, this constraint usually improves the computation cost since it also reduces the number of feasible solutions. The formulation in (13) changes the minimization of two objective functions $T_1(\boldsymbol{x}, \rho_0)$ and $T_2(\boldsymbol{x}, \rho_0)$ into the minimization of one and constraining the other, and thus the name of *hybrid solution* approach.

We provide an illustration of the proposed design using a simple bipartite network of 20 nodes. For simplicity, the covariate z_i is a one-dimensional vector taking value from $\{-1, 1\}$. We set the correlation parameter ρ_0 as 0.5 and parameter α as 0.001 for the proposed approach in (13). The locally optimal design is visualized in Figures 2. The simple bipartite network can be divided into two disjoint sets, and the treatment allocation is orthogonal to the covariate vector, which achieves perfect balance for the covariate and the network. However, perfect balancing may not be achievable for general cases, but small values of T_1 and T_2 can still provide a better-balanced structure of network connection and covariates, respectively.

We first discuss the choice of α in (13). The hybrid problem in (13) is generally computationally efficient to solve for networks with 100-5000 nodes. Therefore, it is feasible to obtain designs through conducting a sensitivity analysis with a series of decreasing α values and stop at the α value when the change of objective values $T(\boldsymbol{x}, \rho_0)$ in (11) is small, or the improvement of precision stops increasing as α decreases. One numerical example is used to demonstrate the performance of the design with respect to different choices of α's in Section 6.1.

At last, we remark on the choice of ρ_0 in this new formulation (13). Like $T(\boldsymbol{x}, \rho_0)$, the new objective function $T_2(\boldsymbol{x}, \rho_0)$ in (13) is also a quadratic form of $\boldsymbol{x}$. Therefore, Proposition S2 still holds for any correlation between $T_2(\boldsymbol{x}, \rho_0)$ and $T(\boldsymbol{x}, \rho)$ for any pair of (ρ_0, ρ). The quality of design with a given ρ_0 can be assessed on any possible true value of ρ using the analytic correlation between $T_2(\boldsymbol{x}, \rho_0)$ and $T_2(\boldsymbol{x}, \rho)$ similar to the discussion near the end of Section 4. Particularly, in

the special case without any covariates, i.e., $\boldsymbol{F} = \mathbf{1}_n$,

$$T_2(\boldsymbol{x}, \rho) = (1-\rho)\frac{(\boldsymbol{x}^\top \boldsymbol{m})^2}{\sum_{i=1}^n m_i},$$

where $\boldsymbol{m} = (m_1, \ldots, m_n)^\top$ with m_i be the number of adjacent neighbors of the i-th user. Therefore, $\text{cor}_{\boldsymbol{x}}(T_2(\boldsymbol{x}, \rho_0), T_2(\boldsymbol{x}, \rho)) = 1$ for any ρ_0 and ρ. It indicates that there is no loss to replace an unknown true ρ with a given ρ_0 in this special case.

6 Numerical Study

The purpose of optimal design is to reduce the variance (or equivalently, improve the precision) of the estimated treatment effect $\hat{\theta}$ in (1). Since the optimal value of the design criterion can not be obtained directly, computing the classical measure "design efficiency" is not feasible. Alternatively, we evaluate the quality of design by computing the percentage of the improvement in precision compared to the expected precision of random balanced designs.

Proposition 1. *Consider a random balanced design $\boldsymbol{x}$. The marginal distribution of each x_i is $\Pr(x_i = 1) = \Pr(x_i = -1) = 0.5$ and $\sum_{i=1}^n x_i$ follows the balance condition, i.e., $-1 \leq \sum_{i=1}^n x_i \leq 1$. The expected precision of the random balanced design is*

$$\mathbb{E}_{\boldsymbol{x}}\left(\sigma^{-2}\boldsymbol{x}^\top \boldsymbol{K}\boldsymbol{x}\right) = \sigma^{-2}\text{tr}(\boldsymbol{K}\boldsymbol{C}), \tag{14}$$

where $\boldsymbol{C}$ is an $n \times n$ matrix with all of the diagonal entries equal to 1 and all of the off-diagonal entries equal to a fixed constant c. The value of c is $-(n-1)^{-1}$ if n is even and it is $-n^{-1}$ if n is odd. Here $\text{tr}(\cdot)$ denotes the trace of a matrix. The expectation in (14) is taken with respect to the probability distribution of $\boldsymbol{x}$.

The proof of the above proposition is given in the Supplement. For any given design $\boldsymbol{x}_0$, the percentage of the improvement in precision with respect to the expected precision of the random balancedized design can be expressed by

$$\text{PIP}(\boldsymbol{x}_0) = \frac{\sigma^{-2}\boldsymbol{x}_0^\top \boldsymbol{K}\boldsymbol{x}_0 - \mathbb{E}_{\boldsymbol{x}}\left(\sigma^{-2}\boldsymbol{x}^\top \boldsymbol{K}\boldsymbol{x}\right)}{\sigma^{-2}\boldsymbol{x}_0^\top \boldsymbol{K}\boldsymbol{x}_0} = 1 - \frac{\text{tr}(\boldsymbol{K}\boldsymbol{C})}{\boldsymbol{x}_0^\top \boldsymbol{K}\boldsymbol{x}_0}. \tag{15}$$

For short, we denote this percentage of improvement in precision by $\text{PIP}(\boldsymbol{x}_0)$. According to (7), the calculation of matrix $\boldsymbol{K}$ involves the network correlation parameter ρ. Since (15) is used to evaluate the design $\boldsymbol{x}_0$, naturally, we should use the true value of the network correlation, denoted by ρ_t, to compute $\text{PIP}(\boldsymbol{x}_0)$. In the following simulation study, ρ_t is part of the simulation settings.

In Section 6.1, we evaluate the robustness of the proposed design approach to different choices of α and ρ_0. In Section 6.2, we evaluate the advantages of the optimal design with network connection under different scenarios. In both subsections, we generate synthetic datasets, where the edges of the network are independently generated from a Bernoulli distribution with a constant probability, which is called *network density*. If there are isolated nodes appear in the generated network, we connect each of them with a randomly selected neighbor to remove isolation and ensure that $m_i \geq 1$ in (4). Each node is associated with a p-dimensional covariates whose entries are randomly generated from $\{-1, 1\}$ with equal probabilities. To stabilize the results, we generate 10 copies of datasets and report the results in boxplots.

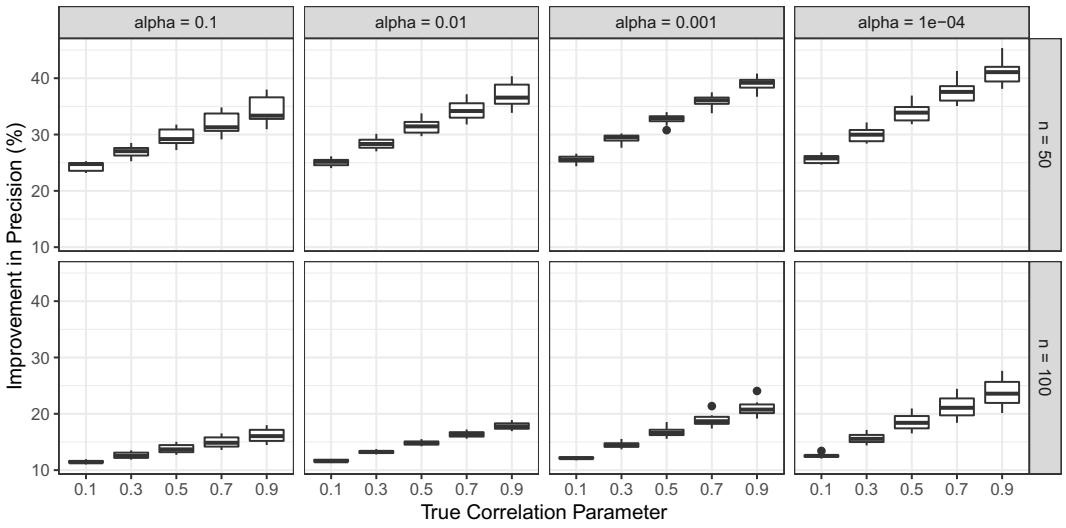

Figure 3: PIP($\boldsymbol{x}$) of the locally optimal designs with $p = 10$ and network density 0.08.

6.1 Robustness on the Choices of α and ρ_0

In this subsection, we consider two versions of the proposed hybrid design approach.

1. Locally optimal design: the optimal design obtained by solving the optimization problem in (11). We specify the mean of the prior distribution to be 0.5, i.e., $\rho_0 = 0.5$.

2. True optimal design: the optimal design obtained by maximizing the objective in (8) with the true network correlation value ρ_t.

We use the hybrid approach in (13) to obtain both the locally and true optimal designs. The comparison between the locally optimal design and the true optimal design shows the gap of replacing the true design criterion $T(\boldsymbol{x}, \rho_t)$ by its practical surrogate $T(\boldsymbol{x}, \rho_0)$. For both designs, we use Gurobi (Gurobi Optimization, 2015) to solve the optimization problem, and the run-time is limited to 500 seconds.

First, we evaluate the performance of the locally optimal design with $\rho_0 = 0.5$ with different choices of α. In this case, we fix $p = 10$ and the network density is 0.08. Each boxplot in Figure 3 shows the PIP($\boldsymbol{x}$) values of 10 datasets. We can detect a slightly bigger PIP($\boldsymbol{x}$) for smaller α values. However, this trend diminishes when $\alpha = 0.001$ and $\alpha = 0.0001$. Therefore, we set $\alpha = 0.001$ for the rest of the paper. As noted earlier, it is computationally efficient to compute the optimal design and PIP($\boldsymbol{x}$) value. In practice, it is possible to obtain the optimal design for a sequence of α values, and choose the one when further decreasing α does not increase PIP.

The second simulation is to evaluate the robustness of the locally optimal design to the choice of ρ_0. We fix $\alpha = 0.001$ and the network density be 0.08. As discussed in Section 4, it is expected that the difference between the locally optimal design and the true optimal design is small. Figure 4 confirms this. It shows the boxplots of the differences between PIP of the two designs for the same data. Each boxplot is based on 10 replications. According to Figure 4, the differences are mostly under 3%. Since the locally optimal design with different ρ_0 performs similarly to the true optimal design with ρ_t in terms of PIP, we use the locally optimal design with $\rho_0 = 0.5$ for the rest of this section.

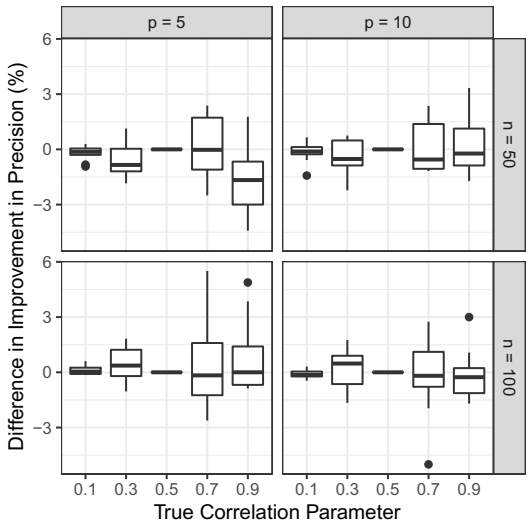

Figure 4: The differences of PIP between the locally optimal design with $\rho_0 = 0.5$ and the true optimal design with ρ_t for $n = 50, 100$ and $p = 5, 10$.

6.2 The Advantage of Considering Network Connection

In this subsection, we address the advantage of considering network connection in the design procedure. First, we comment on the influence of the network on the performance of the proposed optimal design. Essentially, the proposed locally optimal design aims to maximize $T(\boldsymbol{x}, \rho_0)$, which is equal to $m - T_1 - T_2$. Recall that m is twice the total number of edges and it increases as the network expands in the number of nodes n and/or network density. Therefore, for large and dense network, m can dominate the objective function $T(\boldsymbol{x}, \rho_0)$. Since the PIP (15) is calculated based on $T(\boldsymbol{x}, \rho_0)$, the advantage of the proposed optimal design would appear to be marginal. When n is not large and the network connection is sparse, the advantage of the proposed optimal design would be more significant. In the following, the locally optimal design is compared with the optimal design in (2) that does not consider network connections. The latter is obtained using Gurobi (Gurobi Optimization, 2015) and the run-time is also limited to 500 seconds, the same as the locally optimal design.

In Figure 5, we compare the locally optimal designs with network and the optimal design without network under different network densities. For each synthetic dataset, we obtain the two different optimal designs (i.e., with and without network connection) and obtain their respective PIP values. The boxplots are PIP values for 10 synthetic datasets under the same ρ_t, n, and network density setting. The results indicate that by incorporating the network structure, the proposed locally optimal design significantly outperforms the optimal design without a network connection, and this advantage is more prominent when network density is lower, the network size n is smaller, and the true value of correlation ρ_t is larger.

Similarly, in Figure 6, we expand such comparison to more cases of $n = 50, 100, 500, 1000$. In addition to the PIP in (15), we also include the improvement in $T_1(\boldsymbol{x}, \rho)$ and $T_2(\boldsymbol{x}, \rho)$ with respect to the expected T_1 and T_2 of the random balanced designs. Although for the locally optimal design with network, PIP value drops from 40% to 5% as n increases from 50 to 1000, the improvements in T_1 and T_2 do not decrease with the network size n. For an instance, the proposed optimal design for the cases with $n = 1000$ gives that $m \approx 20,000$, T_1 varies from -300 to -20, and T_2 varies from 0 to 200. As discussed earlier, the main reason is that m dominates the percentage of improvement

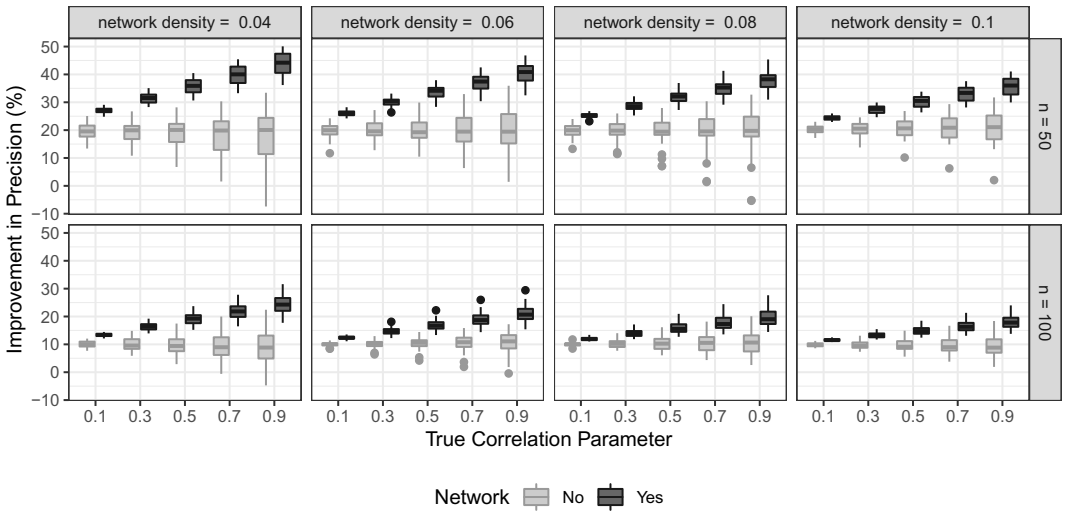

Figure 5: The PIP values of optimal designs with and without network with $p = 10$.

in precision when n is large and/or the network is dense.

7 Case Study

The case study is based on a real dataset from Rozemberczki et al. (2018), which is collected from the music streaming service (November 2017) with a total number of 47538 users from Hungary. The dataset contains the information of friendship networks of the users, as well as their covariates information, representing the users' preference (recorded by 1 or 0) to 84 distinct music genres. To decide if the update of the music recommender algorithm improves the baseline algorithm, a controlled experiment can be conducted, much similar to the application context given in Section 1. The outcome of each user can be the total time the user listening to the recommended music or a more direct metric commonly used by the company. The estimation of the treatment effect θ in (1) would reveal which one of the two versions of the recommender algorithm outperforms the other. Both the social network of users and their covariates are relevant in assessing different algorithms for music recommender systems. The non-interference assumption is proper for this case study since we assume the experiment is conducted without users' awareness.

To evaluate the performance of the proposed design approaches, we repeatedly randomly sample sub-networks with 2000 and 3000 users from the complete data of 47538 users. Among those, around half of the users are isolated from other users (i.e., no network connections at all). This number is large due to the subset sampling of the original complete network. The CAR model does not work for isolated users, since m_i has to be larger than zero. For simplicity, we remove those isolated users and the size of the remaining networks is approximately 1000 or 2000. In practice, all the isolated users can still be kept in the experiment and split into two groups via a covariate balancing measure. The densities of the resulting sub-networks range from 0.001 to 0.002. Although the complete data contains 84 distinct genres as the covariates, many of them are linearly dependent. Also, because of subset sampling, many covariates of the subsets become constants. Thus, we keep the first 20 covariates to remove the potential singularity issue. In the numerical study, we set p from 5 to 20.

We first compute the PIP values given in (15). For the locally optimal design, we set $\alpha = 0.001$

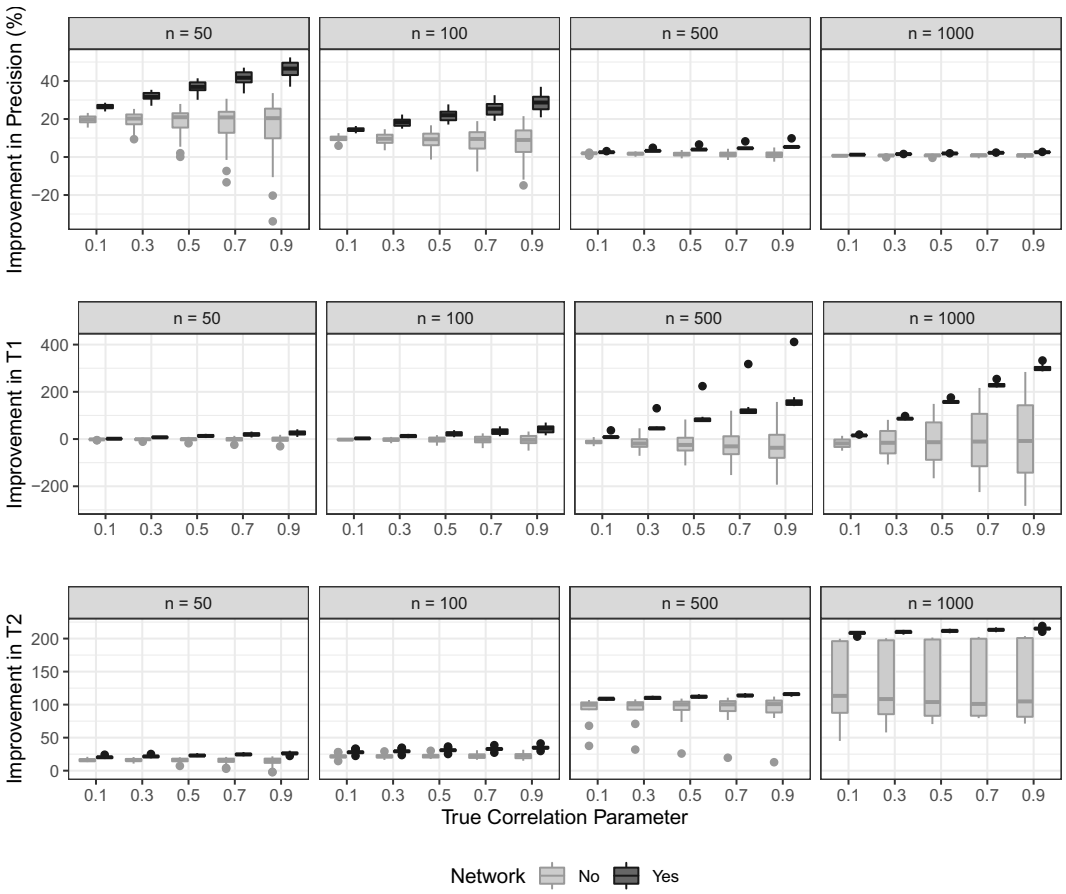

Figure 6: The PIPs of the locally optimal design with network and the optimal design without network with $p = 10$ and network density 0.02.

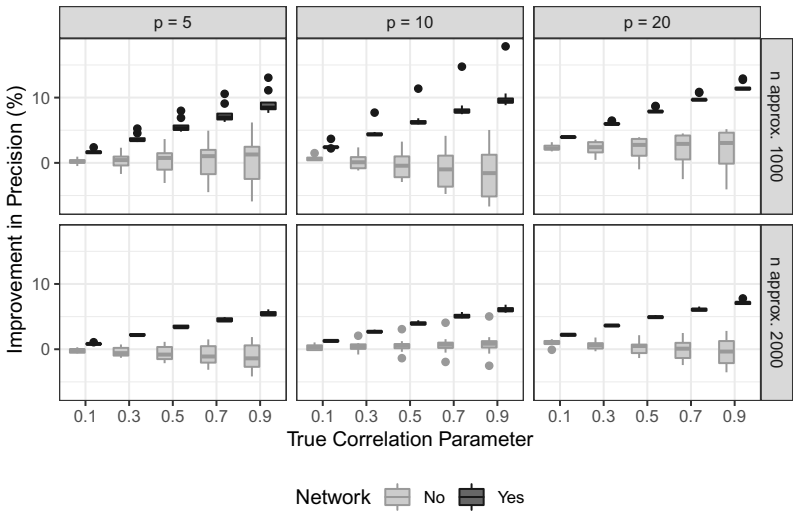

Figure 7: Boxplots of percentages of PIP values of two kinds of optimal designs for case study.

and $\rho_0 = 0.5$. The true correlation parameter ρ_t is varied from 0.1 to 0.9. We include the optimal design without network connection in (2) for comparison. For each n and p, ten subsets are randomly sampled from the complete data. The results are shown in Figure 7. The results based on the real data are different from synthetic datasets in many aspects. For instance, the distributions of covariates and networks are more complex. Particularly, we compute the proportions of 1's for each covariate, and it ranges in $[0.05, 0.85]$. The correlations between different covariates are in $[0.04, 0.97]$, so some of the covariates are highly correlated. Still, the results in Figure 7 show a similar pattern to the ones from synthetic networks, which indicates that the proposed approach is effective for real data sets as well, despite the more complicated network structure and covariates distributions.

Next, we create a pseudo experiment by simulating the outcome data and compare the two methods empirically. In reality, the network correlation coefficient for different users may not be the same. Therefore, to generate the outcome data, we set the covariance matrix of the CAR model to be $\sigma^2(\boldsymbol{D} - \boldsymbol{PWP})^{-1}$, where $\boldsymbol{P}$ is a diagonal matrix with entries $\sqrt{\rho_1}, \ldots, \sqrt{\rho_n}$. The heterogenous correlation coefficients $\rho_1, \ldots, \rho_n$ are sampled from the uniform distribution $U(0, 1)$. We simulate the outcomes from the CAR model in (5) under this covariance structure, and then fit a CAR model with a single unknown correlation coefficient and estimate the treatment effect θ in (1). We sample sub-networks with approximately 1000 users and take the first $p = 5$ or 20 covariates. The true treatment effect θ and the variance σ are specified to be 1. In addition to the locally optimal design and the optimal design without the network, we also generate 10 random balanced designs. For each design, we use the simulated outcome to obtain the estimate $\hat{\theta}$ based on the CAR model. Repeating this procedure 100 times, we compute the mean squared errors (MSEs) for each design approach. During each of the 100 times of simulation for each sampled sub-network and p covariates, we obtain 12 MSEs values (2 optimal designs and 10 random designs) and then compute the empirical percentiles of the MSEs of two optimal designs respectively from the MSEs of the 10 random designs. If the empirical percentile of the MSE from an optimal design is smaller than 0.5, it means that the MSE of the optimal design is superior to more than 50% of the random designs in terms of reducing the MSE. Notice that the resulting empirical percentiles vary from different sub-datasets in each simulation. For each p, we generate 25 random sub-datasets. The

empirical percentiles of MSEs of the two optimal designs are shown in the boxplots in Figure 8 for all the 25 sub-networks and two p values. The results show that the optimal design without the network does not outperform the random balanced designs. For the proposed locally optimal design with the network, the empirical percentiles are mostly below 0.5, which strongly indicates its advantage over the other two alternatives in terms of reducing MSE.

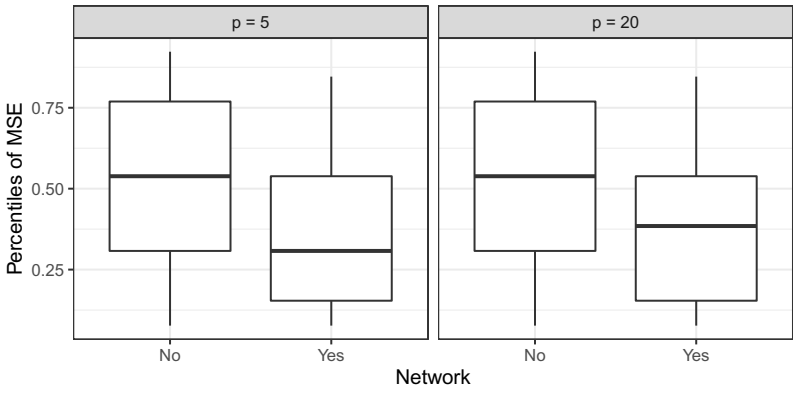

Figure 8: The percentiles of MSEs of the two optimal design approaches (i.e., with and without network) based on the MSEs of 10 completely randomized designs.

8 Conclusion

In this paper, we propose a model-based optimal design approach to include both covariates and network dependence for the experiments of A/B tests. A linear additive model is used to include the covariates information and the CAR model is used to model the network correlation between test subjects. A hybrid approach is proposed to solve the optimization problem and construct the locally optimal design. Both simulation and real data are used to compare the performances of the proposed locally optimal design with the exact optimal design and other alternative approaches. The proposed design performances reasonably well compared with other approaches in terms of variance reduction to random designs. The proposed optimal design relies on the CAR linear additive model including both covariates information and network correlation. Similarly to all optimal design approaches, the validity of the model assumption is crucial. Although we have shown the proposed design has some degree of robustness to the choice of the correlation parameter, if the experimenter thinks the CAR-based additive model assumption does not apply to the potential data to be collected, we recommend the rerandomization approaches proposed by Morgan and Rubin (2012) and Morgan and Rubin (2015) or completely randomized design if the sample size is sufficiently large.

We would like to point out a few directions for future research. First, this work is limited to the CAR model assumption and the network is much simpler than real social networks. But the proposed design approach can be applied to more sophisticated parametric models. For example, the network can become directional and weighted, which can be specified by the adjacency matrix. Other than the CAR model, the Spatial Auto-Regressive (SAR) model can be used. The network correlation parameter ρ can be different for different subjects as in the pseudo experiment we show in Section 7. Second, for the extremely large networks, there may be a time or economic cost to involve as many test subjects as possible. In this case, the optimal design proposed here can be extended to the optimization problem of simultaneous selection of test subjects and treatment

assignment. Some computational efficient approximation algorithms need to be adapted to solve this problem for large networks. Third, the proposed design relies on the observed covariates, which might be inaccurate depending on the data source. To make the design robust to inaccurate covariates, we may incorporate the uncertainty of those covariates, and develop a hierarchy model that can characterize the uncertainty. How to design treatment allocation under this situation would be an interesting topic. Moreover, it is also important to investigate designs when the number of treatment settings is more than two, particularly when the experiment involves multiple factors.

Acknowledgments

The authors thank Professor Roshan V. Joseph for handling this article as the editor and thank the Associate Editor and two reviewers for their valuable comments.

Funding

This research is supported by XXX.

Supplementary materials

The supplementary materials include proofs, derivations, and extra examples. They also include the codes for all the examples.

References

Atkinson, A. C. and Bailey, R. (2001), "One hundred years of the design of experiments on and off the pages of Biometrika," *Biometrika*, 88, 53–97.

Atkinson, A. C. and Donev, A. N. (1992), *Optimum experimental designs*, Oxford Science Publications, London.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2014), *Hierarchical Modeling and Analysis for Spatial Data*, New York: Chapman and Hall/CRC, 2nd ed.

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

— (2018b), "Model-assisted design of experiments in the presence of network-correlated outcomes," *Biometrika*, 105, 849–858.

Belotti, P., Kirches, C., Leyffer, S., Linderoth, J., Luedtke, J., and Mahajan, A. (2013), "Mixed-integer nonlinear optimization," *Acta Numerica*, 22, 1–131.

Bertsimas, D., Johnson, M., and Kallus, N. (2015), "The power of optimization over randomization in designing experiments involving small samples," *Operations Research*, 63, 868–876.

Besag, J. (1974), "Spatial interaction and the statistical analysis of lattice systems," *Journal of the Royal Statistical Society: Series B (Methodological)*, 36, 192–225.

Bhat, N., Farias, V. F., Moallemi, C. C., and Sinha, D. (2020), "Near-Optimal AB Testing," *Management Science*.

Chaloner, K. and Verdinelli, I. (1995), "Bayesian experimental design: A review," *Statistical Science*, 273–304.

Cressie, N. A. C. (1993), *Statistics for spatial data*, New York: Wiley, revised edition ed.

Drovandi, C. C. and Tran, M.-N. (2018), "Improving the Efficiency of Fully Bayesian Optimal Design of Experiments Using Randomised Quasi-Monte Carlo," *Bayesian Analysis*, 13, 139 – 162.

Eckles, D., Karrer, B., and Ugander, J. (2016), "Design and Analysis of Experiments in Networks: Reducing Bias from Interference," *Journal of Causal Inference*, 5, 20150021.

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*, pp. 399–409.

Gurobi Optimization, I. (2015), "Gurobi optimizer reference manual," *URL http://www. gurobi. com*.

Imbens, G. W. and Rubin, D. B. (2015), *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*, New York: Cambridge University Press.

Kallus, N. (2018), "Optimal a priori balance in the design of controlled experiments," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80, 85–112.

Kiefer, J. (1961), "Optimum designs in regression problems, II," *The Annals of Mathematical Statistics*, 298–325.

Kohavi, R., Tang, D., and Xu, Y. (2020), *Trustworthy online controlled experiments: A practical guide to a/b testing*, Cambridge University Press.

Koutra, V. (2017), "Designing experiments on networks," Ph.D. thesis, University of Southampton.

Li, Y., Kang, L., and Huang, X. (2021), "Covariate balancing based on kernel density estimates for controlled experiments," *Statistical Theory and Related Fields*, 5, 102–113.

Martin, R. (1986), "On the design of experiments under spatial correlation," *Biometrika*, 73, 247–277.

Morgan, K. L. and Rubin, D. B. (2012), "Rerandomization to improve covariate balance in experiments," *The Annals of Statistics*, 40, 1263–1282.

— (2015), "Rerandomization to balance tiers of covariates," *Journal of the American Statistical Association*, 110, 1412–1421.

Nandy, P., Basu, K., Chatterjee, S., and Tu, Y. (2020), "A/B testing in dense large-scale networks: design and inference," *Advances in Neural Information Processing Systems*, 33.

Parker, B. M., Gilmour, S. G., and Schormans, J. (2017), "Optimal design of experiments on connected units with application to social networks," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 3, 455–480.

Phan, T. Q. and Airoldi, E. M. (2015), "A natural experiment of social network formation and dynamics," *Proceedings of the National Academy of Sciences*, 112, 6595–6600.

Pokhilko, V., Zhang, Q., Kang, L., and Darcy, P. M. (2019), "D-Optimal Design for Network A/B Testing," *Journal of Statistical Theory and Practice*, 13, 61.

Rozemberczki, B., Davies, R., Sarkar, R., and Sutton, C. (2018), "GEMSEC: Graph Embedding with Self Clustering," .

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

— (2005), "Causal inference using potential outcomes: Design, modeling, decisions," *Journal of the American Statistical Association*, 100, 322–331.

Rue, H. and Held, L. (2005), *Gaussian Markov Random Fields: Theory and Applications*, New York: Chapman and Hall/CRC.

Ryan, E. G., Drovandi, C. C., McGree, J. M., and Pettitt, A. N. (2016), "A Review of Modern Computational Algorithms for Bayesian Optimal Design," *International Statistical Review*, 84, 128–154.

Ryan, E. G., Drovandi, C. C., Thompson, M. H., and Pettitt, A. N. (2014), "Towards Bayesian experimental design for nonlinear models that require a large number of sampling times," *Computational Statistics & Data Analysis*, 70, 45–60.

Ver Hoef, J. M., Hanks, E. M., and Hooten, M. B. (2018), "On the relationship between conditional (CAR) and simultaneous (SAR) autoregressive models," *Spatial statistics*, 25, 68–85.

Supplement: Proofs, Derivations and Extra Example

S1. Proof of Theorem 1

Proof. Because $T(\boldsymbol{x}, \rho) = m - T_1(\boldsymbol{x}, \rho) - T_2(\boldsymbol{x}, \rho)$, we investigate the derivatives T_1 and T_2 with respect to ρ. For T_1,

$$T_1(\boldsymbol{x}, \rho) = \rho \boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{x},$$
$$\frac{\partial T_1(\boldsymbol{x}, \rho)}{\partial \rho} = \boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{x}, \quad \frac{\partial^2 T_1(\boldsymbol{x}, \rho)}{\partial \rho^2} = 0.$$

To derive the derivatives for T_2, we first introduce some notation to shorten the formulas. Let $\boldsymbol{A} := \boldsymbol{F}^\top (\boldsymbol{D} - \rho \boldsymbol{W}) F$, $\boldsymbol{A}_1 := \frac{\partial \boldsymbol{A}^{-1}}{\partial \rho}$, and $\boldsymbol{A}_2 := \frac{\partial^2 \boldsymbol{A}^{-1}}{\partial \rho^2}$. Following the calculus of matrix,

$$\boldsymbol{A}_1 = -\boldsymbol{A}^{-1} \frac{\partial \boldsymbol{A}}{\partial \rho} \boldsymbol{A}^{-1} = -\boldsymbol{A}^{-1} \frac{\partial \boldsymbol{F}^\top \boldsymbol{D} \boldsymbol{F} - \rho \boldsymbol{F}^\top \boldsymbol{W} \boldsymbol{F}}{\partial \rho} \boldsymbol{A}^{-1} = \boldsymbol{A}^{-1} \boldsymbol{F}^\top \boldsymbol{W} \boldsymbol{F} \boldsymbol{A}^{-1},$$
$$\boldsymbol{A}_2 = \frac{\partial \boldsymbol{A}_1}{\partial \rho} = \frac{\partial \boldsymbol{A}^{-1}}{\partial \rho} \boldsymbol{F}^\top \boldsymbol{W} \boldsymbol{F} \boldsymbol{A}^{-1} + \boldsymbol{A}^{-1} \boldsymbol{F}^\top \boldsymbol{W} \boldsymbol{F} \frac{\partial \boldsymbol{A}^{-1}}{\partial \rho} = 2\boldsymbol{A}^{-1} \boldsymbol{F}^\top \boldsymbol{W} \boldsymbol{F} \boldsymbol{A}^{-1} \boldsymbol{F}^\top \boldsymbol{W} \boldsymbol{F} \boldsymbol{A}^{-1}.$$

Using the new notation,

$$\begin{aligned} T_2(\boldsymbol{x}, \rho) &= \boldsymbol{x}^\top (\boldsymbol{D} - \rho \boldsymbol{W}) \boldsymbol{F} \boldsymbol{A}^{-1} \boldsymbol{F}^\top (\boldsymbol{D} - \rho \boldsymbol{W}) \boldsymbol{x} \\ &= \underbrace{\boldsymbol{x}^\top \boldsymbol{D} \boldsymbol{F} \boldsymbol{A}^{-1} \boldsymbol{F}^\top \boldsymbol{D} \boldsymbol{x}}_{\text{Term 1}} - 2 \underbrace{\rho \boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{F} \boldsymbol{A}^{-1} \boldsymbol{F}^\top \boldsymbol{D} \boldsymbol{x}}_{\text{Term 2}} + \underbrace{\rho^2 \boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{F} \boldsymbol{A}^{-1} \boldsymbol{F}^\top \boldsymbol{W} \boldsymbol{x}}_{\text{Term 3}}. \end{aligned}$$

The first order derivative of the three terms with respect to ρ are

$$\begin{aligned} \frac{\partial \text{Term 1}}{\partial \rho} &= \boldsymbol{x}^\top \boldsymbol{D} \boldsymbol{F} \frac{\partial \boldsymbol{A}^{-1}}{\partial \rho} \boldsymbol{F}^\top \boldsymbol{D} \boldsymbol{x} = \boldsymbol{x}^\top \boldsymbol{D} \boldsymbol{F} \boldsymbol{A}_1 F^\top \boldsymbol{D} \boldsymbol{x} \\ \frac{\partial \text{Term 2}}{\partial \rho} &= \boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{F} \boldsymbol{A}^{-1} \boldsymbol{F}^\top \boldsymbol{D} \boldsymbol{x} + \rho \boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{F} \frac{\partial \boldsymbol{A}^{-1}}{\partial \rho} \boldsymbol{F}^\top \boldsymbol{D} \boldsymbol{x} \\ &= \boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{F} \boldsymbol{A}^{-1} \boldsymbol{F}^\top \boldsymbol{D} \boldsymbol{x} + \rho \boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{F} \boldsymbol{A}_1 \boldsymbol{F}^\top \boldsymbol{D} \boldsymbol{x} \\ \frac{\partial \text{Term 3}}{\partial \rho} &= 2\rho \boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{F} \boldsymbol{A}^{-1} \boldsymbol{F}^\top \boldsymbol{W} \boldsymbol{x} + \rho^2 \boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{F} \frac{\partial \boldsymbol{A}^{-1}}{\partial \rho} \boldsymbol{F}^\top \boldsymbol{W} \boldsymbol{x} \\ &= 2\rho \boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{F} \boldsymbol{A}^{-1} \boldsymbol{F}^\top \boldsymbol{W} \boldsymbol{x} + \rho^2 \boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{F} \boldsymbol{A}_1 \boldsymbol{F}^\top \boldsymbol{W} \boldsymbol{x}. \end{aligned}$$

To combine the three derivatives,

$$\frac{\partial T_2(\boldsymbol{x}, \rho)}{\partial \rho} = \boldsymbol{x}^\top (\boldsymbol{D} - \rho \boldsymbol{W}) \boldsymbol{F} \boldsymbol{A}_1 \boldsymbol{F}^\top (\boldsymbol{D} - \rho \boldsymbol{W}) \boldsymbol{x} - 2\boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{F} \boldsymbol{A}^{-1} \boldsymbol{F}^\top (\boldsymbol{D} - \rho \boldsymbol{W}) \boldsymbol{x}$$

The derivative of $T(\boldsymbol{x}, \rho)$ is,

$$\begin{aligned} \frac{\partial T(\boldsymbol{x}, \rho)}{\partial \rho} &= -\boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{x} - \boldsymbol{x}^\top (\boldsymbol{D} - \rho \boldsymbol{W}) \boldsymbol{F} \boldsymbol{A}_1 \boldsymbol{F}^\top (\boldsymbol{D} - \rho \boldsymbol{W}) \boldsymbol{x} + 2\boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{F} \boldsymbol{A}^{-1} \boldsymbol{F}^\top (\boldsymbol{D} - \rho W) \boldsymbol{x} \\ &= -\boldsymbol{x}^\top \boldsymbol{W} \left[\boldsymbol{I}_n - \boldsymbol{F} \boldsymbol{A}^{-1} \boldsymbol{F}^\top (\boldsymbol{D} - \rho \boldsymbol{W}) \right] \boldsymbol{x} \\ &\quad - \boldsymbol{x}^\top \left[(\boldsymbol{D} - \rho \boldsymbol{W}) \boldsymbol{F} \boldsymbol{A}^{-1} \boldsymbol{F}^\top - \boldsymbol{I}_n \right] \boldsymbol{W} \boldsymbol{F} \boldsymbol{A}^{-1} \boldsymbol{F}^\top (\boldsymbol{D} - \rho \boldsymbol{W}) \boldsymbol{x} \\ &= -\boldsymbol{x}^\top \left[\boldsymbol{I}_n - \boldsymbol{F} \boldsymbol{A}^{-1} \boldsymbol{F}^\top (\boldsymbol{D} - \rho \boldsymbol{W}) \right]^\top \boldsymbol{W} \left[\boldsymbol{I}_n - \boldsymbol{F} \boldsymbol{A}^{-1} \boldsymbol{F}^\top (\boldsymbol{D} - \rho \boldsymbol{W}) \right] \boldsymbol{x}. \end{aligned}$$

It is interesting to notice that $\boldsymbol{s} := \left[\boldsymbol{I}_n - \boldsymbol{F}\boldsymbol{A}^{-1}\boldsymbol{F}^\top(\boldsymbol{D} - \rho\boldsymbol{W})\right]\boldsymbol{x}$ can be considered as the residuals of regression model $\boldsymbol{x} = \boldsymbol{F}\boldsymbol{\beta} + \boldsymbol{v}$, where $\boldsymbol{v}$ is the vector with mean equal to $\boldsymbol{0}$ and covariance matrix $\boldsymbol{D} - \rho\boldsymbol{W}$. By the definition of the adjacency matrix,

$$\frac{\partial T(\boldsymbol{x}, \rho)}{\partial \rho} = -\sum_{w_{i,j}=1} s_i s_j.$$

Thus, the sign of $\partial T(\boldsymbol{x}, \rho)/\partial\rho$ is uncertain and is possible to be either positive or negative.

Next, we compute the second order derivative of $T_2(\boldsymbol{x}, \rho)$ with respect to ρ.

$$\begin{aligned}
\frac{\partial^2 \text{Term 1}}{\partial \rho^2} &= \boldsymbol{x}^\top \boldsymbol{D}\boldsymbol{F}\frac{\partial \boldsymbol{A}_1}{\partial \rho}\boldsymbol{F}^\top \boldsymbol{D}\boldsymbol{x} = \boldsymbol{x}^\top \boldsymbol{D}\boldsymbol{F}\boldsymbol{A}_2\boldsymbol{F}^\top \boldsymbol{D}\boldsymbol{x}, \\
\frac{\partial^2 \text{Term 2}}{\partial \rho^2} &= \boldsymbol{x}^\top \boldsymbol{W}\boldsymbol{F}\frac{\partial \boldsymbol{A}^{-1}}{\partial \rho}\boldsymbol{F}^\top \boldsymbol{D}\boldsymbol{x} + \boldsymbol{x}^\top \boldsymbol{W}\boldsymbol{F}\boldsymbol{A}_1 F^\top \boldsymbol{D}\boldsymbol{x} + \rho\boldsymbol{x}^\top \boldsymbol{W}\boldsymbol{F}\frac{\partial \boldsymbol{A}_1}{\partial \rho}\boldsymbol{F}^\top \boldsymbol{D}\boldsymbol{x} \\
&= 2\boldsymbol{x}^\top \boldsymbol{W}\boldsymbol{F}\boldsymbol{A}_1\boldsymbol{F}^\top \boldsymbol{D}\boldsymbol{x} + \rho\boldsymbol{x}^\top \boldsymbol{W}\boldsymbol{F}\boldsymbol{A}_2\boldsymbol{F}^\top \boldsymbol{D}\boldsymbol{x}, \\
\frac{\partial^2 \text{Term 3}}{\partial \rho^2} &= 2\boldsymbol{x}^\top \boldsymbol{W}\boldsymbol{F}\boldsymbol{A}^{-1}\boldsymbol{F}^\top \boldsymbol{W}\boldsymbol{x} + 2\rho\boldsymbol{x}^\top \boldsymbol{W}\boldsymbol{F}\frac{\partial \boldsymbol{A}^{-1}}{\partial \rho}\boldsymbol{F}^\top \boldsymbol{W}\boldsymbol{x} + 2\rho\boldsymbol{x}^\top \boldsymbol{W}\boldsymbol{F}\boldsymbol{A}_1\boldsymbol{F}^\top \boldsymbol{W}\boldsymbol{x} \\
&\quad + \rho^2\boldsymbol{x}^\top \boldsymbol{W}\boldsymbol{F}\frac{\partial \boldsymbol{A}_1}{\partial \rho}\boldsymbol{F}^\top \boldsymbol{W}\boldsymbol{x} \\
&= 2\boldsymbol{x}^\top \boldsymbol{W}\boldsymbol{F}\boldsymbol{A}^{-1}\boldsymbol{F}^\top \boldsymbol{W}\boldsymbol{x} + 4\rho\boldsymbol{x}^\top \boldsymbol{W}\boldsymbol{F}\boldsymbol{A}_1\boldsymbol{F}^\top \boldsymbol{W}\boldsymbol{x} + \rho^2\boldsymbol{x}^\top \boldsymbol{W}\boldsymbol{F}\boldsymbol{A}_2\boldsymbol{F}^\top \boldsymbol{W}\boldsymbol{x}.
\end{aligned}$$

Let $\boldsymbol{C} := \boldsymbol{F}\boldsymbol{A}^{-1}\boldsymbol{F}^\top \boldsymbol{W}\boldsymbol{F}$. Then

$$\begin{aligned}
\frac{\partial^2 T_2(\boldsymbol{x}, \rho)}{\partial \rho^2} &= 2\boldsymbol{x}^\top(\boldsymbol{D} - \rho\boldsymbol{W})\boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{C}^\top(\boldsymbol{D} - \rho\boldsymbol{W})\boldsymbol{x} - 4\boldsymbol{x}^\top \boldsymbol{W}\boldsymbol{F}\boldsymbol{A}^{-1}\boldsymbol{C}^\top(\boldsymbol{D} - \rho\boldsymbol{W})\boldsymbol{x} \\
&\quad + 2\boldsymbol{x}^\top \boldsymbol{W}\boldsymbol{F}\boldsymbol{A}^{-1}\boldsymbol{F}^\top \boldsymbol{W}\boldsymbol{x} \\
&= 2\boldsymbol{x}^\top \left[(\boldsymbol{D} - \rho\boldsymbol{W})\boldsymbol{C} - \boldsymbol{W}\boldsymbol{F}\right]\boldsymbol{A}^{-1}\left[\boldsymbol{C}^\top(\boldsymbol{D} - \rho\boldsymbol{W}) - \boldsymbol{F}^\top \boldsymbol{W}\right]\boldsymbol{x}.
\end{aligned}$$

For any $\rho \in (0, 1)$, it is apparent that $\boldsymbol{D} - \rho\boldsymbol{W}$ is the Laplacian matrix of the weighted undirected graph with the constant weight ρ for each edge, and it is also clear that $\boldsymbol{D} - \rho\boldsymbol{W}$ is a positive definite matrix. We assume $\boldsymbol{F}$ is a full rank matrix so that the regression model is valid. So $\boldsymbol{A}$ and $\boldsymbol{A}^{-1}$ are both positive definite. Thus, $\frac{\partial^2 T_2(\boldsymbol{x},\rho)}{\partial \rho^2} \geq 0$ and $\frac{\partial^2 T(\boldsymbol{x},\rho)}{\partial \rho^2} \leq 0$ for any $\rho \in (0, 1)$. The design criterion $T(\boldsymbol{x}, \rho)$, which is to be maximized, is concave. □

S2. The Gap between $\mathbb{E}[T(\boldsymbol{x}, \rho)]$ and $T(\boldsymbol{x}, \rho_0)$

We randomly generate a network of size $n = 50$. For each pair of nodes, an edge will connect the two with a probability of $1/4$ and the existence of the edge is independent of any other random variables. The covariate z_i is generated from a one-dimensional normal distribution $N(0, 10^2)$ and z_i's are independent of each other and the network structure. The prior distribution of ρ is uniform distribution in $[0, 1]$ and $\rho_0 = \mathbb{E}(\rho) = 1/2$. We randomly generate 400 completely randomized designs $\boldsymbol{x}_l$ for $l = 1, \ldots, 400$ and calculate $T(\boldsymbol{x}_l, \rho_0)$, whose histogram is plotted in the left panel of Figure S1. For any given design $\boldsymbol{x}_l$, we randomly samples ρ_i for $i = 1, \ldots, 200$ and calculate $T(\boldsymbol{x}_l, \rho_i)$. The mean $\mathbb{E}[T(\boldsymbol{x}_l, \rho)]$ is approximated by the sample mean of $T(\boldsymbol{x}_l, \rho_i)$'s. The histogram of the gap $T(\boldsymbol{x}_l, \rho_0) - \mathbb{E}[T(\boldsymbol{x}_l, \rho)]$ for all the random designs is plotted in the right panel of Figure S1. Based on the two histograms, the gap $T(\boldsymbol{x}, \rho_0) - \mathbb{E}[T(\boldsymbol{x}, \rho)]$ is relatively small compared to

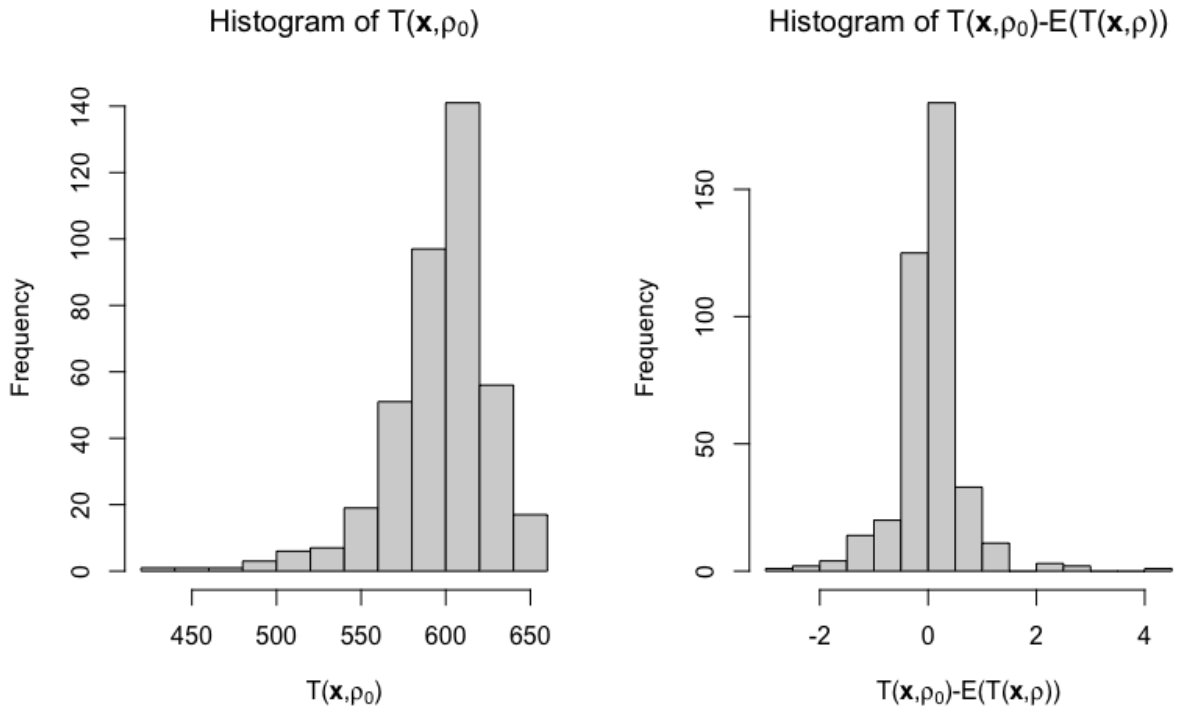

Figure S1: Histogram of $T(\boldsymbol{x}, \rho_0)$ and the gap $T(\boldsymbol{x}, \rho_0) - \mathbb{E}[T(\boldsymbol{x}, \rho)]$

the range of $T(\boldsymbol{x}, \rho_0)$. Thus, it is reasonable to use the surrogate local design criterion $T(\boldsymbol{x}_l, \rho_0)$ to replace $\mathbb{E}[T(\boldsymbol{x}, \rho)]$ for this simple example.

In more general case, Proposition S1 provides the analytic gap between $T(\boldsymbol{x}, \rho_0)$ and $\mathbb{E}[T(\boldsymbol{x}, \rho)]$. Its proof is provided in the Supplement. Proposition S1 also provides two different upper bounds of the gap. Which one of the two upper bounds is larger depends on the adjacency matrix $\boldsymbol{W}$ and ρ_0. Regrettably, since both the upper bounds are independent of the design $\boldsymbol{x}$, they are too large to have any practical guidance, even though they might still be attainable for certain extreme design $\boldsymbol{x}$. For the above simulation example, since the skewness of uniform distribution is 0, the two upper bounds of (19) and (20) are calculated as 902.4 and 650.1, respectively. They are much larger than the range shown in the histogram in Figure S1. On the other hand, the two upper bounds increase as the size and density of the network become larger. Therefore, for large and dense networks we should be more careful applying the locally optimal design.

Proposition S1. *The difference between* $T(\boldsymbol{x}, \rho_0)$ *and* $\mathbb{E}(T(\boldsymbol{x}, \rho))$ *is*

$$T(\boldsymbol{x}, \rho_0) - \mathbb{E}(T(\boldsymbol{x}, \rho)) = \frac{1}{2} \frac{\partial^2 T_2(\boldsymbol{x}, \rho)}{\partial \rho^2}\bigg|_{\rho=\rho_0} \mathrm{var}(\rho) - \mathbb{E}(O(\rho - \rho_0)^3), \tag{16}$$

where

$$\frac{1}{2} \frac{\partial^2 T_2(\boldsymbol{x}, \rho)}{\partial \rho^2}\bigg|_{\rho=\rho_0} = \boldsymbol{s}^\top \boldsymbol{W}\boldsymbol{F} \left[\boldsymbol{F}^\top (\boldsymbol{D} - \rho_0 \boldsymbol{W})\boldsymbol{F}\right]^{-1} \boldsymbol{F}^\top \boldsymbol{W}\boldsymbol{s}, \tag{17}$$

$$\textit{and} \quad \boldsymbol{s} := \left[\boldsymbol{I}_n - \boldsymbol{F}(\boldsymbol{F}^\top (\boldsymbol{D} - \rho_0 \boldsymbol{W})\boldsymbol{F})^{-1} \boldsymbol{F}^\top (\boldsymbol{D} - \rho_0 \boldsymbol{W})\right] \boldsymbol{x}. \tag{18}$$

An upper bound of the gap $T(\boldsymbol{x},\rho_0)-\mathbb{E}(T(\boldsymbol{x},\rho))$ *is*

$$T(\boldsymbol{x},\rho_0)-\mathbb{E}(T(\boldsymbol{x},\rho))\leq \min\left\{n\lambda_{\max}(\boldsymbol{D}-\rho_0\boldsymbol{W}),(1+\rho_0)m\right\}\frac{|\lambda(\boldsymbol{W})|^2_{\max}\mathrm{var}(\rho)}{\lambda^2_{\min}(\boldsymbol{D}-\rho_0\boldsymbol{W})}-\mathbb{E}\left[O(\rho-\rho_0)^3\right], \tag{19}$$

where $\lambda_{\min}(\boldsymbol{D}-\rho_0\boldsymbol{W})$ *and* $\lambda_{\max}(\boldsymbol{D}-\rho_0\boldsymbol{W})$ *are the minimum and maximum eigenvalues of the Laplacian matrix* $\boldsymbol{D}-\rho_0\boldsymbol{W}$*, which is positive definite for* $\rho_0\in(0,1)$*,* $|\lambda(\boldsymbol{W})|_{\max}$ *is the spectrum radius of* $\boldsymbol{W}$*, and* $m=\sum_{i=1}^n m_i$*. Based on Theorem 2, an alternative upper bound (20) holds asymptotically with probability of* $100(1-\alpha)\%$ *and* $\alpha\in(0,1)$*,*

$$T(\boldsymbol{x},\rho_0)-\mathbb{E}(T(\boldsymbol{x},\rho))\leq (m+z_\alpha\sqrt{m})\frac{|\lambda(\boldsymbol{W})|^2_{\max}\mathrm{var}(\rho)}{\lambda^2_{\min}(\boldsymbol{D}-\rho_0\boldsymbol{W})}-\mathbb{E}\left[O(\rho-\rho_0)^3\right], \tag{20}$$

where $z_\alpha=\Phi^{-1}(\alpha)$ *is the upper* α *quantile of the standard normal distribution.*

Lemma S1. *Let* $\boldsymbol{A}$ *be an* $n\times n$ *real symmetric positive definite matrix. For any vector* $\boldsymbol{x}\in\mathbb{R}^n$*,* $\lambda_{\min}(\boldsymbol{A})||\boldsymbol{x}||_2^2\leq \boldsymbol{x}^\top\boldsymbol{A}\boldsymbol{x}\leq\lambda_{\max}(\boldsymbol{A})||\boldsymbol{x}||_2^2$*. The equality holds if* $\boldsymbol{x}=\boldsymbol{0}$ *or* $\boldsymbol{A}=a\boldsymbol{I}_n$ *for* $a\geq 0$*.*

Proof. Because $\boldsymbol{A}$ is a real symmetric positive definite matrix, via eigendecomposition, $\boldsymbol{A}=\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^{-1}$, where $\boldsymbol{\Lambda}=\mathrm{diag}\{\lambda_1,\ldots,\lambda_n\}$ is a diagonal matrix of the eigenvalues of $\boldsymbol{A}$, $\boldsymbol{Q}$ is the square $n\times n$ matrix whose ith column is the eigenvector corresponding to eigenvalue λ_i. Also, $\boldsymbol{Q}^\top=\boldsymbol{Q}^{-1}$. Denote $\boldsymbol{l}:=\boldsymbol{Q}^\top\boldsymbol{x}$.

$$\boldsymbol{x}^\top\boldsymbol{A}\boldsymbol{x}=\boldsymbol{x}^\top\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^\top\boldsymbol{x}=\boldsymbol{l}^\top\boldsymbol{\Lambda}\boldsymbol{l}=\sum_{i=1}^n\lambda_i l_i^2,$$

$$\lambda_{\min}(\boldsymbol{A})||\boldsymbol{l}||_2^2=\lambda_{\min}(\boldsymbol{A})\sum_{i=1}^n l_i^2\leq\sum_{i=1}^n\lambda_i l_i^2\leq\lambda_{\max}(\boldsymbol{A})\sum_{i=1}^n l_i^2=\lambda_{\max}(\boldsymbol{A})||\boldsymbol{l}||_2^2.$$

Here $\lambda_{\max}(\boldsymbol{A})$ and $\lambda_{\min}(\boldsymbol{A})$ are the maximum and minimum eigenvalues of $\boldsymbol{A}$, and since $\boldsymbol{A}$ is positive definite, $\lambda_{\min}(\boldsymbol{A})>0$. The norm $||\cdot||_2$ is the l_2-norm of a vector, and $||\boldsymbol{l}||_2^2=\boldsymbol{l}^\top\boldsymbol{l}=\boldsymbol{x}^\top\boldsymbol{Q}\boldsymbol{Q}^\top\boldsymbol{x}=||\boldsymbol{x}||_2^2$. Thus the lemma is proved. □

Lemma S2. *Let* $\boldsymbol{A}$ *be an* $n\times n$ *real symmetric matrix. For any vector* $\boldsymbol{x}\in\mathbb{R}^n$*,* $|\boldsymbol{x}^\top\boldsymbol{A}\boldsymbol{x}|\leq|\lambda(\boldsymbol{A})|_{\max}||\boldsymbol{x}||_2^2$*.*

Proof. For any real symmetric matrix, based on eigenvalue decomposition, $\boldsymbol{A}=\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^\top$, where $\boldsymbol{\Lambda}=\mathrm{diag}\{\lambda_1,\ldots,\lambda_n\}$ is a diagonal matrix of the eigenvalues of $\boldsymbol{A}$, and $\boldsymbol{Q}$ is the $n\times n$ orthogonal matrix as above. Denote $\boldsymbol{l}:=\boldsymbol{Q}^\top\boldsymbol{x}$.

$$|\boldsymbol{x}^\top\boldsymbol{A}\boldsymbol{x}|=|\boldsymbol{x}^\top\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^\top\boldsymbol{x}|=|\boldsymbol{l}^\top\boldsymbol{\Lambda}\boldsymbol{l}|=|\sum_{i=1}^n\lambda_i l_i^2|\leq\sum_{i=1}^n|\lambda_i|l_i^2\leq|\lambda(\boldsymbol{A})|_{\max}||\boldsymbol{l}||_2^2=|\lambda(\boldsymbol{A})|_{\max}||\boldsymbol{x}||_2^2.$$

Here $|\lambda(\boldsymbol{A})|_{\max}=\max_{i=1,\ldots,n}|\lambda|_i$. □

Proof of Proposition S1

Proof. Using Taylor expansion, we have

$$T(\boldsymbol{x},\rho)=T(\boldsymbol{x},\rho_0)+\left.\frac{\partial T(\boldsymbol{x},\rho)}{\partial\rho}\right|_{\rho=\rho_0}(\rho-\rho_0)+\frac{1}{2}\left.\frac{\partial^2 T(\boldsymbol{x},\rho)}{\partial\rho^2}\right|_{\rho=\rho_0}(\rho-\rho_0)^2+O((\rho-\rho_0)^3).$$

Apply expectation on both side of the equaiton with respet the priori $p(\rho)$, we have

$$\begin{aligned}\mathbb{E}\left[T(\boldsymbol{x},\rho)\right] &= T(\boldsymbol{x},\rho_0) + \left.\frac{\partial T(\boldsymbol{x},\rho)}{\partial \rho}\right|_{\rho=\rho_0} \mathbb{E}\left[\rho-\rho_0\right] + \frac{1}{2}\left.\frac{\partial^2 T(\boldsymbol{x},\rho)}{\partial \rho^2}\right|_{\rho=\rho_0} \mathbb{E}\left[(\rho-\rho_0)^2\right] + \mathbb{E}\left[O((\rho-\rho_0)^3)\right] \\ &= T(\boldsymbol{x},\rho_0) + \frac{1}{2}\left.\frac{\partial^2 T(\boldsymbol{x},\rho)}{\partial \rho^2}\right|_{\rho=\rho_0} \mathrm{var}(\rho) + \mathbb{E}\left[O((\rho-\rho_0)^3)\right].\end{aligned}$$

From the proof of Theorem 1, we have that

$$\frac{\partial^2 T(\boldsymbol{x},\rho)}{\partial \rho^2} = -\frac{\partial^2 T_2(\boldsymbol{x},\rho)}{\partial \rho^2}.$$

Thus we obtain the gap between $T(\boldsymbol{x},\rho_0)$ and $\mathbb{E}\left[T(\boldsymbol{x},\rho)\right]$ in (16). Also in proof of Theorem 1,

$$\frac{1}{2}\left.\frac{\partial^2 T_2(\boldsymbol{x},\rho)}{\partial \rho^2}\right|_{\rho=\rho_0} = \boldsymbol{s}^\top \boldsymbol{W}\boldsymbol{F}\boldsymbol{A}^{-1}\boldsymbol{F}^\top \boldsymbol{W}\boldsymbol{s},$$

where

$$\begin{aligned}\boldsymbol{A} &= \boldsymbol{F}^\top(\boldsymbol{D}-\rho_0\boldsymbol{W})\boldsymbol{F}, \\ \boldsymbol{s} &= \left[\boldsymbol{I}_n - \boldsymbol{F}\boldsymbol{A}^{-1}\boldsymbol{F}^\top(\boldsymbol{D}-\rho_0\boldsymbol{W})\right]\boldsymbol{x}.\end{aligned}$$

From the definition of $\boldsymbol{s}$, we can see that

$$\begin{aligned}\boldsymbol{s}^\top(\boldsymbol{D}-\rho_0\boldsymbol{W})\boldsymbol{s} &= \boldsymbol{x}^\top\left[(\boldsymbol{D}-\rho_0\boldsymbol{W}) - (\boldsymbol{D}-\rho_0\boldsymbol{W})\boldsymbol{F}\boldsymbol{A}^{-1}\boldsymbol{F}^\top(\boldsymbol{D}-\rho_0\boldsymbol{W})\right]\boldsymbol{x} \\ &\le \boldsymbol{x}^\top(\boldsymbol{D}-\rho_0\boldsymbol{W})\boldsymbol{x}.\end{aligned}$$

From Lemma S1, since $\boldsymbol{D}-\rho_0\boldsymbol{W}$ is a real symmetric positive definite matrix as $\rho_0 \in (0,1)$,

$$\lambda_{\min}(\boldsymbol{D}-\rho_0\boldsymbol{W})||\boldsymbol{s}||_2^2 \le \lambda_{\max}(\boldsymbol{D}-\rho_0\boldsymbol{W})||\boldsymbol{x}||_2^2 = \lambda_{\max}(\boldsymbol{D}-\rho_0\boldsymbol{W})n.$$

On the other hand, $\boldsymbol{x}^\top(\boldsymbol{D}-\rho_0\boldsymbol{W})\boldsymbol{x} \le (1+\rho_0)m$. Thus,

$$||\boldsymbol{s}||_2^2 \le \frac{1}{\lambda_{\min}(\boldsymbol{D}-\rho_0\boldsymbol{W})}\min\{n\lambda_{\max}(\boldsymbol{D}-\rho_0\boldsymbol{W}),(1+\rho_0)m\}.$$

According to Theorem 2, $\boldsymbol{x}^\top\boldsymbol{W}\boldsymbol{x}/\sqrt{m}$ converges in distribution to the standard normal distribution. Therefore, with probability of $100(1-\alpha)\%$, $\boldsymbol{x}^\top\boldsymbol{W}\boldsymbol{x} \ge -z_\alpha\sqrt{m}$, asymptotically. Here z_α is the upper α quantile of the standard normal distribution, i.e., $z_\alpha = \Phi^{-1}(1-\alpha)$. So we can obtain an asymptotic upper bound,

$$\boldsymbol{s}^\top(\boldsymbol{D}-\rho_0\boldsymbol{W})\boldsymbol{s} \le \boldsymbol{x}^\top(\boldsymbol{D}-\rho_0\boldsymbol{W})\boldsymbol{x} = \boldsymbol{x}^\top\boldsymbol{D}\boldsymbol{x} - \rho_0\boldsymbol{x}^\top\boldsymbol{W}\boldsymbol{x} = m - \rho_0\boldsymbol{x}^\top\boldsymbol{W}\boldsymbol{x} \le m + z_\alpha\sqrt{m},$$

which holds with probability of $100(1-\alpha)\%$. Consequently, an asymptotic upper bound for $||s||_2^2$ is

$$||\boldsymbol{s}||_2^2 \le \frac{1}{\lambda_{\min}(\boldsymbol{D}-\rho_0\boldsymbol{W})}(m+z_\alpha\sqrt{m})$$

with probability of $100(1-\alpha)\%$.

It is easy to see that the matrix

$$\boldsymbol{I}_n - (\boldsymbol{D}-\rho_0\boldsymbol{W})^{1/2}\boldsymbol{F}\boldsymbol{A}^{-1}\boldsymbol{F}^\top(\boldsymbol{D}-\rho_0\boldsymbol{W})^{1/2}$$

is a projection matrix, and thus

$$\begin{aligned}
&\boldsymbol{s}^\top \boldsymbol{W}\boldsymbol{F}\boldsymbol{A}^{-1}\boldsymbol{F}^\top \boldsymbol{W}\boldsymbol{s} \\
=&\boldsymbol{s}^\top \boldsymbol{W}(\boldsymbol{D}-\rho_0\boldsymbol{W})^{-1/2}(\boldsymbol{D}-\rho_0\boldsymbol{W})^{1/2}\boldsymbol{F}\boldsymbol{A}^{-1}\boldsymbol{F}^\top(\boldsymbol{D}-\rho_0\boldsymbol{W})^{-1/2}(\boldsymbol{D}-\rho_0\boldsymbol{W})^{1/2}\boldsymbol{W}\boldsymbol{s} \\
\leq&\boldsymbol{s}^\top \boldsymbol{W}(\boldsymbol{D}-\rho_0\boldsymbol{W})^{-1}\boldsymbol{W}\boldsymbol{s} \leq \lambda_{\min}^{-1}(\boldsymbol{D}-\rho_0\boldsymbol{W})||\boldsymbol{W}\boldsymbol{s}||_2^2 \\
\leq&\lambda_{\min}^{-1}(\boldsymbol{D}-\rho_0\boldsymbol{W})||\boldsymbol{W}||_2^2||\boldsymbol{s}||_2^2 = \lambda_{\min}^{-1}(\boldsymbol{D}-\rho_0\boldsymbol{W})|\lambda(\boldsymbol{W})|_{\max}^2||\boldsymbol{s}||_2^2
\end{aligned}$$

The first inequality is due to Lemma S2. Here $|\lambda(\boldsymbol{W})|_{\max} = ||\boldsymbol{W}||_2$ is the spetrum radius of $\boldsymbol{W}$. Combining the previous steps we obtain the upper bound of the gap in (19). □

S3. Proposition S2 and Its Proof

Proposition S2. *Let $x_1, \ldots, x_n$ of $\boldsymbol{x}$ are independent and identically distributed random variables from the discrete distribution with $\Pr(x_i = 1) = \Pr(x_i = -1) = 0.5$. For any two symmetric and non-zero $n \times n$ matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, we have that*

$$cor_{\boldsymbol{x}}(\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}, \boldsymbol{x}^\top \boldsymbol{B}\boldsymbol{x}) = \frac{\sum_{i<j} a_{ij}b_{ij}}{\sqrt{\sum_{i<j} a_{ij}^2}\sqrt{\sum_{i<j} b_{ij}^2}}, \tag{21}$$

where a_{ij} and b_{ij} are the (i,j)-th entries of matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ respectively.

Consider two $n \times n$ symmetric matrices $\boldsymbol{A}$ and $\boldsymbol{B}$. For random designs, we have that $\mathbb{E}(x_i) = 0$, $\text{var}(x_i) = 1$, and $\text{cov}(x_i, x_j) = 0$ for $i \neq j$. Therefore, $\text{cov}(\boldsymbol{x}) = \boldsymbol{I}_n$ and

$$\begin{aligned}
\text{cov}(\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}, \boldsymbol{x}^\top \boldsymbol{B}\boldsymbol{x}) &= \mathbb{E}(\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}\boldsymbol{x}^\top \boldsymbol{B}\boldsymbol{x}) - \mathbb{E}(\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x})\mathbb{E}(\boldsymbol{x}^\top \boldsymbol{B}\boldsymbol{x}) \\
&= \mathbb{E}\left(\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}\boldsymbol{x}^\top \boldsymbol{B}\boldsymbol{x}\right) - \text{tr}(\boldsymbol{A})\text{tr}(\boldsymbol{B})
\end{aligned}$$

Note that

$$\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}\boldsymbol{x}^\top \boldsymbol{B}\boldsymbol{x} = (\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}) \otimes (\boldsymbol{x}^\top \boldsymbol{B}\boldsymbol{x}) = (\boldsymbol{x}^\top \otimes \boldsymbol{x}^\top)(\boldsymbol{A} \otimes \boldsymbol{B})(\boldsymbol{x} \otimes \boldsymbol{x}).$$

Then

$$\begin{aligned}
\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}\boldsymbol{x}^\top \boldsymbol{B}\boldsymbol{x} &= \text{tr}(\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}\boldsymbol{x}^\top \boldsymbol{B}\boldsymbol{x}) = \text{tr}((\boldsymbol{x}^\top \otimes \boldsymbol{x}^\top)(\boldsymbol{A} \otimes \boldsymbol{B})(\boldsymbol{x} \otimes \boldsymbol{x})) \\
&= \text{tr}((\boldsymbol{A} \otimes \boldsymbol{B})(\boldsymbol{x} \otimes \boldsymbol{x})(\boldsymbol{x}^\top \otimes \boldsymbol{x}^\top)),
\end{aligned}$$

and thus

$$\mathbb{E}(\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}\boldsymbol{x}^\top \boldsymbol{B}\boldsymbol{x}) = \mathbb{E}(\text{tr}(\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}\boldsymbol{x}^\top \boldsymbol{B}\boldsymbol{x})) = \text{tr}((\boldsymbol{A} \otimes \boldsymbol{B})\mathbb{E}((\boldsymbol{x} \otimes \boldsymbol{x})(\boldsymbol{x}^\top \otimes \boldsymbol{x}^\top)))$$

We need to derive $\mathbb{E}((\boldsymbol{x} \otimes \boldsymbol{x})(\boldsymbol{x}^\top \otimes \boldsymbol{x}^\top))$. Note that $(\boldsymbol{x} \otimes \boldsymbol{x})(\boldsymbol{x}^\top \otimes \boldsymbol{x}^\top) = (\boldsymbol{x}\boldsymbol{x}^\top) \otimes (\boldsymbol{x}\boldsymbol{x}^\top)$ is an $n \times n$ block matrix, and the i,j-th block is $x_i x_j \boldsymbol{x}\boldsymbol{x}^\top$. The diagonal blocks are $\mathbb{E}(x_i^2 \boldsymbol{x}\boldsymbol{x}^\top) = \boldsymbol{I}_n$ ($\mathbb{E}(x_i^4) = 1$). If $i \neq j$, $\mathbb{E}(x_i x_j \boldsymbol{x}\boldsymbol{x}^\top) = \boldsymbol{e}_i\boldsymbol{e}_j^\top + \boldsymbol{e}_j\boldsymbol{e}_i^\top$, where $\boldsymbol{e}_i$ is the element vector with i-th entry equal to 1 others 0 and $\boldsymbol{e}_i\boldsymbol{e}_j^\top + \boldsymbol{e}_j\boldsymbol{e}_i^\top$ is a matrix with (i,j)th and (j,i)th entries equal to 1 and the rest entries 0. Therefore, the resulting $n \times n$ block matrix should have diagonal blocks be an $n \times n$ identity matrix, and the (i,j)-th off-diagonal block be $\boldsymbol{e}_i\boldsymbol{e}_j^\top + \boldsymbol{e}_j\boldsymbol{e}_i^\top$. So we can decompose the block matrix to be

$$\begin{aligned}
&\mathbb{E}((\boldsymbol{x} \otimes \boldsymbol{x})(\boldsymbol{x}^\top \otimes \boldsymbol{x}^\top)) = \boldsymbol{I}_n \otimes \boldsymbol{I}_n + \sum_{i \neq j}(\boldsymbol{e}_i\boldsymbol{e}_j^\top) \otimes (\boldsymbol{e}_i\boldsymbol{e}_j^\top + \boldsymbol{e}_j\boldsymbol{e}_i^\top) \\
=&\boldsymbol{I}_n \otimes \boldsymbol{I}_n + \sum_{i \neq j}(\boldsymbol{e}_i\boldsymbol{e}_j^\top) \otimes (\boldsymbol{e}_i\boldsymbol{e}_j^\top) + \sum_{i \neq j}(\boldsymbol{e}_i\boldsymbol{e}_j^\top) \otimes (\boldsymbol{e}_j\boldsymbol{e}_i^\top)
\end{aligned}$$

Then

$$
\begin{aligned}
&\operatorname{tr}\left[(\boldsymbol{A}\otimes\boldsymbol{B})\mathbb{E}[(\boldsymbol{x}\otimes\boldsymbol{x})(\boldsymbol{x}^\top\otimes\boldsymbol{x}^\top)]\right]\\
=&\operatorname{tr}\left[(\boldsymbol{A}\otimes\boldsymbol{B})(\boldsymbol{I}_n\otimes\boldsymbol{I}_n)\right]+\operatorname{tr}\left[\sum_{i\neq j}(\boldsymbol{A}\otimes\boldsymbol{B})[(\boldsymbol{e}_i\boldsymbol{e}_j^\top)\otimes(\boldsymbol{e}_i\boldsymbol{e}_j^\top)]\right]+\operatorname{tr}\left[\sum_{i\neq j}(\boldsymbol{A}\otimes\boldsymbol{B})[(\boldsymbol{e}_i\boldsymbol{e}_j^\top)\otimes(\boldsymbol{e}_j\boldsymbol{e}_i^\top)]\right]\\
=&\operatorname{tr}\left[\boldsymbol{A}\otimes\boldsymbol{B}\right]+\sum_{i\neq j}\operatorname{tr}\left[(\boldsymbol{A}\otimes\boldsymbol{B})[(\boldsymbol{e}_i\boldsymbol{e}_j^\top)\otimes(\boldsymbol{e}_i\boldsymbol{e}_j^\top)]\right]+\sum_{i\neq j}\operatorname{tr}\left[(\boldsymbol{A}\otimes\boldsymbol{B})[(\boldsymbol{e}_i\boldsymbol{e}_j^\top)\otimes(\boldsymbol{e}_j\boldsymbol{e}_i^\top)]\right]\\
=&\operatorname{tr}(\boldsymbol{A})\operatorname{tr}(\boldsymbol{B})+\sum_{i\neq j}\operatorname{tr}\left[(\boldsymbol{A}\boldsymbol{e}_i\boldsymbol{e}_j^\top)\otimes(\boldsymbol{B}\boldsymbol{e}_i\boldsymbol{e}_j^\top)\right]+\sum_{i\neq j}\operatorname{tr}\left[(\boldsymbol{A}\boldsymbol{e}_i\boldsymbol{e}_j^\top)\otimes(\boldsymbol{B}\boldsymbol{e}_j\boldsymbol{e}_i^\top)\right]\\
=&\operatorname{tr}(\boldsymbol{A})\operatorname{tr}(\boldsymbol{B})+2\sum_{i\neq j}\operatorname{tr}\left[(\boldsymbol{A}\boldsymbol{e}_i\boldsymbol{e}_j^\top)\right]\operatorname{tr}\left[(\boldsymbol{B}\boldsymbol{e}_i\boldsymbol{e}_j^\top)\right]\\
=&\operatorname{tr}(\boldsymbol{A})\operatorname{tr}(\boldsymbol{B})+4\sum_{i<j}\boldsymbol{A}_{ij}\boldsymbol{B}_{ij},
\end{aligned}
$$

where $\boldsymbol{A}_{ij}$ is the ij-th entry of matrix $\boldsymbol{A}$. Then

$$\operatorname{cov}(\boldsymbol{x}^\top\boldsymbol{A}\boldsymbol{x},\boldsymbol{x}^\top\boldsymbol{B}\boldsymbol{x})=4\sum_{i<j}\boldsymbol{A}_{ij}\boldsymbol{B}_{ij}$$

Accordingly,

$$\operatorname{cor}(\boldsymbol{x}^\top\boldsymbol{A}\boldsymbol{x},\boldsymbol{x}^\top\boldsymbol{B}\boldsymbol{x})=\frac{\sum_{i<j}\boldsymbol{A}_{ij}\boldsymbol{B}_{ij}}{\sqrt{\sum_{i<j}\boldsymbol{A}_{ij}^2}\sqrt{\sum_{i<j}\boldsymbol{B}_{ij}^2}}$$

S4. Proof of Theorem 2

We first provide a useful Lemma.

Lemma S3. *Let X and Y be two random variables taking values from $\{-1,1\}$. If $\operatorname{cov}(X,Y)=0$, then X and Y are independent.*

Proof. Let U and V be two Bernoulli random variables. We first show that if $\operatorname{cov}(U,V)=0$, then U and V are independent.

Notice that

$$\Pr(\{U=1\}\text{and}\{V=1\})=\Pr(UV=1)=\mathbb{E}(UV)$$
$$\mathbb{E}(U)=\Pr(U=1)$$

and

$$\mathbb{E}(V)=\Pr(V=1).$$

If $\operatorname{cov}(U,V)=0$,

$$\Pr(\{U=1\}\text{and}\{V=1\})-\Pr(U=1)\Pr(V=1)=\mathbb{E}(UV)-\mathbb{E}(U)\mathbb{E}(V)=0.$$

Similarly, we can show that

$$\Pr(\{U=0\}\text{and}\{V=1\})-\Pr(U=0)\Pr(V=1)=0,$$
$$\Pr(\{U=0\}\text{and}\{V=0\})-\Pr(U=0)\Pr(V=0)=0,$$

and

$$\Pr(\{U=1\}\text{and}\{V=0\}) - \Pr(U=1)\Pr(V=0) = 0,$$

which demonstrate that U and V are independent.

For X and Y, we have that $X = 2U - 1$ and $Y = 2V - 1$. The independence of U and V indicates the independence of X and Y. Also,

$$\text{cov}(X, Y) = 4\text{cov}(U, V).$$

Thus, the conclusion holds. □

Proof. Recall that $w_{ii} = 0$ for $i = 1, \ldots, n$. Therefore, we only need to consider the terms $w_{ij}x_ix_j$ with $i \neq j$. Notice that

$$\text{cov}(x_ix_j, x_{i'}x_{j'}) = \mathbb{E}(x_ix_jx_{i'}x_{j'}) - \mathbb{E}(x_ix_j)\mathbb{E}(x_{i'}x_{j'}) = 0$$

for $i \neq i'$ and $j \neq j'$. Also,

$$\text{cov}(x_ix_j, x_ix_{j'}) = \mathbb{E}(x_i^2x_jx_{j'}) - \mathbb{E}(x_ix_j)\mathbb{E}(x_ix_{j'}) = 0$$

for $j \neq j'$. According to Lemma S3, we have that x_ix_j and $x_ix_{j'}$ are independent, and x_ix_j and $x_{i'}x_{j'}$ are independent. Thus, $w_{ij}x_ix_j$'s with $w_{ij} \neq 0$ are i.i.d random variables with mean

$$\mathbb{E}(w_{ij}x_ix_j) = \mathbb{E}(x_i)\mathbb{E}(x_j) = 0,$$

and variance

$$\text{var}(w_{ij}x_ix_j) = \mathbb{E}(x_i^2x_j^2) - (\mathbb{E}(x_ix_j))^2 = 1.$$

According to the central limit theorem, the conclusion holds. □

S5. Proof of Proposition 1

Proof. Notice that

$$\mathbb{E}\left(\boldsymbol{x}^\top\boldsymbol{K}\boldsymbol{x}\right) = \text{tr}\left[\mathbb{E}\left(\boldsymbol{x}^\top\boldsymbol{K}\boldsymbol{x}\right)\right] = \mathbb{E}\left[\text{tr}(\boldsymbol{x}^\top\boldsymbol{K}\boldsymbol{x})\right] = \mathbb{E}\left[\text{tr}(\boldsymbol{K}\boldsymbol{x}\boldsymbol{x}^\top)\right] = \text{tr}\left[\boldsymbol{K}\mathbb{E}(\boldsymbol{x}\boldsymbol{x}^\top)\right].$$

For completely random design, under the same assumption as in Theorem 2, we have that

$$\mathbb{E}(x_ix_j) = \mathbb{E}(x_i)\mathbb{E}(x_j) = 0 \quad \text{for} \quad i \neq j$$

and $\mathbb{E}(x_i^2) = 1$ for $i = 1, \ldots, n$. Thus, $\mathbb{E}(\boldsymbol{x}\boldsymbol{x}^\top) = \boldsymbol{I}_n$.

Now we consider the case where $\boldsymbol{x}$ is a random balanced design. If n is even, we have that

$$\mathbb{E}\left(x_i\sum_{j=1}^n x_j\right) = 0$$

since the balanced constraint gives $\sum_{j=1}^n x_j = 0$ directly. If n is odd, $n = 2h + 1$ with h be a positive integer. Due to the balance constraint, $\sum_{i=1}^n x_i = 1$ or -1. We have that

$$\mathbb{E}\left(x_i\sum_{j=1}^n x_j\right) = \Pr\left(\sum_{i=1}^n x_i = 1\right)\mathbb{E}\left(x_i\sum_{j=1}^n x_j\Big|\sum_{i=1}^n x_i = 1\right) + \Pr\left(\sum_{i=1}^n x_i = -1\right)\mathbb{E}\left(x_i\sum_{j=1}^n x_j\Big|\sum_{i=1}^n x_i = -1\right)$$

$$= \frac{1}{2}\mathbb{E}\left(x_i\Big|\sum_{j=1}^n x_j = 1\right) - \frac{1}{2}\mathbb{E}\left(x_i\Big|\sum_{j=1}^n x_j = -1\right).$$

Note that

$$\mathbb{E}\left(x_i \middle| \sum_{i=1}^{n} x_i = 1\right) = \Pr\left(x_i = 1 \middle| \sum_{i=1}^{n} x_i = 1\right) - \Pr\left(x_i = -1 \middle| \sum_{i=1}^{n} x_i = 1\right) = \frac{h+1}{2h+1} - \frac{h}{2h+1} = \frac{1}{n},$$

$$\mathbb{E}\left(x_i \middle| \sum_{j=1}^{n} x_j = -1\right) = \Pr\left(x_i = 1 \middle| \sum_{j=1}^{n} x_j = -1\right) - \Pr\left(x_i = -1 \middle| \sum_{j=1}^{n} x_j = -1\right) = \frac{h}{2h+1} - \frac{h+1}{2h+1} = -\frac{1}{n}.$$

Thus, $\mathbb{E}\left(x_i \sum_{j=1}^{n} x_j\right) = 1/n$.

Therefore,

$$\mathbb{E}\left(x_1 \sum_{j=1}^{n} x_j\right) = 1 + (n-1)\mathbb{E}(x_1 x_2)$$

which gives that

$$\mathbb{E}(x_1 x_2) = \begin{cases} -\frac{1}{n-1} & \text{if } n \text{ is even} \\ -\frac{1}{n} \quad if & \text{if } n \text{ is odd} \end{cases}.$$

This conclusion holds for $\mathbb{E}(x_i x_j)$ with any $i \neq j$. □