Unconditional Randomization Tests for Interference

Liang Zhong*

September 10, 2024

Abstract

In social networks or spatial experiments, one unit's outcome often depends on another's treatment, a phenomenon called *interference*. Researchers are interested in not only the presence and magnitude of interference but also its evolution based on factors like distance, neighboring units, and connection strength. However, the non-random nature of these factors and complex correlations across units pose challenges for inference. This paper introduces the *Partial Null Randomization Testing (PNRT)* framework to address these issues. The proposed method is finite-sample valid and applicable without network structure assumptions, utilizing randomization testing and pairwise comparisons. Unlike existing *Conditional Randomization Tests*, PNRT avoids the need for conditioning events, making it more straightforward to implement. Simulations demonstrate the method's desirable power properties and its applicability to general interference scenarios.

JEL Classification: C0, C5.
Keywords: Causal inference, Non-sharp null hypothesis, Dense network.

*Department of Economics, Boston University, 270 Bay State Road, Boston, MA 02215 USA. Email: samzl@bu.edu, Website:https://samzl1.github.io/.

I am indebted to my advisors, Hiroaki Kaido and Jean-Jacques Forneron, for their continuous guidance and support. I would like to thank M. Daniele Paserman, Ivan Fernandez-Val, Pierre Perron, Marc Rysman, Xunkang Tian, Stella Hong, Zhongjun Qu, and the participants of the econometrics seminar at Boston University for their useful comments. All errors are my own.

1 Introduction

In social networks or spatial experiments, the outcome for one unit often depends on the treatment assigned to another, a phenomenon known as *interference*.[1] Researchers are not only interested in the existence and magnitude of such interference but also in how it evolves within a network.[2] For instance, Blattman et al. (2021) examines a large-scale experiment in Bogotá, Colombia, focusing on the impact of a hotspot policing policy on crime by treating each street segment as the unit of analysis. To assess the policy's total welfare impact, it is crucial to evaluate whether interference occurred following treatment assignment, such as crime displacement or deterrence in nearby neighborhoods.[3] Additionally, Blattman et al. (2021) tests the distance of spillover effects to determine which units are sufficiently far from treated ones to serve as a control group in the analysis. As noted by Blattman et al. (2021), standard errors are often underestimated due to complex clustering patterns, suggesting that a design-based approach and randomization inference might be more appropriate in many network settings where the nature of spillovers is unknown.[4] Consequently, recent studies, including Blattman et al. (2021), leverage the *Fisher Randomization Tests(FRT)* for inference, as it is exact in finite samples (Fisher, 1925). However, classical FRT is not guaranteed to be valid when testing for interference (Athey et al., 2018).

In this paper, I introduce a novel unconditional randomization testing framework called *Partial Null Randomization Testing (PNRT)*, designed to detect interference and analyze its evolution within networks. This nonparametric method is finite-sample valid and applicable without requiring assumptions about the network structure, relying solely on the randomness of treatment assignments.[5] Given its robustness, I propose PNRT as a benchmark for network analysis.

An essential concept in implementing the FRT is *imputability*: all potential outcomes are

[1] For example, Bayer et al. (2008) explore the effect of social interactions on labor market outcomes, while Angrist (2014) examine the peer effect on students' achievement.

[2] For example, Bond et al. (2012) investigates whether spillover effects extend beyond users' immediate friends to their friends of friends. Some theoretical work, such as Toulis and Kao (2013), would ex-ante assume away the spillover effects of friends of friends. Rajkumar et al. (2022) studies how job mobility relates to the intensity of links, differentiating between strong and weak ties.

[3] This assumes that interactions pass through neighboring units, resulting in spillover effects.

[4] Blattman et al. (2021) P.2027: "Many urban programs are both place-based and vulnerable to spillovers. This includes efforts to improve traffic flow, beautify blighted streets and properties, foster community mobilization, and rezone land use. The same challenges could arise with experiments in social and family networks."

[5] It is finite sample *exact* in the sense that its probability of false rejection in finite samples will not exceed the user-prescribed nominal probability (Pouliot, 2024).

observed across different treatment assignments under the null (Rosenbaum, 2007; Hudgens and Halloran, 2008). However, testing for interference involves *partially sharp null hypotheses*, which introduce two primary challenges. First, only a subset of potential outcomes is imputable. Second, the set of units with imputable outcomes varies with different treatment assignments. For instance, under the null hypothesis of no spillover effects on streets not treated by hotspot policing, there is no information about treated streets, and the set of streets experiencing spillover effects varies with each treatment assignment.

To address the first challenge, I propose *pairwise imputable statistics*, a bivariate function $T(D^{obs}, D)$ involving the observed assignment D^{obs} and the randomized assignments D. It is restricted to only imputable units under both observed and randomized assignments under the null hypothesis while resembles conventional test statistics defined by Imbens and Rubin (2015) .[6] The critical difference lies in the role of the observed assignment D^{obs}: it not only determines the values of the outcome vector, as in conventional test statistics, but also identifies the set of units that are imputable under D^{obs}. Despite this seeming restriction, I demonstrate that pairwise imputable statistics can accommodate various test statistics commonly used in sharp null hypotheses. For example, in the difference-in-means estimator, we compare the spillover group with the control group, where D^{obs} determines both the units included in the computation and the values of the outcome variable, while D determines group assignments, implicitly excluding treated units under D.

However, the second challenge—variation in the set of imputable units—makes it difficult to guarantee validity when directly using pairwise imputable statistics in FRT. Specifically, p-values are constructed within the fixed set of imputable units following D^{obs}, comparing the observed test statistics $T(D^{obs}, D^{obs})$ with other randomized test statistics $T(D^{obs}, D)$. Nevertheless, $T(D^{obs}, D^{obs})$ belongs to the same distribution as $T(D, D)$, which differs from $T(D^{obs}, D)$ even under the null due to the variation in the set of imputable units across different treatment assignments. This variability makes naive implementations of unconditional randomization testing unable to control size effectively.

To overcome this, I draw inspiration from recent advances in selective inference (Wen et al., 2023; Guan, 2023) and construct PNRT p-values through *pairwise inequality comparisons* between $T(D, D^{obs})$ and $T(D^{obs}, D)$ for each pair of observed and potential assignments (D^{obs}, D). Since pairwise imputable statistics use only imputable units under both observed and randomized assignments, both terms are computable under the partial null hypothesis.

[6] The test statistics often depend on the observed outcome and randomized assignment. As Imbens and Rubin (2015) noted, given the potential outcome function, the observed outcome is a function of the observed assignment.

The validity of the procedure is established through the symmetry of these pairwise comparisons, similar to the conformal lemma in Guan (2023). I propose two types of PNRT: *pairwise comparison-based PNRT* and *minimization-based PNRT*. Theoretically, the pairwise comparison-based PNRT controls type I error below α when the rejection level is $\alpha/2$, and the minimization-based PNRT controls the type I error at the rejection level α. Both methods rely only on the randomness in the treatment assignment and are valid under arbitrary fixed designs and network structures. Moreover, in the case of a sharp null hypothesis, $T(D, D^{obs})$ equals $T(D^{obs}, D^{obs})$ and $T(D^{obs}, D)$ equals $T(D, D)$, so both PNRT procedures encompass FRT as a special case under sharp null hypotheses. Additionally, a multiple hypothesis testing adjustment procedure ensures *Family Wise Error Rate (FWER)* control when defining the "neighborhood" of interference concerning distance measures or tie strengths.

To illustrate PNRT's applicability, I revisit Blattman et al. (2021), which reported significant spillover effects on property crime but not violent crime. A simulation study calibrated to the actual dataset demonstrates PNRT's desirable power properties and its suitability for general interference scenarios. Regarding size control, the pairwise comparison-based PNRT empirically controls type I errors, even at the rejection level α, suggesting that the theoretical result provides a guarantee in the worst-case scenario. Conversely, the classical FRT method tends to over-reject under partial null hypotheses. As for power, the pairwise comparison-based PNRT with a rejection level of α demonstrates superior power compared to alternatives and maintains desirable power even at a rejection level of $\alpha/2$. Moreover, PNRT's reanalysis suggests that contrary to Blattman et al. (2021), the spillover effect might be significant at the 10% level for violent crime, while effects for property crime may be insignificant. This finding could potentially alter the welfare analysis if violent crime is deemed more severe and in need of stricter control.

This paper contributes to two strands of literature. First, it advances causal inference under interference. Unlike model-based approaches that rely on parametric assumptions (Sacerdote, 2001; Bowers et al., 2013; Toulis and Kao, 2013), this work aligns with the randomization-based method (also called *design-based inference*), which uses treatment assignment randomness as the source of uncertainty for inference, treating all potential outcomes as fixed constants (Abadie et al., 2020, 2022). Within the randomization-based method, there are at least two inferential frameworks for causal inference with interference: the Fisherian and Neymanian perspectives (Li et al., 2018). The Neymanian approach focuses on randomization-based unbiased estimation and variance calculation (Hudgens and Halloran, 2008; Aronow and Samii, 2017; Pollmann, 2023), with inference and interval estimation based on normal approximations in asymptotic settings, often requiring sparse networks or local interference.[7]

In contrast, this paper follows the Fisherian perspective, focusing on detecting causal effects with finite-sample exact randomization-based testing (Dufour and Khalaf, 2003; Lehmann and Romano, 2005; Rosenbaum, 2020). Acknowledging FRT's invalidity for testing interference, prior literature has proposed *Conditional Randomization Testing (CRT)*, which restricts the test to a conditioning event involving a subset of units and assignments where the null hypothesis is sharp.[8] Different papers have suggested various procedures for designing these conditioning events to ensure finite-sample exact testing. However, many CRT methods are tailored to specific circumstances, such as clustered interference (Basse et al., 2019, 2024), and cannot be extended to more general settings. Additionally, designing conditioning events to ensure nontrivial power is challenging, often leading to power loss (Puelz et al., 2021). Lastly, conditioning events for general interference are computationally demanding, typically requiring extensive time for implementation. The main contribution of this paper is developing a valid testing procedure free from conditioning events for testing partially sharp null hypotheses. This procedure offers three key advantages: broad applicability, avoidance of complex conditioning events, and straightforward implementation. A simulation study with spatial interference calibrated to Blattman et al. (2021) illustrates PNRT's superior power to CRT, which involves complex conditioning events that restrict power. This advantage is precious given the high cost of collecting information for each unit in network analysis and the often minimal interference effects (Taylor and Eckles, 2018; Breza et al., 2020).

As illustrated in previous literature, such as Athey et al. (2018) and Basse et al. (2019), the confidence intervals for certain causal parameters are constructed by inverting tests. As noted by Basse et al. (2024), this approach provides finite-sample exact tests with minimal model assumptions compared to model-based approaches. Additionally, randomization-based methods can be combined with model-based frameworks, such as the linear-in-means model (Manski, 1993), to potentially increase power or extend the framework beyond random experiments while ensuring test validity (Wu and Ding, 2021; Basse et al., 2024; Borusyak and Hull, 2023).

Second, beyond the network setting, this method extends randomization testing to any partial sharp null hypothesis. Since Neyman et al. (2018) acknowledged the limitation of FRT for testing only sharp null hypotheses, researchers have developed various strategies for

[7] See also, Basse and Airoldi (2018); Viviano (2022); Wang et al. (2023); Vazquez-Bare (2023)

[8] See, for example, Aronow (2012); Athey et al. (2018); Basse et al. (2019); Puelz et al. (2021); Zhang and Zhao (2021); Basse et al. (2024); Hoshino and Yanagi (2023)

different types of weak nulls. For example, Ding et al. (2016), Li et al. (2016), and Zhao and Ding (2020) investigate the null hypothesis of no average treatment effect; Caughey et al. (2023) validate randomization testing for certain classes of test statistics under bounded nulls; Zhang and Zhao (2021) construct conditional randomization testing for partial sharp null following the similar idea as Athey et al. (2018) and Puelz et al. (2021), applying this idea in time-staggered adoption designs. To my knowledge, PNRT is the first procedure to address partial null hypotheses using unconditional testing.

Structure of the paper. First, Section 2 introduces the general setup and establishes all necessary notation. Then, Section 3 presents the PNRT procedure, which includes the pairwise imputable statistics (Section 3.1) and the p-value based on pairwise comparisons (Section 3.2). Section 4 proposes a framework for determining the boundary of interference and adjusting for sequential testing. Next, Section 5 applies the method to a large-scale policing experiment in Bogotá, Colombia, with Section 5.1 reporting the results of a Monte Carlo experiment calibrated to this setting. Finally, Section 6 concludes the paper. The Appendix provides additional empirical and theoretical results, as well as the proofs.

2 Setup and null hypothesis of interest

Consider N units with index $i \in \{1, 2, ..., N\}$, and a treatment assignment vector $D = (D_1, \ldots, D_N) \in \{0, 1\}^N$, where $D_i \in \{0, 1\}$ denote unit i's treatment. Let X be the collected pre-treatment characteristics, such as age and gender. They can be used to control for units' heterogeneity, and I do not attempt to evaluate their effects on the outcome. The treatment assignment is random and follows a known probability distribution $P(D)$ where $P(d) = pr(D = d)$ is the probability of the assignment D taking on the value d. The probability distribution may or may not depend on covariates X: it doesn't depend on X when we have a complete randomization or cluster randomization; it depends on X when we have a stratified randomization design or matched-pair design. Let $Y(d) = (Y_1(d), \ldots, Y_N(d)) \in R^N$ be the potential outcome when the treatment assignment is d, where potential outcome of unit i under assignment d is $Y_i(d)$. I allow unit i's potential outcome to depend on another unit j's treatment assignments, which allows violation of the classic SUTVA proposed by Cox (1958) and enables us to consider situations when spatial/network interference exists.

Throughout the paper, I assume the following objects are observed: 1) the realized vector of treatments for all units in the network, denoted by D^{obs}; 2) the realized outcomes for all of the units, denoted $Y^{obs} \equiv Y(D^{obs}) = (Y_1(D^{obs}), \ldots, Y_N(D^{obs}))$; 3) The $N \times N$ proximity matrix G, where the (i, j)-th component $G_{i,j} \geq 0$, represents a "distance measure" between units i and j, which allowed to be either a continuous or discrete variable. I normalize $G_{i,i} = 0$ for all $i = 1, 2, \ldots, N$, and $G_{i,j} > 0$ for all $i \neq j$. This measure would be context-specific:[9]

Example 1 (Spatial Distance)**.** *Consider settings where units interact locally through shared space, such as street segments in a city in Blattman et al. (2021). $G_{i,j}$ would be the spatial distance between units i and j.*

Example 2 (Network Distance)**.** *Consider settings where units are linked in a social network, such as friends in Facebook in Bond et al. (2012). $G_{i,j}$ measures the distance between units i and j, such that $G_{i,j} = 1$ for friends, $G_{i,j} = 2$ for friend s of friends, etc. $G_{i,j} = \infty$ if i and j are not connected to accommodate the case with disconnected networks and the interest in partial interference, such as cluster interference (Sobel, 2006; Basse et al., 2019).*

Example 3 (Intensity of the Link)**.** *Researchers might not only observe whether two units are linked but also the intensity of the link $int_{i,j}$ between units i and j, such as frequency of interaction or volume of email correspondence (Goldenberg et al. (2009); Bond et al. (2012); Rajkumar et al. (2022)). Following the classic study from Granovetter (1973), one might be interested in how interference differs across the weak and strong ties defined by the intensity measure. Here, denote $i\bar{n}t = max_{i,j \in \{1,\ldots,N\}} int_{i,j}$, one option is to define $G_{i,j} = i\bar{n}t - int_{i,j}$. So, the increase of $G_{i,j}$ implies a weaker connection as in Example 1 and Example 2.*

I adopt a design-based inference approach where D is treated as random, but G, X, P and the potential outcome schedule $Y(\cdot)$ are taken as fixed. For simplicity in notation, they are not treated as arguments of any functions in the rest of the paper.

2.1 Sharp null and partial null hypothesis

In Fisher (1925), randomization testing is introduced with the *sharp null hypothesis*. Formally defined as the following:

Definition 1 (Sharp null hypothesis)**.**

$$H_0 : Y_i(d) = Y_i(d') \text{ for all } i \in \{1, \ldots, N\}, \text{ and any } d, d' \in \{0, 1\}^N$$

[9] Similar to Pollmann (2023), my method can also accommodate the non-network settings. For example, we can consider firms selling differentiated products and define the distance measure as the distance in the product space.

This hypothesis represents a classic null hypothesis, often called the null hypothesis of no treatment effect whatsoever. Under this hypothesis, *all* missing potential outcomes are "observed" (Zhang and Zhao, 2023). This sharp null hypothesis allows the classical Fisher randomization tests to be assessed, as all potential outcomes can be imputed under the null. FRT involves randomly reassigning treatments D to units, calculating the test statistic for each reassignment, and comparing these statistics to the observed value to determine significance. The p-value is constructed as the proportion of D such that the coefficient is higher than the observed coefficient. A detailed discussion can be found in Appendix A.1.

However, this strong null hypothesis is unreasonable in certain cases. Hence, the *partial null hypothesis* is introduced, allowing the potential outcome to differ for certain assignment vectors. Formally defined as follows:

Definition 2 (Partial null hypothesis)**.**

$$H_0 : Y_i(d) = Y_i(d') \textit{ for all } i \in \{1, \dots, N\}, \textit{ and any } d, d' \in \mathcal{D}_i \subsetneq \{0,1\}^N$$

Notice that the set $\mathcal{D}_i$ changes with each i, and it is strictly a subset of $\{0,1\}^N$. As noted in Zhang and Zhao (2023), the missing potential outcomes are only *partially* "observed" under the partial null hypothesis. The sharp null hypothesis in Definition 1 can be thought of as a special case that $\mathcal{D}_i = \{0,1\}^N$ for any i. When considering the existence or evolution of interference within a network, researchers are often interested in whether there is an interference beyond a certain distance ϵ_s that $d_i = 1$ is excluded so that these sets only consider strictly spillover effects, whereas Definition 1 also includes $d_i = 1$.

Definition 3 (Distance interval assignment set)**.** *For unit $i \in \{1, \dots, N\}$, and given distance ϵ_s, distance interval assignment set is*

$$\mathcal{D}_i(\epsilon_s) \equiv \{d \in \{0,1\}^N : \sum_{j=1}^{N} 1\{G_{i,j} \leq \epsilon_s\} d_j = 0\}$$

and given any $d \in \mathcal{D}_i(\epsilon_s)$, I call unit i is in the distance interval (ϵ_s, ∞).

This set includes all treatment assignments where unit i is at least a distance ϵ_s away from any treated units, which I refer to as being "in the distance interval (ϵ_s, ∞) from any treated units." For any $\epsilon_s \geq 0$, since $G_{i,i} = 0$, we have $1\{G_{i,i} \leq \epsilon_s\} = 1$, which implies that $d_i = 0$ in $\mathcal{D}_i(\epsilon_s)$. Specifically, when $\epsilon_s = 0$, $\mathcal{D}_i(0)$ includes all the treatment assignments d such that $d_i = 0$.

For any $\epsilon_s < 0$, given that $G_{i,j} \geq 0$ for all i, j, we have $1\{G_{i,j} \leq \epsilon_s\} = 0$. Therefore, $\mathcal{D}_i(\epsilon_s) = \{0,1\}^N$. If we instead focus on treatment assignments where unit i is within the distance interval $(a, b]$, the corresponding set can be written as $\mathcal{D}_i(a)/\mathcal{D}_i(b)$. Although the method introduced in this paper can generally apply to all partial null hypotheses, the rest of the paper will specifically focus on the following special case of the partial null hypothesis:

Definition 4 (Partial null hypotheses of interference on distance $\epsilon_s \geq 0$)**.**

$$H_0^{\epsilon_s} : Y_i(d) = Y_i(d') \textit{ for all } i \in \{1, \dots, N\}, \textit{ and any } d, d' \in \mathcal{D}_i(\epsilon_s)$$

In other words, this partial null hypothesis asserts that no interference exists beyond distance ϵ_s, where the meaning of distance is context-specific. If $\epsilon_s = 0$, we would test the partial null hypothesis of no interference since $\mathcal{D}_i(0)$ includes all treatment assignments such that $d_i = 0$. This could serve as an alternative to cluster robust standard errors for conducting inference on interference in practice. If $\epsilon_s > 0$, researchers can use this approach to identify the neighborhood of interference or to find a safe comparison group for a later estimation step.

Example 1 (Spatial Distance (cont.))**.** *When units are street segments, for some spatial distance ϵ_s (e.g., 500 meters), $\mathcal{D}_i(\epsilon_s)$ represents the set of treatment assignments where unit i is 500 meters away from any treated street segments. The partial null hypothesis of interference $H_0^{\epsilon_s}$ would test whether a spillover effect exists on an untreated unit 500 meters away from any treated units.*

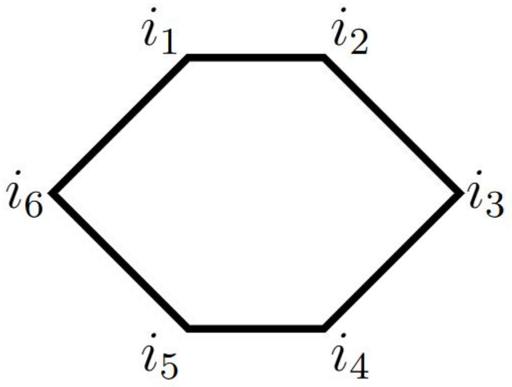
Example 2 (Network Distance (cont.))**.** *Suppose two schools are far apart, with 100 students in each school, and we are interested in the cluster interference of the treatment within the schools, such as the effect of deworming drugs as in Miguel and Kremer (2004). Suppose we don't have access to student-level linkage and are interested in cluster interference within each school. We can consider students as units with a distance of 100 to all other students in the same school and a distance of ∞ to students in the other school. Then, set $\epsilon_s = 0$ to test the existence of cluster interference.*[10]

Example 3 (Intensity of the Link (cont.))**.** *When units are people with cell phones, and we observe the number of text messages between different units, with a maximum of, for example, 50 messages per week, we can construct the "distance" measurement reflecting the*

[10] One might be interested in setting $\epsilon_s = 101$ to test interference across schools. However, as noted by Puelz et al. (2021), such a test might not be feasible due to a lack of power in practice.

intensity of the link as 50 minus the number of messages between units. Then, for a distance $\epsilon_s = 40$, $\mathcal{D}_i(\epsilon_s)$ represents the set of treatment assignments where unit i has fewer than $50 - 40 = 10$ messages with any treated units. The partial null hypothesis of interference $H_0^{\epsilon_s}$ would test whether interference exists for an untreated unit with fewer than 10 messages with any treated units.

A toy example. Consider 6 students as the units connected in a regular polygon, as shown in Figure 1. The units are connected if they are friends with each other. Suppose the outcome of interest, Y, is the number of hours they spend studying each day, and there is a random treatment D as a new version of the textbooks. Only one unit is treated randomly, with $P(d) = 1/6$ for each potential assignment. Let $D^{obs} = (1, 0, 0, 0, 0, 0)$ and $Y^{obs} = (2, 5, 3, 1, 4, 6)$.



$$\Rightarrow \quad G = \begin{pmatrix} 0 & 1 & 2 & 3 & 2 & 1 \\ 1 & 0 & 1 & 2 & 3 & 2 \\ 2 & 1 & 0 & 1 & 2 & 3 \\ 3 & 2 & 1 & 0 & 1 & 2 \\ 2 & 3 & 2 & 1 & 0 & 1 \\ 1 & 2 & 3 & 2 & 1 & 0 \end{pmatrix}$$

Figure 1: Left: Structure of the six units in the toy example. Right: Distance matrix for the toy example.

Hence, when unit i_1 is treated, units i_2 and i_6 are the direct friends of the treated unit, and units i_3 and i_5 are the friends of friends of the treated unit. If, instead, unit i_2 is treated, units i_1 and i_3 are the direct friends of the treated unit, and units i_4 and i_6 are the friends of friends of the treated unit.

If researchers find it possible to have a peer effect and would like to test the existence of it by the partial null hypothesis in Definition 4 with $\epsilon_s = 0$:

$$H_0^0 : Y_i(d) = Y_i(d') \textit{ for all } i \in \{1, \ldots, N\}, \textit{ and any } d, d' \in \{0, 1\}^N \textit{ such that } d_i = d'_i = 0$$

As a partial null hypothesis, the potential outcome schedule aligns with the scenario described in Table 1, where certain potential outcomes may be missing values.

Table 1: Potential outcome schedule under Partial Null

Assignment D	Potential Outcome Y_i					
	i_1	i_2	i_3	i_4	i_5	i_6
$(1,0,0,0,0,0)$	2	5	3	1	4	6
$(0,1,0,0,0,0)$	?	?	3	1	4	6
$(0,0,1,0,0,0)$	?	5	?	1	4	6
$(0,0,0,1,0,0)$	?	5	3	?	4	6
$(0,0,0,0,1,0)$	?	5	3	1	?	6
$(0,0,0,0,0,1)$	?	5	3	1	4	?

Legend: Potential outcome schedule with the partial null hypothesis under Definition 4 for the toy example: Assignment D includes all the potential assignments with the first row as the observed assignment D^{obs}; Potential outcomes are with ? are non-imputable values under the partial null.

A natural idea is to consider a subset of units or assignments for which imputation is possible. However, let's fix the subset of units as all the units with imputable potential outcomes under the null given the observed treatment assignment D^{obs}. There is no guarantee that those units would still be ϵ_s distance away from any treated units. For example, as we can see in table 1, all units i_2 to i_6 are control units under D^{obs} when i_1 is treated, and there is no other potential assignment d that could keep i_2 to i_6 in the control group. Thus, the partial null hypothesis induces a technical barrier when using randomization testing, and naively choosing a *fixed subset* of units wouldn't work.

2.2 Preview of the PNRT procedures for testing the existence of interference

Consider the setting described in Blattman et al. (2021), where we observe a treatment assignment D^{obs}. Suppose we plan to run a regression of the number of crimes on a spillover proximity indicator that the units are within 250 meters of any treated unit. This proximity indicator would be related to the treatment assignment D^{obs}. We may include additional covariates in the regression, but the test statistic of interest will be the coefficient on the indicator.

In the following section, I introduce two *Partial Null Randomization Testing (PNRT)* procedures designed to perform inference for the partial null hypothesis: the pairwise comparison-based PNRT and the minimization-based PNRT. Both methods share a common set of steps:

Step 1. Randomly reassign treatments D to units. For each reassignment D, identify the *subsample* of units that would not be treated under either D^{obs} or D.

Step 2. Using the observed outcome in the *subsample*, run the first regression as if D^{obs} were the treatment assignment to obtain the regression coefficient β. Then, run a second regression as if D were the treatment assignment to obtain the regression coefficient β'.

Step 3. P-value calculation:

For the pairwise comparison-based PNRT, calculate the p-value as the proportion of reassignments D for which $\beta' > \beta$.

For the minimization-based PNRT, first find the minimum value of β across all reassignments D, denoted $\tilde{\beta}$. Then, calculate the p-value as the proportion of reassignments D for which $\beta' \geq \tilde{\beta}$.

The subsequent sections of the paper provide a detailed discussion of why these two procedures lead to valid hypothesis testing.

3 Two types of Partial null randomization testing

3.1 Pairwise imputable statistics

Given some missing potential outcomes, the first technical challenge is constructing the test statistic. In practice, researchers typically have a distance ϵ_c in mind such that any units with distance ϵ_c away from the treated units would not be affected by interference. For example, in a spatial setting, we might be confident that there is no interference if units are $\epsilon_c = 1000$ meters away from any treated units. If we are interested in cluster interference, we might be confident that there is no spillover once ϵ_c is larger than any distance within each cluster, meaning there is no interference across different clusters.

Thus, a natural choice of test statistics would involve a comparison between units in the distance interval $(\epsilon_s, \epsilon_c]$ from treated units replacing the treated group in the class test statistic and the units in the distance interval (ϵ_c, ∞) from treated units as a pure control group. If the researcher doesn't have such a distance ϵ_c in mind, section 4.2 proposes a sequential testing procedure to help select the appropriate ϵ_c. In fact, even if ϵ_c is misspecified and doesn't offer a clean comparison group for interference, the proposed testing procedure remains valid, although it might affect the power of the test.

As illustrated above, using a fixed subset of units is not ideal, especially when we have different units that are imputable under $H_0^{\epsilon_s}$ for different D^{obs}.Therefore, it is essential to pay special attention only to units imputable under $H_0^{\epsilon_s}$ given our observed information. For notational simplicity, let's fix ϵ_s and ϵ_c for the rest of the section.

Definition 5 (Imputable Units). *Given $d \in \{0,1\}^N$ and partial null hypothesis $H_0^{\epsilon_s}$, $\mathbb{I}(d) \equiv \{i \in \{1,\dots,N\} : d \in \mathcal{D}_i(\epsilon_s)\} \in 2^{\{1,\dots,N\}}$ is called Imputable Units under treatment assignment d.*

When we are interested in the sharp null H_0 instead of the partial null $H_0^{\epsilon_s}$, based on Definition 1, we can treat $\mathcal{D}_i$ as $\{0,1\}^N$. As a result, $\mathbb{I}(d) = \{1,\dots,N\}$ for any d under the null H_0, which makes sense as all the units would be imputable under the sharp null. In general, given $H_0^{\epsilon_s}$ and observed treatment D^{obs}, $\mathbb{I}(D^{obs})$ contains all the units we can use for testing: any outcome from units outside this set would not add further information to the test because the potential outcome we observe is not imputable under the partial null hypothesis. For example, if $\epsilon_s = 0$, $\mathcal{D}_i(\epsilon_s)$ would contain all the assignment d such that $d_i = 0$. Hence, $\mathbb{I}(D^{obs})$ would include all the units that were not treated under D^{obs}. In general, $\mathbb{I}(d) \neq \mathbb{I}(d')$ for any $d \neq d'$. For example, when testing the existence of spillover effects among friends, different people have different friends, so the set of "friends" of the treated people would be different for different treatment assignments. In practice, it could be the case that the set $\mathbb{I}(D^{obs})$ is empty, which is generally determined by the structure of the network and the partial null hypothesis of interest. If no units satisfy these criteria, then I would recommend not rejecting the null hypothesis at all.

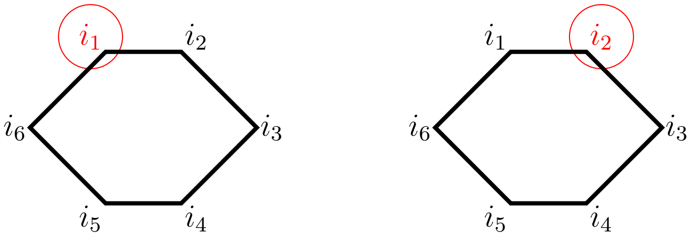


Figure 2: Left: Unit i_1 is treated and marked in red. Units i_2 to i_6 are all imputable and marked as black. Right: Unit i_2 is treated and marked in red. Units i_1, i_6 to i_3 are all imputable and marked as black.

A toy example (cont.) Under H_0^0, the *imputable units* for d can be written as $\mathbb{I}(d) \equiv \{i \in \{1,\dots,N\} : d_i = 0\} \in 2^{\{1,\dots,N\}}$. Specifically, as shown in Figure 2, when unit i_1 is treated, all units i_2 to i_6 belong to the imputable units set; when unit i_2 is treated, unit i_1 and units i_3 to i_6 would belong to the imputable unit set. For the rest of the discussion in the toy example, I would use $\epsilon_c = 1$.

To help define the test statistics later on, we need to define:

Definition 6 (Imputable Outcome Vector)**.** *For $d, d' \in \{0,1\}^N$ and partial null hypothesis $H_0^{\epsilon_s}$, $Y_{\mathbb{I}(d)}(d') \equiv \{Y_i(d')\}_{i \in \mathbb{I}(d)}$ is called the imputable Outcome Vector defined on imputable units for d with the potential outcome value under the treatment assignment d'.*

For the potential outcome vector $Y(d')$ given the treatment assignment d', $Y_{\mathbb{I}(d)}(d')$ is a subvector of it, and the units that are included are determined by d. If the null hypothesis is a sharp null, as we illustrated before $\mathbb{I}(d) = \{1, \ldots, N\}$, then $Y_{\mathbb{I}(d)}(d') = Y(d')$. Due to the partial null, different d implies a different set of units in the imputable outcome vector, and when $d' = D^{obs}$, $Y(d') = Y^{obs}$. This allows us further to define our core idea on the test statistics:

Definition 7 (Pairwise Imputable Statistics)**.** *Let $T : R^N \times \{0,1\}^N \times \{0,1\}^N \to R \cup \{\infty\}$ be a measurable function, and given partial null hypothesis $H_0^{\epsilon_s}$. T is said to be the pairwise imputable statistics if $T(Y_{\mathbb{I}(d)}(d), d') = T(Y_{\mathbb{I}(d)}(d'), d')$, for any $d, d' \in \{0,1\}^N$ such that $Y_i(d) = Y_i(d')$ for all $i \in \mathbb{I}(d) \cap \mathbb{I}(d')$.*

The set $\mathbb{I}(d) \cap \mathbb{I}(d')$ in Definition 7 resembles the set H in the Definition 1 of Zhang and Zhao (2023). Intuitively, it excludes all the units that are not imputable under the partial null hypothesis in the test statistics. At first glance, the *pairwise imputable statistics* seems to restrict the form of the test statistics we can use. However, it turns out to be general enough to include the test statistics we often use. For example, the classic difference-in-mean can be defined as:

$$T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), D) = \|\bar{Y}_{\mathbb{I}(D^{obs})}(D^{obs})_{\{i:D \in \mathcal{D}_i(\epsilon_s)/\mathcal{D}_i(\epsilon_c)\}} - \bar{Y}_{\mathbb{I}(D^{obs})}(D^{obs})_{\{i:D \in \mathcal{D}_i(\epsilon_c)\}}\|$$

where

$$\bar{Y}_{\mathbb{I}(D^{obs})}(D^{obs})_{\{i:D \in \mathcal{D}_i(\epsilon_s)/\mathcal{D}_i(\epsilon_c)\}} = \frac{\sum_{i \in \mathbb{I}(D^{obs})} 1\{D \in \mathcal{D}_i(\epsilon_s)/\mathcal{D}_i(\epsilon_c)\} Y_i(D^{obs})}{\sum_{i \in \mathbb{I}(D^{obs})} 1\{D \in \mathcal{D}_i(\epsilon_s)/\mathcal{D}_i(\epsilon_c)\}},$$

which is the mean value of units in the distance interval $(\epsilon_s, \epsilon_c]$, and

$$\bar{Y}_{\mathbb{I}(D^{obs})}(D^{obs})_{\{i:D \in \mathcal{D}_i(\epsilon_c)\}} = \frac{\sum_{i \in \mathbb{I}(D^{obs})} 1\{D \in \mathcal{D}_i(\epsilon_c)\} Y_i(D^{obs})}{\sum_{i \in \mathbb{I}(D^{obs})} 1\{D \in \mathcal{D}_i(\epsilon_c)\}},$$

which is the mean value of units in the distance interval (ϵ_c, ∞). Difference-in-mean estimator is widely used in the literature, such as Basse et al. (2019) and Puelz et al. (2021). The formula is the same as the classic difference in mean when $\mathbb{I}(D^{obs}) = \{1, \ldots, N\}$, and whether i belongs to distance interval $(\epsilon_s, \epsilon_c]$ or (ϵ_c, ∞) depends on D. In practice, it could be the case that one of the mean values is undefined as no unit i in $\mathbb{I}(D^{obs})$ belongs to one of these two intervals, then we further define $T = \infty$.

In addition, one can try the rank statistics. Formally, following Imbens and Rubin (2015) to define rank as

$$\begin{aligned} R_i &\equiv R_i(Y_{\mathbb{I}(D^{obs})\cap\mathbb{I}(D)}(D^{obs})) \\ &= \sum_{j\in\mathbb{I}(D^{obs})\cap\mathbb{I}(D)} 1\{Y_j(D^{obs}) < Y_i(D^{obs})\} + 0.5 * (1 + \sum_{j\in\mathbb{I}(D^{obs})\cap\mathbb{I}(D)} 1\{Y_j(D^{obs}) = Y_i(D^{obs})\}) \\ &- \frac{1 + \|\mathbb{I}(D^{obs}) \cap \mathbb{I}(D)\|}{2} \end{aligned}$$

Hence,

$$T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), D) = \|\bar{R}_{\{i:D\in\mathcal{D}_i(\epsilon_s)/\mathcal{D}_i(\epsilon_c)\}} - \bar{R}_{\{i:D\in\mathcal{D}_i(\epsilon_c)\}}\|$$

When $Y_i(D^{obs}) = Y_i(D)$ for all $i \in \mathbb{I}(D^{obs})\cap\mathbb{I}(D)$, $R_i(Y_{\mathbb{I}(D^{obs})\cap\mathbb{I}(D)}(D^{obs})) = R_i(Y_{\mathbb{I}(D^{obs})\cap\mathbb{I}(D)}(D))$, so the R_i remains the same. Hence, $T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), D) = T(Y_{\mathbb{I}(D^{obs})}(D), D)$ and satisfies Definition 7.

See section 5 of Imbens and Rubin (2015) for a detailed discussion on the choice of statistics in the randomization testing, and section 5 of Athey et al. (2018) for a detailed discussion on other choices of T in different network settings. One can also use the regression coefficient of interest illustrated in Hoshino and Yanagi (2023). Although the method is valid even without using information on covariates, incorporating covariate adjustments in practice might increase power (Wu and Ding, 2021). See the Appendix D for a detailed discussion. For the sharp null, because all the units are imputable regardless of the treatment assignment d, $\mathbb{I}(d)\cap\mathbb{I}(d') = \{1, \ldots, N\}$ for any d and d', all the formula above would be the same as the classical formula defined in Imbens and Rubin (2015).

A toy example (cont.) Consider the test statistics:

$$T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), D) = \|\bar{Y}_{\mathbb{I}(D^{obs})}(D^{obs})_{\{i:D\in\mathcal{D}_i(0)/\mathcal{D}_i(1)\}} - \bar{Y}_{\mathbb{I}(D^{obs})}(D^{obs})_{\{i:D\in\mathcal{D}_i(1)\}}\|$$

As illustrated in the left-hand side of Figure 3, when $D = D^{obs}$ with unit i_1 being treated, the first term $\bar{Y}_{\mathbb{I}(D^{obs})}(D^{obs})_{\{i:D\in\mathcal{D}_i(0)/\mathcal{D}_i(1)\}}$ would be the mean outcome value of both i_2 and i_6; the second term $\bar{Y}_{\mathbb{I}(D^{obs})}(D^{obs})_{\{i:D\in\mathcal{D}_i(1)\}}$ would be the mean outcome value of i_3 to i_5. On the right-hand side of Figure 3, when the randomized treatment assignment D is unit i_2 being treated, although i_1 and i_3 are both in the distance interval $(0, 1]$, the first term would only use the value of i_3 because i_1 is not in $\mathbb{I}(D^{obs})$; the second term would be the mean

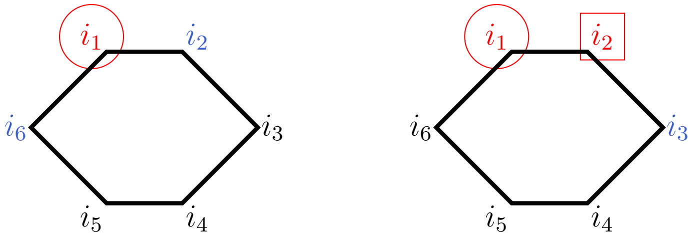


Figure 3: Left: D^{obs}: Unit i_1 being treated (marked as red circle); $D = D^{obs}$, so i_2 and i_6 in the distance interval $(0, 1]$ and both counted in the first term of the difference-in-mean estimator (marked as blue). Right: D^{obs}: Unit i_1 being treated (marked as red circle); D: Unit i_2 being treated (marked as red square), so i_1 and i_3 in the distance interval $(0, 1]$, but only i_3 counted in the first term of the difference-in-mean estimator (marked as blue).

outcome value of i_4 to i_6.

Following the Definition 7 of pairwise imputable statistics, we can have a property to calculate test statistics using only the observed information:

Proposition 1. *Suppose the partial null hypothesis $H_0^{\epsilon_s}$ holds. Suppose $T(Y_{\mathbb{I}(d)}(d), d')$ is a pairwise imputable statistics. Then, $T(Y_{\mathbb{I}(d)}(d), d') = T(Y_{\mathbb{I}(d)}(d'), d')$ for any $d, d' \in \{0, 1\}^N$.*

Proof of Proposition 1. For any $d, d' \in \{0, 1\}^N$. Consider any $i \in \mathbb{I}(d) \cap \mathbb{I}(d')$. By the Definition 5 of *Imputable Units*, under $H_0^{\epsilon_s}$, we have $Y_i(d) = Y_i(d')$. Hence, by the Definition 7 of pairwise imputable statistics, $T(Y_{\mathbb{I}(d)}(d), d') = T(Y_{\mathbb{I}(d)}(d'), d')$. □

By proposition 1, let $d = D^{obs}$ and $d' = D$, we have $T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), D) = T(Y_{\mathbb{I}(D^{obs})}(D), D)$ under the null $H_0^{\epsilon_s}$, which ensures we observe a counterfactual test statistics for comparison. How about we follow the same steps as in FRT to conduct the testing?

A toy example (cont.) If we replace everything is FRT with the new pairwise imputable statistics $T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), D)$, we can obtain Table 2 with all the test statistics. As we can see, following the definition of p-value in the FRT, $pval(D^{obs}) = P(T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), D) \geq T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), D^{obs}))$ respect to $D \sim P(D)$. Hence, the p-value equals $1/6$ as $T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), D^{obs})$ is the largest number in the entire column. However, is it a valid testing procedure? The answer turns out to be NO!

Although we use pairwise imputable statistics, naively constructing the p-value defined in FRT does not guarantee the validity of the test. For validity, similar to FRT, we need the following condition under the partial null hypothesis:

$$T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), D) \sim T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), D^{obs})$$

Table 2: Naive FRT in the toy example

Dist. to the Treated Unit	Potential Outcome Y_i						$T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), D)$
	i_1	i_2	i_3	i_4	i_5	i_6	
$(0, 1, 2, 3, 2, 1)$	2	5	3	1	4	6	17/6
$(1, 0, 1, 2, 3, 2)$	🟥	?	3	1	4	6	2/3
$(2, 1, 0, 1, 2, 3)$	🟥	5	?	1	4	6	2
$(3, 2, 1, 0, 1, 2)$	🟥	5	3	?	4	6	2
$(2, 3, 2, 1, 0, 1)$	🟥	5	3	1	?	6	1/2
$(1, 2, 3, 2, 1, 0)$	🟥	5	3	1	4	?	1

Legend: Dist. to the Treated Unit: the minimum distance of each unit to the treated units. j means unit is distance j away from the treated units and belongs to the distance interval $(j-1, j]$ for $j = 1, 2, 3$. 0 means the unit itself is being treated in the randomized D. Potential Outcome Y_i: Potential outcome of each unit under the null H_0^0 with red ? as missing values. Unit i_1 doesn't belong to set $\mathbb{I}(D^{obs})$, so the whole column is marked as red. Blue cells are the units used to calculate the mean value in the first term of the test statistics. $T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), D)$: test statistics under different D and fixing D^{obs} that unit i_1 is treated.

where the left-hand side is induced by the randomness of D with D^{obs} fixed, and the right-hand side is induced by the randomness of D^{obs}.

By Proposition 1, under the null, we have the left-hand side:

$$T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), D) = T(Y_{\mathbb{I}(D^{obs})}(D), D)$$

Due to the randomness of the experimental design, we also have the right-hand side:

$$T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), D^{obs}) \sim T(Y_{\mathbb{I}(D)}(D), D)$$

Therefore, to ensure the validity of the test, we need:

$$T(Y_{\mathbb{I}(D^{obs})}(D), D) \sim T(Y_{\mathbb{I}(D)}(D), D)$$

However, this is not guaranteed under the partial null hypothesis because $\mathbb{I}(D^{obs}) \neq \mathbb{I}(D)$ in general. Different units have different neighbors in practice, leading to different sets of imputable units for different treatment assignments, as discussed in the toy example. In contrast, when testing the sharp null hypothesis, $\mathbb{I}(D^{obs}) = \{1, \ldots, N\} = \mathbb{I}(D)$ and the validity trivially holds.

To address the challenges arising from the variation in imputable unit sets, previous literature suggests a remedy by designing a conditioning event formed by a fixed subset of imputable units, called *focal units*, and a fixed subset of assignments, called *focal assignments*. Then using Conditional Randomization Tests (CRT) by conducting FRT within the conditioning event. See a detailed discussion in Appendix A.2. However, in practice, using conditioning events naturally introduces two drawbacks.

First, as Zhang and Zhao (2023) pointed out, there is a trade-off between the sizes of focal units and focal assignments: a larger subset of treatment assignments often comes with a smaller subset of experimental units. This inevitably leads to a loss of information, with fewer units and assignments within conditioning events, potentially affecting the power of the test. Second, constructing the conditioning event adds a layer of computational burden. This raises the question: can unconditional randomization testing be valid in finite samples?

Fortunately, this paper demonstrates that the answer is YES! While previous literature embeds the idea of carefully designing a fixed subset of units to maintain the validity of randomization testing, my method avoids fixing the subset of units during implementation. Instead, it maintains valid testing through a carefully designed p-value calculation.

3.2 P-value with pairwise comparison

Inspired by the recent works of Wen et al. (2023) and Guan (2023) from the selective inference literature, the key idea is to compute p-values directly by summing pairwise inequality comparisons between $T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), d^r)$ and $T(Y_{\mathbb{I}(d^r)}(D^{obs}), D^{obs})$. When the null hypothesis is false, $T(Y_{\mathbb{I}(d^r)}(D^{obs}), D^{obs})$ would maintain a relatively large value across different d^r since the distance interval for each unit is fixed by D^{obs}. The change in d^r only alters the set of units used in the test statistics. Therefore, we would expect a small p-value, as the probability that $T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), d^r)$ is larger than $T(Y_{\mathbb{I}(d^r)}(D^{obs}), D^{obs})$ is low. I refer to any randomization testing with p-values constructed through this pairwise comparison idea as *Partial Null Randomization Testing (PNRT)*. Formally, I call the procedure *pairwise comparison-based PNRT*, with the p-value defined below:

Definition 8 (P-Value for pairwise comparison-based PNRT). *Define $pval^{pair}(D^{obs}) : \mathbb{D} \rightarrow [0, 1]$ as $pval^{pair}(D^{obs}) = P(T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), D) \geq T(Y_{\mathbb{I}(D)}(D^{obs}), D^{obs}))$ respect to $D \sim P(D)$*

In practice, we can calculate this p-value with the following algorithm, where the p-value is calculated as the mean value of $1 + R$ draws due to using $d = D^{obs}$ for $r = 0$, so there are $R + 1$ draws:

Comparison to the inner vs. outer ring strategy One popular strategy for testing interference at some distance ϵ_s is the inner vs. outer ring strategy. The idea is that outer ring units' outcome values are not affected by the treatment's interference and could approximate the control level potential outcome in the treated group. For example, Blattman et al. (2021)

Algorithm 1 Pairwise comparison-based PNRT Procedure

```
Inputs      : Test statistic $T = T(Y(d), d)$, observed assignment $D^{obs}$, observed
              outcome $Y^{obs}$, treatment assignment mechanism $P(D)$, and size $\alpha$.
for $r = 1$ to $R$ do
    Randomly sample: $d^r \sim P(D)$, Store $T_r \equiv T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), d^r)$.
    Store $T_r^{obs} \equiv T(Y_{\mathbb{I}(d^r)}(D^{obs}), D^{obs})$.
end
Output      : P-value: $\hat{pval}^{pair} = (1 + \sum_{r=1}^{R} 1\{T_r \geq T_r^{obs}\})/(1 + R)$.
              Reject if $\hat{pval}^{pair} \leq \alpha/2$.
```

incorporate a similar idea when attempting to pinpoint the distance of the spillover effect. They first calculate the average mean value across *different units* in the inner ring and the average mean value across *different units* in the outer ring, then test whether there is a systematic difference between the two groups.[11]

However, as noted by Pollmann (2023), this strategy requires assumptions beyond the random experiment: First, as discussed in Aronow (2012), even in a random experiment, the distance of each unit to the treated units is not random. Thus, the outer ring units might systematically differ from the inner ring units across different treatment assignments. Second, as highlighted by Pollmann (2023), even if each unit is equally likely to be in the inner or outer rings, we need to make functional form assumptions on the potential outcome to eliminate the bias from such a comparison. Overall, the outer ring units may not possess potential control outcomes comparable to those of the inner ring units without further assumptions, potentially leading to biased results.

The idea behind the partial null hypothesis in Definition 4 is to assess interference by directly testing the value of *the same unit's* potential outcome whenever it is at least distance ϵ_s away from the treated units. A key advantage of the partial null hypothesis is that it directly addresses the unit-level potential outcome rather than the average outcome across different units that might not be compatible even with the random experiment. The critical contribution of this paper is to show that we can test interference with only the assumption of random treatment assignment.

[11] Blattman et al. (2021) use an F-test for the proposed mean difference of outcomes variables "Perceived risk" and "Crime incidence". Results can be found in Blattman et al. (2021)'s online appendix subsection A.2.

A toy example (cont.) Using the same difference-in-mean estimator as before,

$$T(Y_{\mathbb{I}(D)}(D^{obs}), D^{obs}) = \|\bar{Y}_{\mathbb{I}(D)}(D^{obs})_{\{i:D^{obs}\in\mathcal{D}_i(0)/\mathcal{D}_i(1)\}} - \bar{Y}_{\mathbb{I}(D)}(D^{obs})_{\{i:D^{obs}\in\mathcal{D}_i(1)\}}\|$$

Table 3: PNRT in the toy example

Dist. to the Treated Unit	Potential Outcome Y_i						$T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), D)$	$T(Y_{\mathbb{I}(D)}(D^{obs}), D^{obs})$
	i_1	i_2	i_3	i_4	i_5	i_6		
$(0, 1, 2, 3, 2, 1)$	2	5	3	1	4	6	17/6	17/6
$(1, 0, 1, 2, 3, 2)$	?	?	3	1	4	6	2/3	10/3
$(2, 1, 0, 1, 2, 3)$	?	5	?	1	4	6	2	3
$(3, 2, 1, 0, 1, 2)$	?	5	3	?	4	6	2	2
$(2, 3, 2, 1, 0, 1)$	?	5	3	1	?	6	1/2	7/2
$(1, 2, 3, 2, 1, 0)$	?	5	3	1	4	?	1	7/3

Legend: Dist. to the Treated Unit: the minimum distance of each unit to the treated units. j means unit is distance j away from the treated units and belongs to the distance interval $(j-1, j]$ for $j = 1, 2, 3$. 0 means the unit itself is being treated in the randomized D. Potential Outcome Y_i: Potential outcome of each unit under the null H_0^0 with red ? as missing values. Units i_2 and i_6 are in the distance interval $(0, 1]$ under D^{obs}, so the two columns are marked as deep blue. Units i_3 to i_5 are in the distance interval $(1, \infty)$ under D^{obs}, so the three columns are marked as light blue.

According to Table 3, only when D has unit i_1 and unit i_4 being treated, $T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), D) \geq T(Y_{\mathbb{I}(D)}(D^{obs}), D^{obs})$. So, $pval^{pair} = 2/6$. In practice, similar to Guan (2023), we can use 1/2 to discount the number of equalities and decrease the p-value without compromising the validity of the test. Additionally, in the simulation, I tried using a uniform random number multiplied by the number of equalities, and the test remained valid.

The validity of Algorithm 1 is implied by the symmetric between $T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), d^r)$ and $T(Y_{\mathbb{I}(d^r)}(D^{obs}), D^{obs})$ under $H_0^{\epsilon_s}$. Intuitively, given the D^{obs} and d^r, both terms are restricted to units $i \in \mathbb{I}(D^{obs}) \cap \mathbb{I}(d^r)$ by Definition 7. Additionally, by Proposition 1, under the null, consider $d = D$ and $d' = D^{obs}$, $T(Y_{\mathbb{I}(D)}(D^{obs}), D^{obs}) = T(Y_{\mathbb{I}(D)}(D), D^{obs})$ which is the counterfactual value of $T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), D)$ by flipping the observed assignment and randomized assignment between D and D^{obs}. Hence, the pairwise comparison is symmetric and implies the following theorem:

Theorem 1. *Suppose the partial null hypothesis $H_0^{\epsilon_s}$ is true. Then the p-value $pval^{pair}$ in Definition 8 constructed by the pairwise comparison-based PNRT Algorithm 1 satisfied $P(pval^{pair}(D^{obs}) \leq \alpha/2) \leq \alpha$, for any $\alpha > 0$, where the probability is with respect to $D^{obs} \sim P(D)$.*

Proof and a discussion in the case with too many potential treatment assignments can be found in Appendix B.

The primary limitation of Theorem 1 is that, when rejecting the null hypothesis at significance level α, the probability of a false rejection is bounded by 2α instead of α. A straightforward approach to address this is to reject the null when the p-value is below $\alpha/2$ rather than α. Another possible method, inspired by Wen et al. (2023), involves adopting a more conservative testing procedure.

3.3 Minimization-based PNRT

The key idea is to construct $\tilde{T}(D^{obs}) = min_{d\in\{0,1\}^N}(T(Y_{\mathbb{I}(d)}(D^{obs}), D^{obs}))$, and the define the following p-value:

Definition 9 (P-Value for minimization-based PNRT)**.** *Define $pval^{min}(D^{obs}) : \mathbb{D} \to [0,1]$ as $pval^{min}(D^{obs}) = P(T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), D) \geq \tilde{T}(D^{obs}))$ respect to $D \sim P(D)$*

In practice, we can calculate this value with the following algorithm, where the p-value is calculated as the mean value of $1 + R$ draws due to using $d = D^{obs}$ for $r = 0$, so there are $R + 1$ draws:

```
Algorithm 2 Minimization-based PNRT Procedure
Inputs      : Test statistic $T = T(Y(d), d)$, observed assignment $D^{obs}$, observed
              outcome $Y^{obs}$, treatment assignment mechanism $P(D)$, and size $\alpha$.
for $r = 1$ to $R$ do
    Randomly sample: $d^r \sim P(D)$, Store $T_r \equiv T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), d^r)$.
     Store $T_r^{obs} \equiv T(Y_{\mathbb{I}(d^r)}(D^{obs}), D^{obs})$.
end
Compute     : $\tilde{T}^\star(D^{obs}) = min_{r=1,...,R}(T_r^{obs}))$
Output      : P-value: $\hat{pval}^{min} = (1 + \sum_{r=1}^{R} 1\{T_r \geq \tilde{T}^\star(D^{obs})\})/(1 + R)$.
              Reject if $\hat{pval}^{min} \leq \alpha$.
```

As shown in table 3, in the toy example, $\tilde{T}(D^{obs}) = 2$, so $pval^{min} = 1/2$ with two other equal numbers. The key difference between minimization-based PNRT and pairwise comparison-based PNRT is that by taking the minimization, we ensure the size control as shown in theorem 2.

Theorem 2. *Suppose the partial null hypothesis $H_0^{\epsilon_s}$ is true. Then the p-value $pval^{min}$ defined in the minimization-based PNRT algorithm 2 is a valid p-value, i.e. $P(pval^{min}(D^{obs}) \leq \alpha) \leq \alpha$, for any $\alpha \in [0,1]$, where the probability is with respect to $D^{obs} \sim P(D)$.*

Proof can be found in Appendix B.

Too many potential treatment assignments. When N is large, finding the minimum and constructing $\tilde{T}(D^{obs})$ can be challenging. However, simulations in section 5.1 show that Algorithm 2 with $R = 1000$, while not computing the true $\tilde{T}(D^{obs})$, remains conservative and ensures validity.

To guarantee the validity of Algorithm 2 when the number of units is large, one approach is to use optimization methods to find an approximation $\tilde{T}^R(D^{obs})$ of the minimization $\tilde{T}(D^{obs})$ so that $\tilde{T}(D^{obs}) \geq \tilde{T}^R(D^{obs}) - \eta_R$ with probability $1 - \eta$. Hence, we only need to adjust the rejection level to $\tilde{\alpha}$ to satisfy $\alpha = \tilde{\alpha}(1 - \eta) + \eta$, ensuring the test remains valid. However, this method adds a computational burden.

Alternatively, one can combine the Conditional Randomization Tests (CRT) with PNRT to reduce the number of potential treatment assignments, making it easier to find $\tilde{T}(D^{obs})$ within the conditioning event. For example, as pointed out by Athey et al. (2018) and Zhang and Zhao (2023), researchers often trim the potential assignment space to all treatment assignments with the same number of treated units as the observed assignment. The testing procedure remains valid as a two-stage process: first, formulate the number of treated units by the observed assignment, and second, conduct testing within the trimmed assignment space. Using PNRT in this case allows for a much larger set of focal units, potentially increasing power.

Trade-off in the conservatism. To avoid computational difficulties while maintaining the validity of the test, researchers can simply use pairwise comparison-based PNRT as outlined in Algorithm 1 with a rejection level of $\alpha/2$. The later simulation shows that this straightforward adjustment has higher power than minimization-based PNRT, and it is actually a conservative way to ensure validity in the worst-case scenario, as there are cases where the rejection level α is valid. I will leave the detailed discussion to Section 5.1.

3.4 Comparison to previous literature

As illustrated earlier, the key difference is that the units included in $\mathbb{I}(d)$ vary across different assignments d, utilizing all imputable units for testing. The procedure in Owusu (2023) shares a similar property but is more complicated to implement, involving tuning parameters, and is only valid asymptotically. PNRT is easy to implement without any tuning parameters and is valid in finite samples.

In the case of the sharp null, we know $\mathbb{I}(d) = \{1, \ldots, N\}, \quad \forall d \in \{0,1\}^N$, hence $\tilde{T}(D^{obs}) = min_{d \in \{0,1\}^N}(T(Y_{\mathbb{I}(d)}(D^{obs}), D^{obs})) = T(Y(D^{obs}), D^{obs}), \quad \forall D^{obs} \in \{0,1\}^N$. Therefore, both pairwise comparison-based PNRT and minimization-based PNRT coincide with the classical Fisher Randomization Tests (FRT). The proposed method nests FRT but ensures validity under partial null by allowing the set of units included in test statistics to vary across different assignments.

In the case of partial null, compared to conditional randomization tests, $\mathbb{I}(D^{obs})$ is the counterpart of $\mathbb{N}_{\mathbb{U}}$, and $\{0,1\}^N$ is the counterpart of $\mathbb{D}_{\mathbb{U}}$. Recognizing PNRT as conditional randomization tests with the "conditioning event" $(\mathbb{I}(D^{obs}), \{0,1\}^N)$, PNRT has a larger size of the event but construct p-value differently. However, restricting the focal assignment set to $\mathbb{D}_{\mathbb{U}}$, and fixing the focal units set $\mathbb{N}_{\mathbb{U}}$ in place of $\mathbb{I}(D^{obs})$, the p-value constructed follow Definition 8 and Definition 9 would coincide with the original CRT. Thus, the larger "conditioning event" might lead to a higher power if the extra potential assignments and units are useful. If including all the assignments from $\{0,1\}^N$ is not optimal, combining PNRT with CRT could avoid missing test statistics and selecting more relevant assignments to increase power (Lehmann and Romano, 2005; Hennessy et al., 2015). How to leverage the flexibility introduced by PNRT to optimize power performance is left for future research.

Additionally, since the value of $T(Y_{\mathbb{I}(d)}(D^{obs}), D^{obs})$ only depends on $\mathbb{I}(d)$ for a fixing D^{obs}, it might be the case that when we have large N, the variation of $\mathbb{I}(d)$ is very small, making pairwise comparison-based PNRT similar to minimization-based PNRT. In that case, pairwise comparison-based PNRT would achieve the asymptotic validity control of α rather than 2α in the finite sample case. A formal proof might be worth exploring in the future.

4 Framework to determine the boundary of interference

In practice, researchers may seek to estimate a sequence of partial null hypotheses at varying distances, ϵ_s, to identify the neighborhood of interference. This approach can be instrumental in selecting a pure control distance or evaluating how interference evolves with distance. To this end, we consider a sequence of distance thresholds:

$$\epsilon_0 < \epsilon_1 < \epsilon_2 < \cdots < \epsilon_K < \infty$$

where $K \geq 1$ is chosen to include the setting introduced in earlier sections. For instance, if the objective is to test for the presence of interference, one could set $K = 1$ with $\epsilon_0 = \epsilon_s = 0$ and $\epsilon_1 = \epsilon_c$.

Utilizing this sequence of distances, we can test a sequence of null hypotheses as defined in Definition 4, where $\epsilon_s \in \{\epsilon_0, \ldots, \epsilon_K\}$. However, it is important to note that not all distance levels will yield nontrivial power.

Firstly, there is a trade-off between the number of thresholds tested and the power of each test. While testing more thresholds can provide a deeper understanding of how interference varies with distance, it may also diminish the power to detect interference, particularly if certain threshold groups lack sufficient units. Based on simulation exercises, I recommend ensuring that each observed exposure level includes at least 20 units to maintain nontrivial power at a significance level of $\alpha = 0.05$.

Secondly, in some instances, ϵ_K may represent the maximum distance in the network, leaving no further options for ϵ_c. While it is still possible to test $H_0^{\epsilon_K}$, it may be necessary to explore alternative variations, such as the number of nearby treated units, as suggested by Hoshino and Yanagi (2023), to construct a test statistic with nontrivial power. For simplicity, this section will focus on testing $H_0^{\epsilon_k}$ for $k \leq K - 1$.

A toy example (cont.) Based on the above setup, one might consider setting $K = 3$ with $(\epsilon_0, \epsilon_1, \epsilon_2, \epsilon_3) = (0, 1, 2, 3)$. However, it may only be feasible to test H_0^0, H_0^1, and H_0^2, as testing the partial null hypothesis H_0^3 requires at least one unit to be at a distance greater than 3 from any treated unit, which is not the case in this example. Therefore, for this toy example, I would set $K = 2$ with $(\epsilon_1, \epsilon_2) = (1, 2)$.

Following Definition 4 of $H_0^{\epsilon_s}$, the multiple hypotheses we consider have a nested structure:

Proposition 2. *Suppose there exists $\bar{K} \geq 0$ such that, for any $k \leq \bar{K} - 1$, $H_0^{\epsilon_k}$ is false, and $H_0^{\epsilon_{\bar{K}}}$ is true. Then, $H_0^{\epsilon_k}$ is true for any $k \geq \bar{K}$.*

Proof of Proposition 2. By Definition 4, if $H_0^{\epsilon_{\bar{K}}}$ is true, then $Y_i(d) = Y_i(d')$ for all $i \in \{1, \ldots, N\}$ and any $d, d' \in \mathcal{D}_i(\epsilon^{\bar{K}})$.

Observe that, for any $i \in 1, \ldots, N$, by Definition 3:

$$\mathcal{D}_i(\epsilon_0) \supset \mathcal{D}_i(\epsilon_1) \supset \cdots \supset \mathcal{D}_i(\epsilon_K)$$

Hence, for any $k \geq \bar{K}$, and any $d, d' \in \mathcal{D}_i(\epsilon_k) \subseteq \mathcal{D}_i(\epsilon_{\bar{K}})$, $Y_i(d) = Y_i(d')$ *for all* $i \in \{1, \ldots, N\}$. By Definition 4, $H_0^{\epsilon_k}$ is true for any $k \geq \bar{K}$. □

By Proposition 2, interference would exist only up to a certain boundary. Given this nested structure, we would prefer an inference method that helps determine such boundaries by rejecting the null hypothesis up to some distance while not rejecting it beyond that point. However, in practice, we might encounter a situation where $H_0^{\epsilon_k}$ cannot be rejected, but $H_0^{\epsilon_{k+1}}$

can be. This could occur either because the test lacks the power to reject the false null $H_0^{\epsilon_k}$, or because the test erroneously rejects the true null $H_0^{\epsilon_{k+1}}$ due to multiple hypothesis testing. To address the issue of multiple hypothesis testing, I propose controlling for the *family-wise error rate (FWER)* to mitigate the risk of over-rejecting true null hypotheses:

Definition 10 (Family-wise error rate over all $H_0^{\epsilon_k}$ for $k = 0, \ldots, K-1$.)**.** *Suppose there exist $\bar{K} \geq 0$, such that for any $k \leq \bar{K} - 1$, $H_0^{\epsilon_k}$ is false, and $H_0^{\epsilon_{\bar{K}}}$ is true. Define $FWER = P(\exists k \geq \bar{K}, H_0^{\epsilon_k}$ is rejected).*

The definition of FWER in Definition 10 is motivated by the nested structure of $H_0^{\epsilon_k}$, wherein the null hypothesis is true for any $k \geq \bar{K}$. The critical question is how we should reject all the $H_0^{\epsilon_k}$ when determining the boundary while still controlling for FWER.

4.1 A valid procedure to determine the neighborhood of interference

A major challenge in testing how interference evolves with distance lies in addressing the issue of multiple hypothesis testing when conducting a series of tests to identify the neighborhood of interference. To manage the increased error rate that arises from multiple tests, and drawing inspiration from Meinshausen (2008) and subsection 15.4.4 of Lehmann and Romano (2005), I propose Algorithm 3.

Algorithm 3 Sequential Testing Procedure

```
Inputs      : Test statistic $T = T(Y(d), d)$, observed assignment $D^{obs}$, observed
              outcome $Y^{obs}$, treatment assignment mechanism $P(D)$.
Set         : $\hat{K} = 0$.
for $k = 0$ to $K - 1$ do
    Testing $H_0^{\epsilon_k}$ using PNRT procedure, collect $pval^k$.
    If $pval^k \leq \alpha$, set $\hat{K} = k + 1$ and reject $H_0^{\epsilon_k}$.
    If $pval^k > \alpha$, Break
end
Output      : Significant spillover within distance $\epsilon_{\hat{K}}$.
```

Algorithm 3 is designed to control the FWER while taking advantage of the nested structure of sequential hypothesis testing. Unlike typical multiple-hypothesis testing procedures, such as the Bonferroni-Holm procedure, which would reject the null at a level smaller than α, this algorithm does not require adjusting the significance level and potentially increases power compared to traditional multiple-hypothesis testing adjustments (Meinshausen, 2008).

Moreover, if the unadjusted p-values increase as k increases, indicating that interference diminishes with distance, there is no loss of power compared to not making multiple hypothesis testing adjustments since we would naturally not reject any null beyond a certain distance. When using the pairwise comparison-based PNRT for each k, rejecting at the $\alpha/2$ level ensures size control. For the partial null hypothesis $H_0^{\epsilon_k}$, a natural choice for ϵ_c is ϵ_{k+1}. Theorem 3 guarantees the FWER control of Algorithm 3.

Theorem 3. *The Sequential Testing Procedure constructed by Algorithm 3 controls the familywise error rate at α.*

Proof of Theorem 3. Without loss of generality, consider minimization-based PNRT below. The same proof holds when using the pairwise comparison-based PNRT with a rejection level of $\alpha/2$.

Suppose for any $k < \bar{K}$, $H_0^{\epsilon_k}$s are false, and $H_0^{\epsilon_{\bar{K}}}$ is true. Then, by Algorithm 3, if there exist $k \geq \bar{K}$ that $H_0^{\epsilon_k}$ is rejected, it must be the case that $H_0^{\epsilon_{\bar{K}}}$ is rejected. Thus, by Definition 10:

$$FWER = P(pval^1 \leq \alpha, pval^2 \leq \alpha, \ldots, pval^{\bar{K}} \leq \alpha) \leq P(pval^{\bar{K}} \leq \alpha) \leq \alpha.$$

because $H_0^{\epsilon_{\bar{K}}}$ is true. □

A toy example (cont.) Algorithm 3 can be implemented in two steps: First, collect $pval^0$ for H_0^0 and reject H_0^0 if $pval^0 \leq \alpha$. If H_0^0 is not rejected, report that no significant interference was found. If H_0^0 is rejected, proceed to the second step, collect $pval^1$ for H_0^1, and reject H_0^1 if $pval^1 \leq \alpha$. If H_0^1 is rejected, report significant interference within distance 2; if H_0^2 is not rejected, report significant interference within distance 1.

4.2 Rational of using FWER

In practice, FWER is not the only criterion for controlling error rates in multiple-hypothesis testing. As Anderson (2008) points out, it may also be worthwhile to consider false discovery rate (FDR) control in exploratory analyses, as it allows for a small number of Type I errors in exchange for greater power than FWER control. A looser adjustment algorithm might be of interest in future work. However, if a policymaker aims to implement a policy in a distant area, expecting a positive far-distant interference effect, a high FWER might lead to overly optimistic assumptions about the interference boundary. Therefore, FWER can still be helpful by providing a conservative distance threshold, which better accounts for interference when calculating the expected welfare change.

Additionally, this procedure has the advantage of helping pre-test the pure control group and ensuring post-inference validity even after this first-step pre-testing.

A procedure to help select pure control group As discussed in Section 3.1, we typically need a "safe distance" ϵ_c to construct a pure control group. But how should we choose this distance? A natural candidate could be ϵ_K, representing the furthest distance that maintains nontrivial power for testing. However, it may be tempting to decrease this distance to include more units in the pure control group, thereby increasing the power of the test. The key challenges are: (1) determining which ϵ_c to choose, and (2) addressing the post-model selection inference issue highlighted by Leeb and Pötscher (2005).

Theorem 4 guarantees the validity of $H_0^{\epsilon_k}$ for any k, even after choosing ϵ_c using Algorithm 3 as a first step. It ensures that subsequent inference using any method, including PNRT or other asymptotic-based approaches, remains valid when applying the pre-testing rule to obtain $pval(D^{obs})$. Therefore, in practice, researchers might consider using the distance $\epsilon_{\hat{K}}$ obtained from Algorithm 3 as the threshold ϵ_c for the subsequent analysis:

Theorem 4. *For any $H_0^{\epsilon_k}$ with $k = 0, \ldots, K-1$, suppose the partial null hypothesis $H_0^{\epsilon_k}$ is true. Then, given the observed assignment D^{obs}, consider a two-step pre-testing procedure:*

Step 1: Obtain $\epsilon_{\hat{K}}$ from Algorithm 3.

Step 2: Use $\epsilon_c = \epsilon_{\hat{K}}$ to test $H_0^{\epsilon_k}$ and obtain $pval(D^{obs})$ using any inference method.

The two-step procedure constructed above satisfied $P(pval(D^{obs}) \leq \alpha) \leq \alpha$.

Proof of Theorem 4 Suppose that for any $k \leq \bar{K}$, $H_0^{\epsilon_k}$s are false, and $H_0^{\epsilon_{\bar{K}+1}}$ is true. Due to the nested structure of $H_0^{\epsilon_k}$, it is true for any $k > \bar{K}$. To validate the testing procedure with the added pre-testing step, we only need to ensure that the p-value $P(pval(D^{obs}) \leq \alpha) \leq \alpha$ for any $H_0^{\epsilon_{\tilde{k}}}$ that $\tilde{k} > \bar{K}$, which can be split into two terms:

$$\begin{aligned} P(pval(D^{obs}) \leq \alpha) = &\, p(\hat{K} \geq \tilde{k}+1) P(pval(D^{obs}) \leq \alpha | \hat{K} \geq \tilde{k}+1) \\ &+ p(\hat{K} < \tilde{k}+1) P(pval(D^{obs}) \leq \alpha | \hat{K} < \tilde{k}+1) \end{aligned}$$

Following Algorithm 3, we would reject any $H_0^{\epsilon_k}$ with $k < \hat{K}$, and failed to reject any $H_0^{\epsilon_k}$ with $k \geq \hat{K}$. So, when $\hat{K} \geq \tilde{k}+1$, it must be the case that $H_0^{\epsilon_{\bar{K}+1}}$ is rejected as $\tilde{k} > \bar{K}$.

Hence,

$$p(\hat{K} \geq \tilde{k}+1) \leq P(pval^{\bar{K}+1} \leq \alpha) \leq \alpha;$$

When $\hat{K} < \tilde{k}+1$, we would not reject $H_0^{\epsilon_{\tilde{k}}}$ with or without the pre-testing step. Hence, $P(pval(D^{obs}) \leq \alpha | \hat{K} < \tilde{k}+1) = 0$. So,

$$P(pval(D^{obs}) \leq \alpha) \leq \alpha P(pval(D^{obs}) \leq \alpha | \hat{K} \geq \tilde{k}+1) \leq \alpha$$

□

The rationale is as follows: If the pre-testing procedure does not reject any true null hypothesis, the second-step inference will avoid re-testing these true nulls, thus preventing any false rejections. On the other hand, if the pre-testing does reject some true nulls, there is a minimal chance – less than α, as ensured by the design of Algorithm 3 – that the second-step inference might over-reject due to testing different hypotheses on the same dataset. Consequently, the probability of a false rejection remains below the significance level α. Without this adjustment, we could encounter issues with post-model selection inference. Refer to Appendix C for a more detailed discussion. Still, the chosen ϵ_c could be smaller than the actual boundary due to the conservative nature of Algorithm 3. Therefore, researchers should carefully weigh the benefits of implementing the pre-testing step in their analysis.

5 Application: Reanalysis of crime in Bogotá

In 2016, Bogotá, Colombia, conducted a large-scale experiment described by Blattman et al. (2021). The study involved $136,984$ street segments, with $1,919$ identified as crime "hotspots." Among these hotspots, 756 were randomly assigned to a treatment involving increased daily police patrolling duties from 92 to 169 minutes over eight months. The original study also included an independent intervention to enhance municipal services, which is peripheral to the main focus. The primary outcome of interest was the number of crimes on each street segment, encompassing both property crimes and violent crimes such as assault, rape, and murder.

Figure 4(left) illustrates the distribution of hotspots, showing many are clustered closely together. While only $1,919$ street segments received active treatment, every segment potentially experienced spillover effects, creating a "dense" network that complicates the application of cluster-robust standard errors to address unit correlation. The original paper estimated a negative treatment effect and used Fisher randomization tests (FRT) with a sharp null hypothesis of no effect for inference.

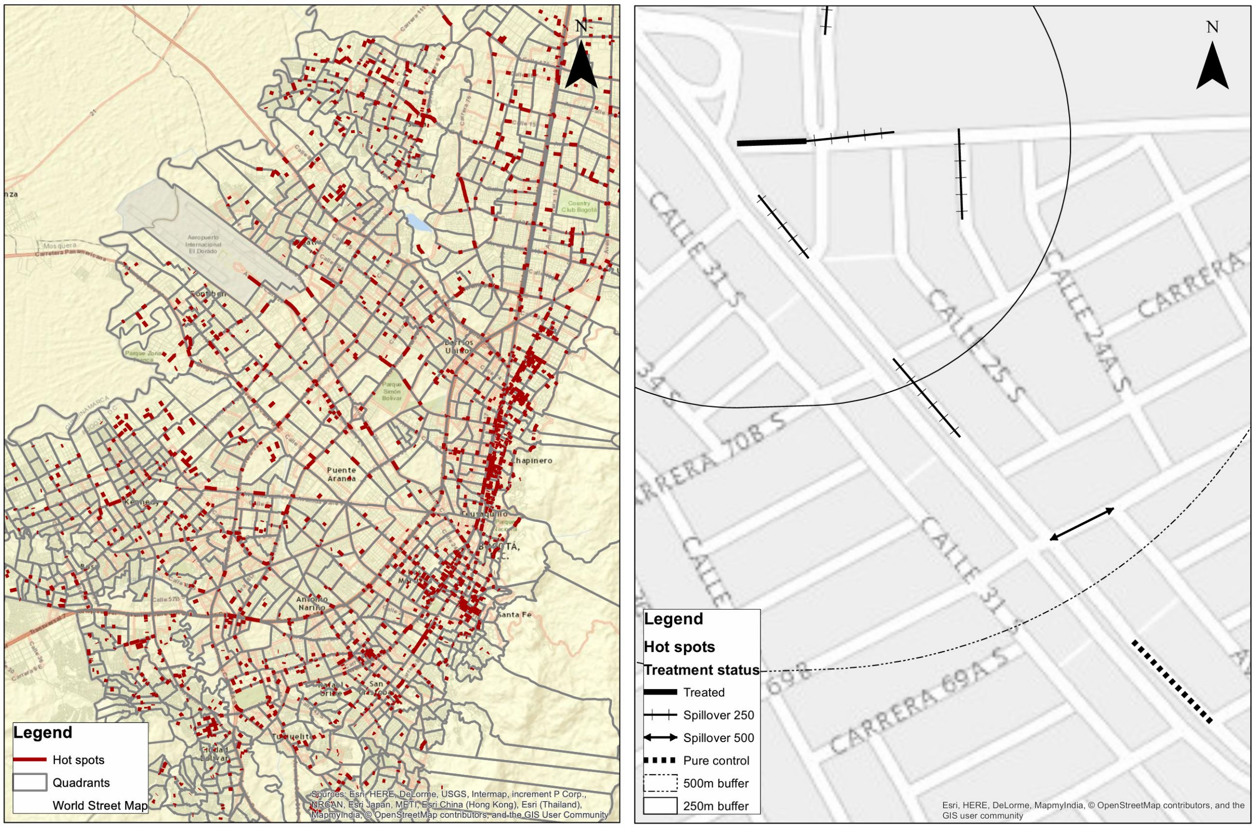


Figure 4: Left: Map of the experimental sample with hotspots street segments in red. Right: An example of assignment to the four experimental conditions. Source: Blattman et al. (2021)

Additionally, to assess the total welfare of the policy, it's crucial to evaluate whether interference occurred following treatment assignment, such as crime displacement or deterrence in nearby neighborhoods. Therefore, the authors aimed to answer the following questions: 1) Does interference exist? 2) If so, what is its direction (displacement or deterrence)? 3) What distance is effective for this interference? Given the challenges in modeling correlation across units within such a dense network, testing a partial null hypothesis, as proposed by Blattman et al. (2021) and Puelz et al. (2021), becomes relevant. I specify the distance threshold sequence $(\epsilon_0, \epsilon_1, \epsilon_2, \epsilon_3) = (0, 125, 250, 500)$ for $K = 3$, where the distance interval $(500, \infty)$ represents a pure control group with no treated units within 500 meters. Figure 4(right) provides an example of different distance intervals identified in Blattman et al. (2021).

5.1 Power comparison of spatial interference: A simulation study

For a comprehensive approach to testing in such a large-scale experiment, it is prudent to preselect the preferred method through a simulation study. Specifically, I generate $N = 1000$ points from a bivariate Gaussian with non-diagonal covariance to simulate the network on a $[0,1] \times [0,1]$ space, including 20 hotspots and 7 randomly treated units, mirroring proportions similar to the original Bogotá study.

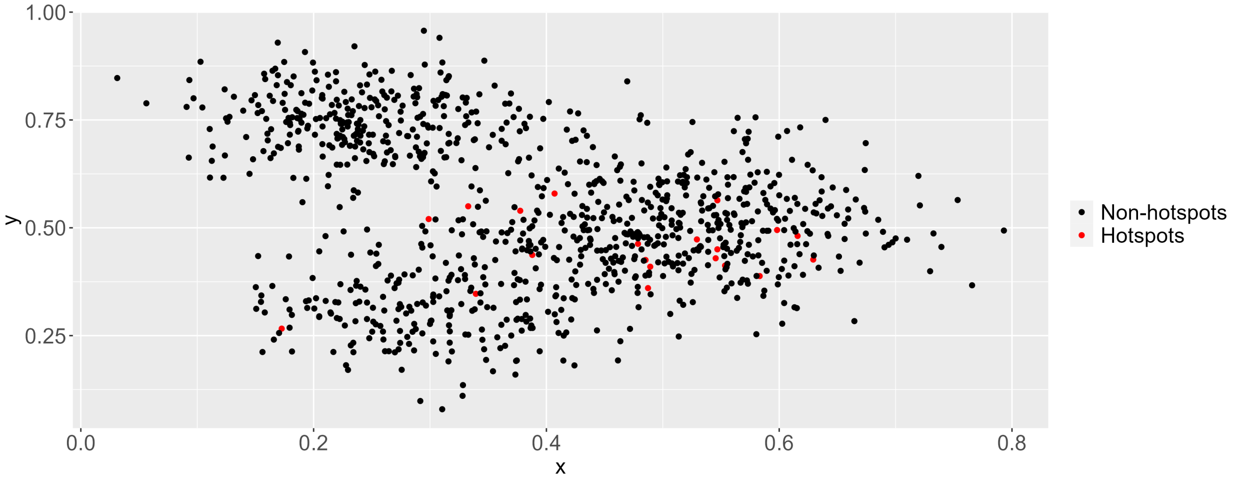


Figure 5: Distribution of the Units

Figure 5 illustrates the distribution of units in this space. To simplify, I focused on two distance thresholds, with $(\epsilon_0, \epsilon_1, \epsilon_2) = (0, 0.1, 0.2)$. Across different treatment assignments, the distance interval $(0, 0.1]$ comprises approximately 420 units, $(0.1, 0.2]$ around 250 units, and the pure control group $(0.2, \infty)$ around 320 units.

Recall that the partial null hypothesis of interest for $k = 0$ and 1:

$$H_0^{\epsilon_k} : Y_i(d) = Y_i(d') \textit{ for all } i \in \{1, \ldots, N\}, \textit{ and any } d, d' \in \mathcal{D}_i(\epsilon_k)$$

The potential outcome schedule was calibrated to match the Bogotá street network using gamma distributions, ensuring they align with the mean and variance of the observed total crimes, as detailed in Table 4. Additionally, I set a negative treatment effect of 1 while ensuring all treated units maintained a nonnegative number of crimes. I also incorporated a decreasing displacement effect with respect to distance levels with a positive τ. Our primary focus is on the spillover effect, τ.

Throughout the analysis, I compared five methods: 1) The classic FRT using the sharp null hypothesis of no effect rather than a partial null, which is also used in Blattman et al.

Table 4: Potential Outcome Schedule in the Simulation

Pure control for "non-hotspots":	$Y_i^C \sim Gamma(0.086, 3.081)$
Pure control for "hotspots":	$Y_i^C \sim Gamma(0.737, 1.778)$
Treated unit:	$Y_i^T = max(Y_i^C - 1, 0)$
Short-range spillover:	$Y_i(d) = Y_i^C + \tau \quad \forall d \in \mathcal{D}_i(0)/\mathcal{D}_i(0.1)$
Long-range spillover:	$Y_i(d) = Y_i^C + 0.5\tau \quad \forall d \in \mathcal{D}_i(0.1)/\mathcal{D}_i(0.2)$

Legend: $Gamma(k, \theta)$: The first element k is the shape parameter; The second element θ is the scale parameter. Y_i^C represents the pure control potential outcome for unit i; Y_i^T represents the potential outcome for unit i when being treated.

(2021) when conducting inference for spillover effect; 2) The Biclique CRT proposed by Puelz et al. (2021), which is considered the benchmark for CRT due to its demonstrated power in simulations involving general interference; 3) The minimization-based PNRT following Algorithm 2 by using the minimum of $T(Y_{\mathbb{I}(d^r)}(D^{obs}), D^{obs})$ across random R assignments rather than solving the actual minimum; 4) The pairwise comparison-based RNRT with rejection based on $\alpha/2$ to ensure validity in the worst case scenario; 5) The pairwise comparison-based RNRT with rejection based on α.

To select the preferred method, two main criteria guide the testing procedure: First, under the scenario of no spillover effect ($\tau = 0$), the partial null hypothesis is true and should be rejected less than or equal to 5% of the time to maintain control over Type I errors. Second, in the presence of a spillover effect ($\tau > 0$), the partial null hypothesis is false and should be rejected as frequently as possible to maximize power. To assess power, I considered 50 τ values spaced equally from 0 to 1, conducting 2,000 simulations for each τ to compute the average rejection rate for each method. See Appendix C for detailed algorithm. I focused on displacement effects and used the non-absolute value difference-in-mean for one-sided testing.

Figure 6(left) shows that only FRT over-rejects the true partial null hypothesis when $\tau = 0$, which is consistent with the observation in Athey et al. (2018) that testing the sharp null of no effect is invalid when the actual interest lies in a partial null hypothesis. In my simulation study, with only 7 units being treated (0.7% of the total units), the rejection rate is around 10%. Surprisingly, the pairwise comparison-based PNRT without any adjustment at the level α still maintains good size control, indicating that the 2α control in the theorem is a worst-case scenario guarantee. Other PNRT algorithms are also valid but more conservative, with rejection rates below 5%. The biclique conditional randomization testing remains a valid method with a rejection rate close to 5%.

Regarding the power of the tests, FRT is excluded from the comparison due to its in-

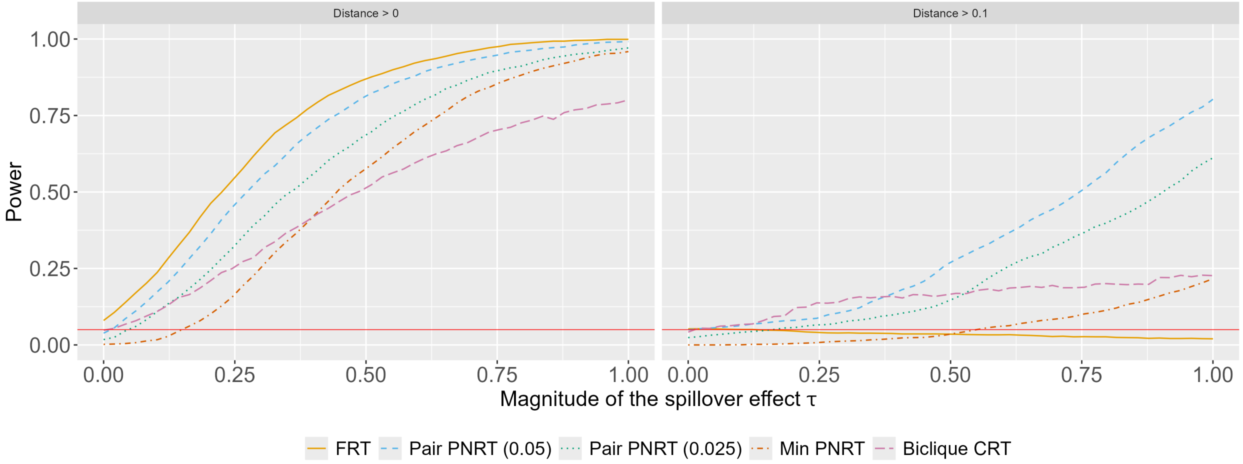


Figure 6: Left: Power Comparison for H_0^0. Right: Power Comparison for $H_0^{0.1}$. The red line represents the size level $\alpha = 0.05$. Min PNRT: Minimization-based PNRT by Algorithm 2. Pair PNRT (0.025): pairwise comparison-based RNRT with rejection based on $\alpha/2$ to ensure validity in the worst-case scenario. Pair PNRT (0.05): pairwise comparison-based RNRT with rejection based on α.

validity. The unadjusted pairwise comparison-based PNRT is the best method, dominating all others across all effect magnitudes τ. However, concerns about its validity in worst-case scenarios might persist. Among methods with theoretical size control, the pairwise comparison-based PNRT with $\alpha/2$ rejection level appears optimal, although it has slightly less power than the biclique CRT for very small τ magnitudes. This trade-off is expected when opting for more conservative testing. The minimization-based PNRT, despite being dominated by pairwise comparison-based PNRT, outperforms biclique CRT, especially for larger spillover effect magnitudes $\tau > 0.5$. Lastly, although valid, the biclique CRT lacks sufficient power, with a rejection rate below 90% even when $\tau = 1$.

Figure 6(right) contrasts the left-hand side. First, all methods are valid under the null, including FRT. This may be because hotspots rarely belong to either exposure level $(0.1, 0.2]$ and $(0.2, \infty)$. Therefore, despite a negative treatment effect, it doesn't affect the test statistics used for testing. As with H_0^0, the pairwise comparison-based PNRT and biclique CRT methods show a rejection rate close to 5%, while the pairwise comparison-based PNRT with a rejection level of $\alpha/2$ and the minimization-based PNRT remain conservative.

Second, all methods exhibit significantly lower power compared to H_0^0. This is largely because only 60% of the units are related to the partial null hypothesis this time, and the effect magnitude is only 0.5τ. Nonetheless, the pairwise comparison-based PNRT procedures still demonstrate power when the spillover magnitude τ is large enough and outperforms the other methods when using the unadjusted rejection level α. The minimization-based PNRT seems too conservative due to the minimization over an extensive amount of treatment assignments. Surprisingly, FRT shows under-rejection and almost no power for any τ. This can be explained by the intuition behind FRT rejection: the p-value is small if the observed test statistics exceed most test statistics from randomized treatment assignments. However, because units in group $(0, 0.1]$ under the observed assignment are included in test statistics for another randomized assignment d, and these units have spillover effect τ, the observed test statistics constructed from $(0.1, 0.2]$ and $(0.2, \infty)$ no longer exhibit extremely high values, even for large τ, resulting in a large p-value. This, combined with the discussion in H_0^0, illustrates that using FRT and testing the sharp null of no effect can lead to either over-rejection or under-rejection in practice. Finally, although the biclique CRT method still has power, it increases much slower than the PNRT methods. This is mainly due to the complex network structure in spatial interference, making finding a good conditioning event challenging.

Overall, the simulation results favor the PNRT procedures, especially the pairwise comparison-based PNRT without size adjustment. Therefore, I used PNRT to replicate the results from Blattman et al. (2021), employing the non-absolute difference-in-mean estimator.

5.2 PNRT on real data

I replicated the results using the publicly available dataset from Blattman et al. (2021), which includes the street-level observed treatment and their distance intervals with distance thresholds: 125m, 250m, and 500m, as well as another $1,000$ pseudo-randomized treatments and their distance intervals used in the original paper to conduct randomization inference. However, there is no data on the longitude/latitude of streets, so I cannot extend the randomization testing beyond the given $1,000$ random treatments. Since displacement effects are crucial as they influence the overall evaluation of the intervention's total welfare, this reanalysis aims to assess whether there is a displacement effect and, if so, at what distance it is significant. In Blattman et al. (2021), the authors found no displacement effect for violent crimes and a marginally significant displacement effect for property crimes. As discussed in previous sections, both using FRT for inference of the partial null and pre-selecting the pure control group without any adjustment is not guaranteed valid and can lead to different conclusions. So, how might these conclusions change if we implement a valid testing approach?

I use the pairwise comparison-based PNRT with the difference-in-mean estimator as

Table 5: Hot Spots Policing: p-values for testing the spillover effect at different distances

	Unadjusted P-values		
	$(0m, \infty)$	$(125m, \infty)$	$(250m, \infty)$
Violent crime			
Pair PNRT	0.047	0.546	0.045
Pair PNRT + reg	0.105	0.719	0.158
Min PNRT	0.074	0.832	0.518
Property crime			
Pair PNRT	0.325	0.346	0.394
Pair PNRT + reg	0.508	0.232	0.619
Min PNRT	0.471	0.809	0.882

Legend: Impact of intensive policing on violent and property crime. Pair PNRT: Pairwise comparison-based PNRT with the difference-in-mean estimator as the test statistic. Min PNRT: Minimization-based PNRT with the difference-in-mean estimator as the test statistic. Pair PNRT + reg: Pairwise comparison-based PNRT with the coefficient from the covariates-included regression, such as police station fixed effects, with inverse propensity weighting as the test statistic.

the test statistic as the main specification of the testing. Still, I also try to assess the robustness of the results when using either minimization-based PNRT with the difference-in-mean estimator as the test statistic or the pairwise comparison-based PNRT with the coefficient from a regression. Compared to the difference-in-mean estimator, the regression approach incorporates two additional factors, following Blattman et al. (2021), with a slight modification:

First, it includes the same covariates, such as control police station fixed effects, except those related to municipal services treatment.[12] Blattman et al. (2021) performed randomization testing by jointly randomizing intensive policing and municipal services rather than holding one fixed. This approach might complicate the interpretation, especially when a simple additive model cannot capture interaction effects between the two interventions. Therefore, I fixed the municipal services intervention and randomized only the intensive policing to isolate its effect.

Second, the original paper suggests using Inverse Propensity Weighting (IPW) in a

[12] The specific set of covariates include the number of crimes in 2012-2015, average patrol time per day, Sq. meters built (100m around) per meter of longitude, distance to the closest shopping center, distance to the closest educational center, distance to closest religious or cultural center, distance to the closest health center, distance to closest additional services office (i.e. justice), distance to closest transport infrastructure (i.e. bus or BRT station), the indicator for industry/commerce zone, the indicator for services sector zone, income level, eligibility of the municipal services, police station indicator, and their intersections with the crime hotspot indicator.

weighted regression. While this approach may still be biased, as it doesn't fully align with the formula provided by Aronow et al. (2020), it helps address the potential imbalance in the spillover group. However, the original specification drops around 50% of observations due to a lack of overlap conditions, potentially affecting the power of the tests. Therefore, I utilized the full sample for the regressions rather than selecting a subsample.

See Appendix D for the robustness check on different methods of incorporating covariates.

Discussion on the difference conclusion. Table 5 reveals a significant displacement effect for violent crimes but not for property crimes. After adjusting for multiple hypothesis testing using Algorithm 3, both pairwise comparison-based PNRT and minimization-based PNRT methods agree on a significant short-range spillover within 125m at the 10% level using the difference-in-mean estimator. Suppose we do not apply the $\alpha/2$ adjustment to the pairwise comparison-based PNRT, as suggested in the simulation study. In that case, the short-range spillover within 125m is significant at the 5% level with the difference-in-mean estimator. It remains marginally significant at the 10% level with the regression coefficient.[13] There is no clear evidence of additional spillover effects beyond 125m for violent crimes and no evidence of spillover effects at any distance for property crimes. The unadjusted p-value of 0.045 for the $(250m, \infty)$ interval using pairwise comparison-based PNRT might suggest a potential spillover effect within this range. However, it could also be a false discovery due to multiple hypothesis testing. Importantly, Table 5 is presented to illustrate the methodology rather than to draw definitive conclusions about the effects of hotspot policing, which would require addressing issues beyond the scope of this paper.[14]

In line with Puelz et al. (2021), p-values tend to increase with the inclusion of covariate adjustments, likely due to the heterogeneous nature of spillover effects. This observation suggests that geographic distance alone may not fully capture the intensity of these effects. In future work, we could enhance the distance measure by incorporating additional factors, such as socioeconomic disparities between street segments, as discussed in Puelz et al. (2021).

[13] This also suggests that the distance interval $(125m, \infty)$ could serve as a more appropriate control group, in contrast to the $(250m, \infty)$ interval used by Blattman et al. (2021).

[14] A potential explanation, consistent with standard economic models of crime, is that violent crime in Bogotá's hotspots may not be solely expressive violence, as implied by Blattman et al. (2021). Instead, some crimes might be highly concentrated with instrumental motives driven by generally mobile criminal rents. By increasing the risk of detection, criminals are deterred from committing crimes in specific locations, but the crime itself may simply relocate rather than be deterred. As pointed out by Blattman et al. (2021), violent crimes are often considered more severe than property crimes, making the potential displacement effect a critical consideration when evaluating the overall welfare impact of the policy intervention.

6 Conclusion

This paper introduces a straightforward testing framework for interference in network settings. The proposed tests are computationally simpler than previous methods while maintaining desirable power and size properties, making them highly practical for applied use.

Beyond network settings, the PNRT method has broader applicability. For instance, Zhang and Zhao (2021) demonstrated that partial null hypotheses are relevant in time-staggered designs. This suggests an intriguing direction for future research: extending the framework to quasi-experimental settings and observational studies. In quasi-experimental designs, a unified framework applicable to time-staggered adoption, regression discontinuity, and network settings would be invaluable (Borusyak and Hull, 2023; Kelly, 2021). In observational studies, incorporating propensity score weighting to create pseudo-random synthetic treatments and conducting sensitivity analyses would be essential, as noted by Rosenbaum (2020).

Although simulation results have shown PNRT to perform favorably compared to CRT, its power properties in broader contexts remain unexplored. Fortunately, theoretical insights from studies such as Basse et al. (2019) and Puelz et al. (2021) have highlighted the power properties of CRT, and Wen et al. (2023) has discussed the near minimax optimality of pairwise p-values. These findings suggest that further investigation into the power properties of the PNRT method could be both feasible and fruitful.

References

Abadie, A., S. Athey, G. W. Imbens, and J. M. Wooldridge (2020): "Sampling-Based versus Design-Based Uncertainty in Regression Analysis," *Econometrica*, 88, pp. 265–296.

——— (2022): "When Should You Adjust Standard Errors for Clustering?*," *The Quarterly Journal of Economics*, 138, 1–35.

Anderson, M. L. (2008): "Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects," *Journal of the American Statistical Association*, 103, 1481–1495.

Angrist, J. D. (2014): "The perils of peer effects," *Labour Economics*, 30, 98–108.

Aronow, P. M. (2012): "A General Method for Detecting Interference Between Units in Randomized Experiments," *Sociological Methods & Research*, 41, 3–16.

Aronow, P. M., D. Eckles, C. Samii, and S. Zonszein (2020): "Spillover Effects in Experimental Data," .

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

Athey, S., D. Eckles, and G. W. Imbens (2018): "Exact p-Values for Network Interference," *Journal of the American Statistical Association*, 113, 230–240.

Basse, G., P. Ding, A. Feller, and P. Toulis (2024): "Randomization Tests for Peer Effects in Group Formation Experiments," *Econometrica*, 92, 567–590.

Basse, G. and A. Feller (2018): "Analyzing Two-Stage Experiments in the Presence of Interference," *Journal of the American Statistical Association*, 113, 41–55.

Basse, G. W. and E. M. Airoldi (2018): "Limitations of Design-based Causal Inference and A/B Testing under Arbitrary and Network Interference," *Sociological Methodology*, 48, 136–151.

Basse, G. W., A. Feller, and P. Toulis (2019): "Randomization tests of causal effects under interference," *Biometrika*, 106, 487–494.

Bayer, P., S. Ross, and G. Topa (2008): "Place of Work and Place of Residence: Informal Hiring Networks and Labor Market Outcomes," *Journal of Political Economy*, 116, 1150–1196.

Blattman, C., D. P. Green, D. Ortega, and S. Tobón (2021): "Place-Based Interventions at Scale: The Direct and Spillover Effects of Policing and City Services on Crime [Clustering as a Design Problem]," *Journal of the European Economic Association*, 19, 2022–2051.

Bond, R. M., C. J. Fariss, J. J. Jones, A. D. I. Kramer, C. Marlow, J. E. Settle, and J. H. Fowler (2012): “A 61-million-person experiment in social influence and political mobilization,” *Nature*, 489, 295–298.

Borusyak, K. and P. Hull (2023): “Nonrandom Exposure to Exogenous Shocks.” *Econometrica : journal of the Econometric Society*, 91 6, 2155–2185.

Bowers, J., M. M. Fredrickson, and C. Panagopoulos (2013): “Reasoning about Interference Between Units: A General Framework,” *Political Analysis*, 21, 97–124.

Breza, E., A. G. Chandrasekhar, T. H. McCormick, and M. Pan (2020): “Using Aggregated Relational Data to Feasibly Identify Network Structure without Network Data,” *American Economic Review*, 110, 2454–84.

Caughey, D., A. Dafoe, X. Li, and L. Miratrix (2023): “Randomisation inference beyond the sharp null: bounded null hypotheses and quantiles of individual treatment effects,” *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85, 1471–1491.

Cox, D. R. (1958): *Planning of Experiments*, New York: Wiley.

Ding, P., A. Feller, and L. Miratrix (2016): “Randomization inference for treatment effect variation,” *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 78, 655–671.

Dufour, J.-M. and L. Khalaf (2003): *Monte Carlo Test Methods in Econometrics*, John Wiley & Sons, Ltd, chap. 23, 494–519.

Fisher, R. A. (1925): “Theory of Statistical Estimation,” *Mathematical Proceedings of the Cambridge Philosophical Society*, 22, 700–725.

Goldenberg, A., A. X. Zheng, S. E. Fienberg, and E. M. Airoldi (2009): “A Survey of Statistical Network Models,” *ArXiv*, abs/0912.5410.

Granovetter, M. S. (1973): “The Strength of Weak Ties,” *American Journal of Sociology*, 78, 1360–1380.

Guan, L. (2023): “A conformal test of linear models via permutation-augmented regressions,” .

Hennessy, J. P., T. Dasgupta, L. W. Miratrix, C. W. Pattanayak, and P. Sarkar (2015): “A Conditional Randomization Test to Account for Covariate Imbalance in Randomized Experiments,” *Journal of Causal Inference*, 4, 61 – 80.

Hoshino, T. and T. Yanagi (2023): “Randomization Test for the Specification of Interference Structure,” .

HUDGENS, M. G. AND M. E. HALLORAN (2008): “Toward Causal Inference with Interference,” *Journal of the American Statistical Association*, 103, 832–842.

IMBENS, G. W. AND D. B. RUBIN (2015): *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*, Cambridge University Press.

KELLY, M. (2021): “Persistence, Randomization, and Spatial Noise,” Working Papers 202124, School of Economics, University College Dublin.

LEEB, H. AND B. M. PÖTSCHER (2005): “MODEL SELECTION AND INFERENCE: FACTS AND FICTION,” *Econometric Theory*, 21, 21–59.

LEHMANN, E. L. E. L. AND J. P. ROMANO (2005): *Testing statistical hypotheses*, Springer texts in statistics., New York: Springer, 3rd ed. ed.

LI, X., P. DING, Q. LIN, D. YANG, AND J. S. LIU (2018): “Randomization Inference for Peer Effects,” *Journal of the American Statistical Association*, 114, 1651 – 1664.

LI, X., P. DING, AND D. B. RUBIN (2016): “Asymptotic theory of rerandomization in treatment–control experiments,” *Proceedings of the National Academy of Sciences*, 115, 9157 – 9162.

MANSKI, C. F. (1993): “Identification of Endogenous Social Effects: The Reflection Problem,” *The Review of Economic Studies*, 60, 531–542.

MEINSHAUSEN, N. (2008): “Hierarchical testing of variable importance,” *Biometrika*, 95, 265–278.

MIGUEL, E. AND M. KREMER (2004): “Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities,” *Econometrica*, 72, 159–217.

NEYMAN, J., K. IWASZKIEWICZ, AND S. KOŁODZIEJCZYK (2018): “Statistical Problems in Agricultural Experimentation,” *Supplement to the Journal of the Royal Statistical Society*, 2, 107–154.

OWUSU, J. (2023): “Randomization Inference of Heterogeneous Treatment Effects under Network Interference,” .

POLLMANN, M. (2023): “Causal Inference for Spatial Treatments,” .

POULIOT, G. (2024): “An Exact t-test,” .

PUELZ, D., G. BASSE, A. FELLER, AND P. TOULIS (2021): “A Graph-Theoretic Approach to Randomization Tests of Causal Effects under General Interference,” *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84, 174–204.

RAJKUMAR, K., G. SAINT-JACQUES, I. BOJINOV, E. BRYNJOLFSSON, AND S. ARAL (2022): “A causal test of the strength of weak ties,” *Science*, 377, 1304 – 1310.

ROSENBAUM, P. (2007): “Interference Between Units in Randomized Experiments,” *Journal* *of the American Statistical Association*, 102, 191–200.

——— (2020): *Design of Observational Studies*, Springer Series in Statistics, Springer International Publishing.

Sacerdote, B. (2001): "Peer Effects with Random Assignment: Results for Dartmouth Roommates*," *The Quarterly Journal of Economics*, 116, 681–704.

Sobel, M. E. (2006): "What Do Randomized Studies of Housing Mobility Demonstrate?" *Journal of the American Statistical Association*, 101, 1398–1407.

Taylor, S. J. and D. Eckles (2018): *Randomized Experiments to Detect and Estimate Social Influence in Networks*, Cham: Springer International Publishing, 289–322.

Toulis, P. and E. Kao (2013): "Estimation of Causal Peer Influence Effects," in *Proceedings of the 30th International Conference on Machine Learning*, ed. by S. Dasgupta and D. McAllester, Atlanta, Georgia, USA: PMLR, vol. 28 of *Proceedings of Machine Learning Research*, 1489–1497.

Vazquez-Bare, G. (2023): "Identification and estimation of spillover effects in randomized experiments," *Journal of Econometrics*, 237, 105237.

Viviano, D. (2022): "Experimental Design under Network Interference," .

Wang, Y., C. Samii, H. Chang, and P. M. Aronow (2023): "Design-Based Inference for Spatial Experiments under Unknown Interference," .

Wen, K., T. Wang, and Y. Wang (2023): "Residual Permutation Test for High-Dimensional Regression Coefficient Testing," .

Wu, J. and P. Ding (2021): "Randomization Tests for Weak Null Hypotheses in Randomized Experiments," *Journal of the American Statistical Association*, 116, 1898–1913.

Zhang, Y. and Q. Zhao (2021): "Multiple conditional randomization tests," *arXiv: Statistics Theory*.

——— (2023): "What is a Randomization Test?" *Journal of the American Statistical Association*, 118, 2928–2942.

Zhao, A. and P. Ding (2020): "Covariate-adjusted Fisher randomization tests for the average treatment effect," *Journal of Econometrics*.

Appendix A Review of the FRT and CRT

A.1 Review of the fisher randomization tests

The original Fisher randomization tests (FRT), as proposed by Fisher (1925), was designed for a binary treatment scenario without interference. In this framework, each $Y_i(d)$ depends solely on d_i, resulting in only two potential outcomes: potential outcome when in the treatment group $Y_i(1)$ and potential outcome when in the control group $Y_i(0)$ for every unit i. The standard approach to testing whether the treatment has an effect typically involves the following null hypothesis:

$$H_0 : Y_i(0) = Y_i(1), i = 1, 2, ..., N,$$

which is a special case of the null hypothesis in Definition 1.

Let $T(Y, D) : R^N \times \mathbb{D} \to \mathbb{R}$ denote a test statistic as the function of Y and D, typically the differences in mean, rank statistics, etc. For instance, an example test statistic could be the absolute difference in means between treated and control units:

$$T(Y^{obs}, D) = \|\bar{Y}^{obs}_{\{i:D_i=1\}} - \bar{Y}^{obs}_{\{i:D_i=0\}}\| \tag{A.1}$$

where $\bar{Y}^{obs}_{\{i:D_i=1\}} = \frac{\sum_{i=1}^{N} 1\{D_i=1\}Y_i}{\sum_{i=1}^{N} 1\{D_i=1\}}$, $\bar{Y}^{obs}_{\{i:D_i=0\}} = \frac{\sum_{i=1}^{N} 1\{D_i=0\}Y_i}{\sum_{i=1}^{N} 1\{D_i=0\}}$. In practice, we can also use the non-absolute value version of test statistics for the one-side testing.

Denote $T_{obs} = T(Y^{obs}, D^{obs})$. The p-value is then defined as $pval(D^{obs}) = P(T(Y^{obs}, D) \geq T_{obs})$ respect to $D \sim P(D)$, reflecting a stochastic version of the "proof by contradiction" method discussed by Imbens and Rubin (2015): If there are few potential assignments D with $T(Y^{obs}, D) \geq T_{obs}$, it suggests that observing the current value T_{obs} under the null hypothesis is highly improbable. Consequently, the p-value would be lower, increasing the likelihood of rejecting the null hypothesis. The formal testing procedure can be outlined as follows, where the p-value is calculated as the mean value of $1 + R$ draws due to using $d = D^{obs}$ for $r = 0$, so there are $R + 1$ draws:

Observe that, if H_0 is true, one must have $T_r = T(Y^{obs}, D') = T(Y(D'), D')$ for any $D' \sim P(D')$. Since D' is a random draw from $P(D)$, we have $T_r = T(Y(D'), D') \sim T(Y^{obs}, D^{obs}) = T_{obs}$, leading to an valid test at any level α, where $P\{pval \leq \alpha\} \leq \alpha$, for all $\alpha \in [0, 1]$ when the null hypothesis is true. A formal proof can be found in Basse et al. (2019) and Zhang and Zhao (2023).

Algorithm 4 Fisher Randomization Tests (FRT)

```
Inputs      : Test statistic $T = T(Y(d), d)$, observed assignment $D^{obs}$, observed
              outcome $Y^{obs}$, treatment assignment mechanism $P(D)$, and size $\alpha$.
Compute     : The observed test statistic, $T_{obs} = T(Y^{obs}, D^{obs})$.
for $r = 1$ to $R$ do
    Randomly sample: $d^r \sim P(D)$, Store $T_r \equiv T(Y^{obs}, d^r)$.
end
Output      : P-value: $\hat{pval} = (1 + \sum_{r=1}^{R} 1\{T_r \geq T_{obs}\})/(1 + R)$.
              Reject if p-value$\leq \alpha$.
```

Table A6: Illustration of FRT in the toy example

Assignment D	Potential Outcome Y_i						$T(Y^{obs}, D)$
	i_1	i_2	i_3	i_4	i_5	i_6	
$(1, 0, 0, 0, 0, 0)$	2	5	3	1	4	6	1.8
$(0, 1, 0, 0, 0, 0)$	2	5	3	1	4	6	1.8
$(0, 0, 1, 0, 0, 0)$	2	5	3	1	4	6	0.6
$(0, 0, 0, 1, 0, 0)$	2	5	3	1	4	6	3
$(0, 0, 0, 0, 1, 0)$	2	5	3	1	4	6	0.6
$(0, 0, 0, 0, 0, 1)$	2	5	3	1	4	6	3

Legend: Potential outcome schedule for the toy example: Assignment D includes all the potential assignments with the first row as the observed assignment D^{obs}; Potential outcomes are the same for each potential assignment under the sharp null of no effect; $T(Y^{obs}, D)$ is the absolute value of the difference in mean defined in equation A.1.

A toy example (cont.) Under SUTVA, without any interference, Table A6 illustrates the potential outcome schedule for FRT in the toy example. Following the algorithm 4, the observed test statistics is 1.8. $T(Y^{obs}, D)$ when unit i_4 and i_6 are treated have values of 3, which are larger than 1.8, and have the same value of 1.8 when unit i_2 is treated. So, the P-value is 2/3, and we can use a uniform random variable in practice for the tie-break without influencing the validity of the testing (Lehmann and Romano (2005)). One can also repeat the testing procedure with the non-absolute difference for one-side testing.

What if we use FRT for partial null hypothesis? For the partial null hypothesis, as illustrated in the appendix by Athey et al. (2018), the FRT procedure might over-reject under the null. The main reason is that implementing FRT for a partial null hypothesis mistakenly treats the missing potential outcomes, as illustrated in Table 1, with the sharp null of no effect. Hence, the rejection of the test ignores the variation arising from the treatment effects. For example, Bond et al. (2012) tested for spillovers from a randomly assigned encouragement to vote in the 2010 U.S. elections. This was implemented as a permutation test that implicitly assumed the absence of direct effects. Even though Bond et al. (2012) elsewhere rejected that null hypothesis, Athey et al. (2018) show that such tests can dramatically inflate Type I error rates.

A.2 Review of conditional randomization tests

Pioneered by Aronow (2012) and Athey et al. (2018), recent literature, including Basse et al. (2019) and Puelz et al. (2021), has turned to Conditional Randomization Tests (CRT) to tackle various kinds of network interference settings. The key idea in this line of research is that although the null hypothesis $H_0^{\epsilon_s}$ is not sharp in general, it can be "made sharp" by restricting our attention to a well-chosen conditioning event (Basse et al., 2019): $\mathbb{U} = (\mathbb{N}_{\mathbb{U}}, \mathbb{D}_{\mathbb{U}}) \sim P(\mathbb{U}|D^{obs})$. This event includes a subset of units, called *focal units* ($\mathbb{N}_{\mathbb{U}} \subseteq \{1, \ldots, N\}$), and a subset of assignments, called *focal assignments* ($\mathbb{D}_{\mathbb{U}} \subseteq \{0,1\}^N$), that satisfy Definition A11:

Definition A11 (Conditions for Conditioning Event)**.** *Given partial null hypothesis $H_0^{\epsilon_s}$. For any conditioning Event $\mathbb{U} = (\mathbb{N}_{\mathbb{U}}, \mathbb{D}_{\mathbb{U}}) \in 2^{\{1,\ldots,N\}} \times 2^{\{0,1\}^N}$, we have for any $i \in \mathbb{N}_{\mathbb{U}}$, $Y_i(d) = Y_i(d')$ for any $d, d' \in \mathbb{D}_{\mathbb{U}}$ under $H_0^{\epsilon_s}$.*

Definition A11 combined Definition 3 and Section 4 from Athey et al. (2018), emphasizing the restrictive nature of the conditioning event. Theorem 3 from Basse et al. (2019) reinterprets this condition in the context of exposure mapping, a low-dimensional summary of the treatment assignments, which can be misspecified in practice. Both Athey et al. (2018) and Basse et al. (2019) allow $\mathbb{N}_{\mathbb{U}} \nsubseteq \mathbb{I}(D^{obs})$. However, they need to restrict the set of focal assignments to fix the exposure levels of units not in $\mathbb{I}(D^{obs})$, and they do not use this information in the test statistics. The final implementation in Basse et al. (2019) still uses a set $\mathbb{N}_{\mathbb{U}} \subseteq \mathbb{I}(D^{obs})$.

Different papers have different $P(\mathbb{U}|D^{obs})$ to choose $\mathbb{U}$: Both Aronow (2012) and Athey et al. (2018) only considered conditioning mechanisms of the form $P(\mathbb{U}|D^{obs}) = P(\mathbb{U})$, where conditioning is either random or guided by known auxiliary information but is not conditioned on the observed assignment. This failure to use all the observed information would cause a loss of power. Basse et al. (2019) identified this weakness and proposed a two-step conditional mechanism tailored to cluster interference. They formalized the idea as sampling from a carefully constructed distribution $P(\mathbb{U}|D^{obs})$ and then ran a test conditional on $\mathbb{U}$. Puelz et al. (2021) extended this framework using the *Biclique decomposition method* to construct the conditioning event for general interference, including both clustered and spatial interference. Both Basse et al. (2019) and Puelz et al. (2021) constructed $\mathbb{U} \sim P(\mathbb{U}|D^{obs})$ with CRT restricted to the conditioning event $\mathbb{U}$ and use the restricted test statistic under the conditioning event for further comparison:

Definition A12 (Conditioning Event Restricted Test Statistics)**.** *Let $T^{\mathbb{U}}(Y, d) : R^N \times \{0, 1\}^N \to R$ be a measurable function. $T^{\mathbb{U}}$ is said to be Conditioning Event Restricted Test Statistics on $\mathbb{U}$ if $T^{\mathbb{U}}(Y, d) = T^{\mathbb{U}}(Y', d')$, for any $(Y, Y', d, d') \in R^{2N} \times \{0, 1\}^{2N}$ such that $Y_i = Y_i', d_i = d_i'$ for all $i \in \mathbb{N}_{\mathbb{U}}$*

The test statistics in Definition A12 is similar to the pairwise imputable statistics in Definition 7: The value of the test statistic is only related to the units in $\mathbb{N}_{\mathbb{U}}$. Therefore, the p-value is constructed similarly to FRT by restricting everything on $\mathbb{U}$, following the CRT procedure below, where the p-value is calculated as the mean value of $1 + R$ draws due to using $d = D^{obs}$ for $r = 0$, so there are $R + 1$ draws:

Algorithm 5 Conditional Randomization Testing (CRT)

```
Inputs     : Test statistic $T = T(Y(d), d)$, observed assignment $D^{obs}$, observed
             outcome $Y^{obs}$, treatment assignment mechanism $P(D)$, size $\alpha$, and
             design of conditioning event $P(\mathbb{U}|D^{obs})$.
Draw       : $\mathbb{U} \sim P(\mathbb{U}|D^{obs})$.
Compute    : The observed test statistic, $T^{\mathbb{U}}_{obs} = T^{\mathbb{U}}(Y^{obs}, D^{obs})$.
for $r = 1$ to $R$ do
    Randomly sample: $d^r \sim P(D|\mathbb{U}) \propto P(\mathbb{U}|D)P(D)$, Store $T^{\mathbb{U}}_r \equiv T^{\mathbb{U}}(Y^{obs}, d^r)$.
end
Output     : P-value: $\hat{pval} = (1 + \sum_{r=1}^{R} 1\{T^{\mathbb{U}}_r \geq T^{\mathbb{U}}_{obs}\})/(1 + R)$.
             Reject if p-value$\leq \alpha$.
```

A toy example (cont.) By Definition A11, we need $\forall d \in \mathbb{D}_{\mathbb{U}}, i \in \mathbb{N}_{\mathbb{U}}$, $d_i = 0$ for H_0^0: If there exist i with $d_i = 1$, for any other $d' \in \mathbb{D}_{\mathbb{U}}$, we need $d_i' = 1$. However, by experimental design, only one unit would be treated, so $d = d' \forall d, d' \in \mathbb{D}_{\mathbb{U}}$. This essentially results in one effective treatment, causing no power.

Following Definition A12, difference-in-mean estimator can be used as formula below:

$$T^{\mathbb{U}}(Y^{obs}, D) = \|\bar{Y}_{\mathbb{N}_{\mathbb{U}}}(D^{obs})_{\{i:D \in \mathcal{D}_i(0)/\mathcal{D}_i(1)\}} - \bar{Y}_{\mathbb{N}_{\mathbb{U}}}(D^{obs})_{\{i:D \in \mathcal{D}_i(1)\}}\|$$

Given units i_1 is treated in the observed D^{obs}, one example of a valid $\mathbb{U}$ would be choosing $\mathbb{N}_{\mathbb{U}} = \{i_2, i_4, i_6\}$, and $\mathbb{D}_{\mathbb{U}} = \{(1, 0, 0, 0, 0, 0), (0, 0, 1, 0, 0, 0), (0, 0, 0, 0, 1, 0)\}$. We can construct the potential outcome table as in Table A7.

Table A7: CRT in the toy example

Dist. to the Treated Unit	Potential Outcome Y_i			$T^{\mathbb{U}}(Y^{obs}, D)$
	i_2	i_4	i_6	
$(0, 1, 2, 3, 2, 1)$	5	1	6	4.5
$(2, 1, 0, 1, 2, 3)$	5	1	6	3
$(2, 3, 2, 1, 0, 1)$	5	1	6	1.5

Legend: Dist. to the Treated Unit: the minimum distance of each unit to the treated units. j means unit is distance j away from the treated units and belongs to the distance interval $(j-1, j]$ for $j = 1, 2, 3$. 0 means the unit itself is being treated in the randomized D. Potential Outcome Y_i: Potential outcome of each unit under the null H_0^0 with red ? as missing values. Blue cells are the units used to calculate the mean value in the first term of the test statistics. $T^{\mathbb{U}}(Y^{obs}, D)$: test statistics under different D and fixing D^{obs} that unit i_1 is treated.

According to Algorithm 5, the p-value equals 1/3 since the observed test statistic is the highest. However, in practice, designing $\mathbb{U}$ and $P(\mathbb{U}|D^{obs})$ in more complex settings to ensure nontrivial power can be challenging.

Appendix B Proof of the Theorems

Proof of Theorem 1. Given any $\alpha > 0$, consider the subset of assignment

$$\mathbb{D} \equiv \{D^{obs} | pval^{pair}(D^{obs}) \leq \alpha/2\}.$$

Therefore, we can denote $P(pval^{pair}(D^{obs}) \leq \alpha/2) = \sum_{D^{obs} \in \mathbb{D}} P(D^{obs}) = x$. To prove the theorem, we want to show $x \leq \alpha$.

Denote $H(D^{obs}, D) = 1\{T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), D) \geq T(Y_{\mathbb{I}(D)}(D), D^{obs})\}$, then by construction, $H(D^{obs}, D) + H(D, D^{obs}) \geq 1$.

Under $H_0^{\epsilon_s}$, by proposition 1 and the Definition 8 of p-value,

$$pval^{pair}(D^{obs}) = \sum_{D \in \{0,1\}^N} H(D^{obs}, D) P(D).$$

Now, consider the term

$$\sum_{D^{obs} \in \mathbb{D}} \sum_{D \in \{0,1\}^N} H(D^{obs}, D) P(D) P(D^{obs})$$

On the one hand, it equals to

$$\sum_{D^{obs}\in\mathbb{D}} pval^{pair}(D^{obs})P(D^{obs}) \leq (\alpha/2)(\sum_{D^{obs}\in\mathbb{D}} P(D^{obs})) = x\alpha/2$$

On the other hand, by flipping D and D^{obs} in the same set $\mathbb{D}$:

$$\begin{aligned}\sum_{D^{obs}\in\mathbb{D}}\sum_{D\in\mathbb{D}} H(D^{obs},D)P(D)P(D^{obs}) &= \sum_{D\in\mathbb{D}}\sum_{D^{obs}\in\mathbb{D}} H(D,D^{obs})P(D^{obs})P(D)\\ &= \sum_{D\in\mathbb{D}}\sum_{D^{obs}\in\mathbb{D}} H(D,D^{obs})P(D)P(D^{obs})\\ &= \sum_{D^{obs}\in\mathbb{D}}\sum_{D\in\mathbb{D}} H(D,D^{obs})P(D)P(D^{obs})\end{aligned}$$

Hence, we would have:

$$\begin{aligned}\sum_{D^{obs}\in\mathbb{D}}\sum_{D\in\{0,1\}^N} H(D^{obs},D)P(D)P(D^{obs}) &\geq \sum_{D^{obs}\in\mathbb{D}}\sum_{D\in\mathbb{D}} H(D^{obs},D)P(D)P(D^{obs})\\ &= \sum_{D^{obs}\in\mathbb{D}}\sum_{D\in\mathbb{D}}(H(D,D^{obs})+H(D^{obs},D))P(D)P(D^{obs})/2\\ &\geq \sum_{D^{obs}\in\mathbb{D}}\sum_{D\in\mathbb{D}} P(D)P(D^{obs})/2 = x^2/2\end{aligned}$$

Hence, $x^2/2 \leq x\alpha/2$ implying $x \leq \alpha$. As mentioned before, using 1/2 to discount the number of equalities doesn't affect the validity of the test because $H(D^{obs},D)+H(D,D^{obs}) \geq 1$ would still hold.

□

Too many potential treatment assignments. When the number of units N is large, there would be 2^N potential treatment assignments,which is a large number in practice. In such cases, given D^{obs} and Algorithm 1, we can show that $\|\hat{pval}^{pair} - pval^{pair}(D^{obs})\| = O_p(R^{-1/2})$. Specifically, by $\hat{pval}^{pair} = (1+\sum_{r=1}^{R} 1\{T_r \geq T_r^{obs}\})/(1+R)$ and $d^r \sim P(D)$ independently, we have $E_{d^r}\hat{pval}^{pair} = pval^{pair}(D^{obs})$ and

$$Var(\hat{pval}^{pair}) = Var(1\{T_r \geq T_r^{obs}\})/(1+R) = pval^{pair}(D^{obs})(1-pval^{pair}(D^{obs}))/(1+R)$$

Hence, by Chebyshev's inequality, $\|\hat{pval}^{pair} - pval^{pair}(D^{obs})\| = O_p(R^{-1/2})$.

Proof of Theorem 2. To avoid confusion, denote $P_{D^{obs}}$ as probability respect to D^{obs} and P_D as probability respect to D.

Under the null $H_0^{\epsilon_s}$, by Proposition 1 and setting $d = D$, $d' = D^{obs}$, we have $T(Y_{\mathbb{I}(D)}(D), D^{obs}) = T(Y_{\mathbb{I}(D)}(D^{obs}), D^{obs})$. Hence, we have $\tilde{T}(D^{obs}) = min_{d\in\{0,1\}^N}(T(Y_{\mathbb{I}(d)}(d), D^{obs}))$.

Then, by construction, $\tilde{T}(D^{obs}) \sim \tilde{T}(D) \leq T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), D)$, and

$$pval^{min}(D^{obs}) = P_D(T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), D) \geq \tilde{T}(D^{obs})) \geq P_D(\tilde{T}(D) \geq \tilde{T}(D^{obs}))$$

Therefore,

$$P_{D^{obs}}(pval^{min}(D^{obs}) \leq \alpha) \leq P_{D^{obs}}(P_D(\tilde{T}(D) \geq \tilde{T}(D^{obs})) \leq \alpha)$$

Let U be a random variable with the same distribution as $\tilde{T}(D)$ as induced by $P(D)$, F_U be its cumulative distribution function, we have $P_D(\tilde{T}(D) \geq \tilde{T}(D^{obs})) = 1 - F_U\{\tilde{T}(D^{obs})\}$, which is a random variable induced by $D^{obs} \sim P(D^{obs})$. Hence, $P_D(\tilde{T}(D) \geq \tilde{T}(D^{obs})) = 1 - F_U(U)$, and by the probability integral transformation, $P_D(\tilde{T}(D) \geq \tilde{T}(D^{obs}))$ respect to D^{obs} has a uniform $[0, 1]$ distribution under $H_0^{\epsilon_s}$. So, for any $\alpha \in [0, 1]$

$$P_{D^{obs}}(pval^{min}(D^{obs}) \leq \alpha) \leq P_{D^{obs}}(P_D(\tilde{T}(D) \geq \tilde{T}(D^{obs})) \leq \alpha) \leq \alpha$$

□

Appendix C Other algorithms in the paper

Algorithm in Blattman et al. (2021) for pure control group Other approaches, like the one used by Blattman et al. (2021), often employ a prespecified rule by starting with the null hypothesis $H_0^{\epsilon_{K-1}}$ and collapsing any unrejected nulls into a single control condition. However, this method might encounter issues with post-model-selection inference, leading to over-rejection under the null.

In Blattman et al. (2021), they implement Algorithm 6 with $K = 2$ and $(\epsilon_0, \epsilon_1, \epsilon_2) = (0, 250m, 500m)$.

Following Algorithm 6, the procedure involves two steps: In the first step, we collect $pval^1$ for H_0^{250} and reject H_0^{250} if $pval^1 \leq \alpha$. If H_0^{250} is rejected,the process terminates, and we report that $\epsilon_c = 500$. If H_0^{250} is not rejected, we proceed to the second step by collecting $pval^0$ for H_0^0 and reject H_0^0 if $pval^0 \leq \alpha$. If H_0^0 is rejected, we report $\epsilon_c = 250$; If H_0^0 is not

Algorithm 6 A procedure for pure control group

```
Inputs      : Test statistic $T = T(Y(d), d)$, observed assignment $D^{obs}$, observed
              outcome $Y^{obs}$, treatment assignment mechanism $P(D)$.
Set         : $\hat{K} = K$.
for $k = K - 1$ to 0 do
    Testing $H_0^{\epsilon_k}$ using PNRT procedure, collect $pval^k$.
    If $pval^k \leq \alpha$, reject $H_0^{\epsilon_k}$, Terminate.
    If $pval^k > \alpha$, set $\hat{K} = k$.
end
Output      : Set pure control group with $\epsilon_c = \epsilon_{\hat{K}}$.
```

rejected, we report there is no significant interference whatsoever.

As illustrated in Algorithm 6, this procedure does not incorporate any size adjustment for multiple hypothesis testing issues. Consequently, it is possible to over-reject the partial null hypothesis, leading to an ϵ_c larger than the true distance for the pure control group. Specifically, for any given $\tilde{k}$ where $H_0^{\tilde{k}}$ is true, the probability $p(\hat{K} \geq \tilde{k} + 1)$ exceeds α. For example, if there is no spillover and $\tilde{k} = 0$, $p(\hat{K} \geq 1) > \alpha$ due to multiple hypothesis testing. In extreme cases, if $P(pval(D^{obs}) \leq \alpha | \hat{K} \geq \tilde{k} + 1)$ is close to 1, the true null hypothesis could be over-rejected using this pre-selection procedure.

Algorithm for simulation exercise in Section 5.1 As outlined in Algorithm 7:

Algorithm 7 Simulation Study Procedure

```
Inputs      : 5000 randomly chose assignments as the potential assignments set,
              $\mathbb{D}_S$.
              The biclique decomposition of $\mathbb{D}_S$ for Puelz et al. (2021).
Set         : Spillover effect $\tau$ and corresponding schedule of potential outcomes.
for $s = 1 : S$ do
    Sample $D_s^{obs}$ from $\mathbb{D}_S$, and generate $Y_s^{obs}$.
    Implement the algorithms and collect corresponding $pval(D_s^{obs})$ using $R = 1000$.
    Average the # of rejections to get the power for that fixed $\tau$.
end
Output      : Power plot of each algorithm.
```

Appendix D Incorporating covariate adjustment

In practice, we often have access to covariates X, and incorporating this information is crucial for enhancing the power of tests, particularly when these covariates are predictive of potential outcomes (Wu and Ding, 2021). Since the choice of test statistic does not affect the validity of the testing procedure for the partial null hypothesis of interest, I propose three approaches for incorporating covariates in the analysis:

The first approach is *PNRT with regression.* As illustrated in the main text, this method involves conducting PNRT using regression coefficients from a simple OLS model as the test statistic. This OLS model includes a binary variable indicating whether a unit receives spillovers at a certain distance and known covariates, such as information about the neighborhood and social center points. A similar approach is discussed in Puelz et al. (2021).

The second approach is *PNRT with residuals outcome.* The key idea here is to use the residuals from a model-based approach, such as regression with covariates of interest, rather than the raw outcome variables. We first obtain predicted values $\hat{Y}_i$ for the sample outcomes and then use the residuals, defined as the difference between observed outcomes and predicted values $\hat{e}_i = Y_i^{obs} - \hat{Y}_i$, for the PNRT procedures as the Y defined in the main text. A similar approach for FRT is proposed by Rosenbaum (2020), with detailed discussion in Sections 7 and 9.2 of Basse and Feller (2018).

The third approach is *PNRT with pairwise residuals.* In this method, for each pair of treatment assignments (D^{obs}, D), we conduct a regression with covariates within the imputable units set to transform the outcomes into residuals before testing and constructing the p-values accordingly. This approach can be viewed as combining the first and second methods.

As shown in Table D8, the p-values are very similar across the different methods, allowing researchers to choose the most practical implementation. Additionally, as discussed in Section C.3 of Basse et al. (2024), one can stratify potential assignments based on covariates to balance the focal units. This is done by stratifying both the permutations and the test statistic by an additional discrete covariate. However, we could not implement and compare p-values from this method due to limitations in the original dataset.

Similar to the findings in Puelz et al. (2021), I observed that p-values increased after controlling for covariates. This suggests that covariates help control spillover effects, indicating that geographic distance alone may be insufficient to capture the intensity of spillovers. This implies the presence of heterogeneous spillover effects that cannot be fully captured by the partial null hypothesis defined at the unit level. In an extreme case, if the spillover effect is perfectly correlated with covariates, the underlying partial null hypothesis would be rejected, as the spillover effect exists. However, regression adjustment might eliminate the nonzero spillover effect, leading to increased p-values under the same partial null hypothesis.

Table D8: P-values for pairwise comparison-based PNRT with different specifications

	Unadjusted P-values		
	$(0m, \infty)$	$(125m, \infty)$	$(250m, \infty)$
Violent crime			
Reg (WLS)	0.105	0.719	0.158
Reg (OLS)	0.156	0.767	0.110
Pair residuals	0.119	0.726	0.142
Residuals outcome	0.114	0.757	0.166
Property crime			
Reg (WLS)	0.508	0.232	0.619
Reg (OLS)	0.494	0.462	0.560
Pair residuals	0.481	0.252	0.565
Residuals outcome	0.455	0.250	0.578

Legend: P-values of pairwise comparison-based PNRT across different methods. Reg (WLS): PNRT with regression, using the coefficient from the covariates-included regression with inverse propensity weighting as the test statistic. Reg (OLS): PNRT with regression, using the coefficient from the covariates-included regression without weighting as the test statistic. Pair residuals: PNRT with pairwise residuals, where residuals are constructed from the pairwise subset regression in the first step. The coefficient from the no-covariates regression with inverse propensity weighting is then used as the test statistic. Residuals outcome: PNRT with residuals outcome, where residuals are constructed for all units in the first step, followed by using the coefficient from the no-covariates regression with inverse propensity weighting as the test statistic.

Researchers should interpret these results cautiously and decide on the null hypothesis of interest beforehand. If a researcher is interested in testing for no spillover effects after controlling for covariates, PNRT can be extended to accommodate the work by Ding et al. (2016). One can refer to Owusu (2023) for investigating heterogeneous effects in network settings. Alternatively, if interested in the weak null of the average effect being equal to zero (see Zhao and Ding (2020); Basse et al. (2024)), one should note that the construction of p-values in PNRT differs from those in CRT and FRT, making classical approaches for weak nulls potentially inapplicable. Further investigation into these differences would be of interest to future research.