Unconditional Randomization Tests for Interference

Liang Zhong*

October 26, 2024

(Click here for the most recent version)

Abstract

When conducting causal inference or designing policy, researchers are often concerned with the existence and extent of interference between units, influenced by factors such as distance, proximity, and connection strength. However, complex correlations across units pose significant challenges for inference. This paper introduces partial null randomization tests (PNRTs), a novel framework for testing interference in experimental settings. PNRTs adopt a design-based approach, combining unconditional randomization testing with pairwise comparisons to enable straightforward implementation and ensure finite-sample validity under minimal assumptions about network structure. To illustrate the method's broad applicability, I apply it to a large-scale experiment by Blattman et al. (2021) in Bogotá, Colombia, which evaluates the impact of hotspot policing on crime using street segments as units of analysis. I find that increasing police patrolling time in hotspots has a significant displacement effect on violent crime but not on property crime. A simulation study calibrated to this dataset further demonstrates the strong power properties of PNRTs and their suitability for 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, Benjamin Marx, Haoran Pan, Lei Ma, Justin Hong, Qingyuan Chai, Mohsen Bayati, Jizhou Liu, and the participants of the 19th Economics Graduate Student Conference, the Econometrics seminar and Empirical Micro workshop at Boston University for their useful comments. All errors are my own.

1 Introduction

The treatment assigned to one unit may influence the outcome of another, a phenomenon known as interference. Recognizing both the existence and extent of interference is crucial for accurate causal inference.[1] For example, when a student hires a tutor, the benefits may extend beyond the individual student to their classmates, potentially influencing the academic performance of friends or even friends of friends. Identifying whether and to what extent such spillovers occur can refine estimates of treatment effects. Additionally, understanding interference can facilitate more efficient resource allocation in policy design. For instance, recognizing that a tutor's influence extends to students' friends can guide the distribution of educational resources within classrooms. Consequently, researchers are often interested in testing for the presence of interference and assessing its extent, often based on factors such as distance, proximity, and the strength of connections between units.[2]

However, testing for interference presents significant econometric challenges, as interference violates the Stable Unit Treatment Value Assumption (SUTVA), a cornerstone of most causal inference frameworks. A common approach is to compare outcomes across varying distances from treated units, yet large sample approximations for test statistics can become intractable due to complex clustering patterns (Kelly, 2021). Also, even in the random experiment, the validity of the approach might require additional assumptions beyond the randomness of the treatment assignment (Aronow, 2012; Pollmann, 2023). These challenges can render traditional methods unreliable or uninformative, underscoring the need for more robust approaches to test interference.

In this paper, I introduce partial null randomization tests (PNRTs), a novel unconditional randomization testing framework designed to detect and analyze interference patterns in experimental settings. PNRTs adopt a design-based approach, leveraging the randomness of treatment assignment as the source of uncertainty for inference while treating potential outcomes as fixed constants (Abadie et al., 2020, 2022). I define the *partial null hypothesis* to assess interference, comparing a unit's potential outcomes under different treatment assignments when the unit is beyond a certain distance threshold, ϵ_s, from any treated unit. This hypothesis is tested using a randomization procedure that reassigns treatments

[1] See, for example, Angrist (2014), Sacerdote (2001), Cai et al. (2015), Paluck et al. (2016), Miguel and Kremer (2004), and Jayachandran et al. (2017).

[2] For example, Bond et al. (2012) investigate whether spillover effects extend beyond immediate friends to friends of friends. Some theoretical work, such as Toulis and Kao (2013), assumes that such higher-order spillovers do not occur. Rajkumar et al. (2022) examine how job mobility relates to link intensity, distinguishing between strong and weak ties.

to units, keeping outcomes fixed, and calculates a carefully designed test statistic for each reassignment. The observed test statistic is then compared to the distribution of randomized statistics to determine significance, with the p-value constructed as the proportion of treatment assignments where the test statistic exceeds the observed value. If the observed statistic sufficiently deviates from the randomized distribution, it provides evidence against the null hypothesis.

PNRTs are finite-sample valid, non-parametric methods that address primary challenges of traditional approaches.[3] The method requires minimal assumptions about network structure, making it suitable for dense networks. Moreover, it relies solely on the randomness of treatment assignment, ensuring validity without further assumptions. Given its robustness, I propose PNRTs as a benchmark method for analyzing interference.

Recent studies, such as Bond et al. (2012) and Blattman et al. (2021), highlight the value of randomization tests for detecting interference. As Blattman et al. (2021) note, a design-based approach and randomization inference may be particularly suitable in network settings where spillover effects are unknown.[4] However, these studies typically rely on Fisher randomization tests (FRTs), which are not always valid when testing for interference (Athey et al., 2018). The core issue lies in the null hypothesis. FRTs are designed to test the *sharp null hypothesis* of no effect and rely on imputability, meaning all potential outcomes are assumed to be known across all treatment assignments under the null (Rosenbaum, 2007; Hudgens and Halloran, 2008). In network settings, this implies that no unit's potential outcome changes regardless of who is treated, thereby excluding both treatment effects and interference under the sharp null. As a result, rejections of the sharp null in FRTs could indicate either nonzero treatment effects or interference, without distinguishing between the two.

In contrast, testing specifically for interference requires isolating potential treatment effects while assessing whether a unit's outcome is influenced by the treatment statuses of others. This involves testing partial null hypotheses, where only a subset of potential outcomes is assumed to be known across different treatment assignments (Zhang and Zhao, 2023). However, this approach introduces two technical challenges. First, only a limited number of potential outcomes—those of a subset of units—are "imputable," meaning that we

[3]It is finite-sample exact, meaning the probability of a false rejection in finite samples does not exceed the user-prescribed nominal rate (Pouliot, 2024).

[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."

can infer their values based on the observed data under the partial null. For example, under the partial null hypothesis of no peer effects in a social network, we can only impute outcomes for non-treated people, while outcomes for treated people remain unknown. Second, the set of units with imputable outcomes varies with each treatment assignment. This variability arises because the set of non-treated units would be different across different treatment assignments. Together, these challenges make it difficult to apply traditional randomization inference methods directly and highlight the need for more specialized approaches.

Acknowledging the limitations of FRTs for testing interference, prior literature has introduced conditional randomization tests (CRTs), which restrict testing to a conditioning event—a subset of units and assignments where the null hypothesis is sharp.[5] Various procedures have been proposed for designing these conditioning events to ensure finite-sample exactness. However, many CRT methods are tailored to specific scenarios, such as clustered interference (Basse et al., 2019, 2024), limiting their generalizability. Additionally, designing conditioning events that ensure non-trivial power is challenging, often resulting in power loss (Puelz et al., 2021). Furthermore, implementing CRTs for general interference can be computationally demanding, requiring significant time and resources.

This paper's primary contribution is the development of valid testing procedures for partial null hypotheses without relying on conditioning events. PNRTs address the two technical challenges posed by partial null hypotheses and offer three key advantages: broad applicability, the elimination of complex conditioning requirements, and straightforward implementation.

To address the first challenge, I propose *pairwise imputable statistics*—a bivariate function $T(D_1, D_2)$, where the first argument, D_1, selects the set of imputable units and the second argument, D_2, determines how those units are grouped or compared. These statistics resemble conventional test statistics as defined by Imbens and Rubin (2015), but are restricted to only the imputable units, as specified by the partial null hypothesis, under both D_1 and D_2. Despite this restriction, pairwise imputable statistics can accommodate various test statistics commonly used under sharp null hypotheses. For example, in the difference-in-means estimator, I would compare the imputable individuals who are friends of at least one treated individual with those who are not friends of any treated individual. Here, D_1 determines which individuals are imputable and included in the computation, while D_2 functions similarly to conventional test statistics in determining group assignments, implicitly

[5]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), and Hoshino and Yanagi (2023).

excluding treated units under D_2. It is important to note that this framing is a simplification of the actual definition, which will be introduced formally later in the paper. The simplified version here helps highlight the key concept of imputability, whereas the formal definition later will involve a more detailed notation reflecting how the test statistic interacts with the imputable outcome vector across treatment assignments.

However, the second challenge—variation in the set of imputable units across different treatment assignments—complicates the use of pairwise imputable statistics in FRTs. Specifically, p-values are constructed by comparing the observed test statistic $T(D^{obs}, D^{obs})$ with randomized test statistics $T(D^{obs}, D)$, where D is drawn randomly from the design distribution. However, $T(D^{obs}, D^{obs})$ belongs to the same distribution as $T(D, D)$, which differs from $T(D^{obs}, D)$, even under the partial null hypothesis, because the set of imputable units is generally different for D and D^{obs}. This variability in the set of imputable units makes it difficult for simple implementations of unconditional randomization tests to effectively control size.

To tackle this challenge, 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 observed and potential assignment pair (D^{obs}, D). The key distinction between these terms lies in how the two arguments are used: $T(D, D^{obs})$ uses the potential assignment D to determine the set of imputable units, while D^{obs} determines how those units are grouped and compared. Conversely, $T(D^{obs}, D)$ uses the observed assignment D^{obs} to select the imputable units, and the potential assignment D defines how those units are grouped or compared. Since pairwise imputable statistics rely only on units that are imputable under both the observed and randomized assignments, both terms are computable under the partial null hypothesis. The validity of this procedure is established through the symmetry of these pairwise comparisons, akin to the conformal lemma in Guan (2023). The proposed method depends solely on the randomness of treatment assignment and is valid for arbitrary fixed designs and network structures. Under the sharp null hypothesis, $T(D, D^{obs})$ equals $T(D^{obs}, D^{obs})$, and $T(D^{obs}, D)$ equals $T(D, D)$, making the PNRT framework a generalization of FRTs for sharp null hypotheses.

Additionally, a multiple hypothesis testing adjustment ensures control of the family-wise error rate (FWER) when defining the "neighborhood" of interference based on distance or tie strength. If the spillover effect is positive, this approach can help policymakers design cost-effective interventions and optimize their implementation. Conversely, if the spillover effect is negative, identifying the range of these effects can aid in evaluating the policy's

overall effectiveness.

To illustrate the applicability of PNRTs, I apply them to a large-scale experiment by Blattman et al. (2021) in Bogotá, Colombia, which evaluates the impact of a hotspot policing policy on crime, using street segments as units of analysis. I assess the policy's overall effectiveness and explore criminal behavior and incentives by examining whether interference—such as crime displacement or deterrence—occurred in nearby neighborhoods following treatment assignment.[6] Additionally, Blattman et al. (2021) test the reach of spillover effects to identify units sufficiently distant from treated areas to serve as control groups. The authors report that increasing police patrolling time in hotspots has a significant displacement effect on property crime but not on violent crime. However, when tested against the alternative of no displacement effect, PNRTs suggest that contrary to Blattman et al. (2021), the displacement effect is marginally significant at the 10% level for violent crime, while the effect on property crime is insignificant. This finding could reshape our understanding of criminal behavior and inform welfare analysis, particularly if violent crime, being more severe, warrants stricter control measures.

A simulation study calibrated to this dataset further demonstrates the strong empirical properties of PNRTs compared to the biclique CRT and the classical FRT, confirming their suitability for general interference scenarios. In this context, we test for the displacement effect, where interference may cause outcomes to shift or "spill over" to neighboring units. In terms of size control, the pairwise comparison-based PNRT successfully maintains type I error rates even at the rejection level of α, indicating robustness under worst-case scenarios. By contrast, the classical FRT may over-reject under partial null hypotheses. Regarding power, the pairwise comparison-based PNRT at the α rejection level outperforms the biclique CRT, which faces power limitations due to complex conditioning events. This advantage is particularly valuable in network analysis, where data collection for each unit is costly and interference effects are often subtle (Taylor and Eckles, 2018; Breza et al., 2020). However, a trade-off exists, as PNRTs may exhibit conservatism under the null.

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 this method, there are at least

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

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 adopts the Fisherian perspective, focusing on detecting causal effects using finite-sample valid, randomization-based tests (Dufour and Khalaf, 2003; Lehmann and Romano, 2005; Rosenbaum, 2020). Pioneering work by Aronow (2012) and Athey et al. (2018) developed CRTs to test interference, with Basse et al. (2019) and Basse et al. (2024) demonstrating that CRTs can be adapted as permutation tests in certain network settings and interference patterns, thereby improving power and enabling computational efficiency. This paper builds on this foundation by introducing an alternative approach that applies broadly, even in cases where designing a conditioning event is challenging. Confidence intervals for specific causal parameters can then be constructed by inverting these tests. As noted by Basse et al. (2024), this approach ensures finite-sample validity with fewer model assumptions than model-based methods.[8]

Second, this paper contributes to the literature on extending randomization testing to non-sharp null hypotheses. While the primary focus is on the partial null hypothesis defined through distance measures, the principles of PNRTs appear to be generalizable beyond network settings. Since Neyman et al. (2018) acknowledged the limitation of FRTs in testing only sharp null hypotheses, researchers have developed various strategies to address weak nulls. For example, Ding et al. (2016), Li et al. (2016), and Zhao and Ding (2020) explore the null hypothesis of no average treatment effect, and Caughey et al. (2023) validate randomization tests for certain classes of test statistics under bounded nulls. Zhang and Zhao (2021) construct CRTs for partial sharp nulls, following a similar approach to Athey et al. (2018) and Puelz et al. (2021), applying this method in time-staggered adoption designs. To my knowledge, the PNRT method is the first procedure to address partial null hypotheses using unconditional testing.

The rest of the paper is structured as follows. Section 2 introduces the general setup and establishes all necessary notation. Section 3 presents the PNRT procedure, which includes

[7]Also see Basse and Airoldi (2018), Viviano (2022), Wang et al. (2023), Vazquez-Bare (2023), Leung (2020), Leung (2022), Shirani and Bayati (2024)

[8]Furthermore, randomization-based methods can be integrated with model-based frameworks, such as the linear-in-means model (Manski, 1993), to increase power or broaden applicability beyond randomized experiments while preserving test validity (Wu and Ding, 2021; Basse et al., 2024; Borusyak and Hull, 2023).

the pairwise imputable statistics and the p-value based on pairwise comparisons. Section 4 proposes a framework for determining the boundary of interference and adjusting for sequential testing. 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 appendix provides additional empirical and theoretical results as well as proofs.

2 Setup and Null Hypothesis of Interest

Consider N units indexed by $i \in \{1, 2, \ldots, N\}$, connected through an undirected network observed by the researcher. The researcher is interested in understanding the extent of interference based on factors such as distance, neighboring units, and connection strength, which are captured by an $N \times N$ proximity matrix G. The (i, j)-th component $G_{i,j} \geq 0$ represents a "distance measure" between units i and j, which can be either a continuous or discrete variable. For normalization, I set $G_{i,i} = 0$ for all $i = 1, 2, \ldots, N$, and assume $G_{i,j} > 0$ for all $i \neq j$. This distance measure is context specific:

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

Example 2 (Network distance)**.** *In social network settings, such as friendships on Facebook (Bond et al., 2012), $G_{i,j}$ measures the distance between units i and j, where $G_{i,j} = 1$ for friends, $G_{i,j} = 2$ for friends of friends, and $G_{i,j} = \infty$ if i and j are not connected. This framework accommodates disconnected networks and captures partial interference, such as cluster-level interference (Sobel, 2006; Basse et al., 2019).*

Example 3 (Link intensity)**.** *Researchers may observe not only whether two units are linked but also the intensity of the link $int_{i,j}$, such as frequency of interaction or volume of email correspondence (Goldenberg et al., 2009; Bond et al., 2012; Rajkumar et al., 2022). Building on the classic study by Granovetter (1973), one might examine how interference differs across weak and strong ties, defined by this intensity measure. Let $\bar{int} = \max_{i,j \in \{1,\ldots,N\}} int_{i,j}$, and define $G_{i,j} = \bar{int} - int_{i,j}$. In this way, an increase in $G_{i,j}$ implies a weaker connection, analogous to Examples 1 and 2.*

Example 4 (Distance in the product space)**.** *In the context of firms selling differentiated products, the units are the products, and the distance measure $G_{i,j}$ can be defined as the*

Euclidean distance between units i and j in a potentially multi-dimensional space of product characteristics, similar to Pollmann (2023). This distance measure can be useful for defining market boundaries, for instance, when a merger authority must determine whether two products belong to the same relevant market. This determination may depend on whether a price change for one product affects the quantity demanded for the other, or whether a price change impacts quantities for products located at greater distances within the product space.

In this paper, I focus on experimental settings where treatment assignment is random and follows a known probability distribution $P(D)$, where $P(d) = \Pr(D = d)$ is the probability that the treatment assignment D equals d. Let X represent the collected pre-treatment characteristics, such as age and gender, which can be used to control for unit heterogeneity. However, I do not attempt to evaluate their direct effects on the outcome. The probability distribution may or may not depend on covariates X. In cases of complete or cluster randomization, it does not depend on X, while in stratified or matched-pair designs, it does.

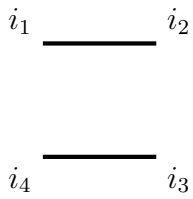
I adopt the potential outcomes framework with a binary treatment assignment vector $D = (D_1, \ldots, D_N) \sim P(D)$, where $D \in \{0,1\}^N$ and $D_i \in \{0,1\}$ denotes unit i's treatment. Let $Y(d) = (Y_1(d), \ldots, Y_N(d)) \in \mathbb{R}^N$ be the potential outcomes under treatment assignment d, where the potential outcome of unit i is $Y_i(d) = Y_i(d_1, \ldots, d_N)$. This allows unit i's potential outcome to depend on the treatment assignment of unit j, violating the classic *Stable Unit Treatment Value Assumption (SUTVA)* proposed by Cox (1958), and accommodating cases where spatial or network interference exists. However, the distance measure between treatment and individuals is unaffected by the treatment.

Throughout the paper, I assume that the following are observed: 1) the realized vector of treatments for all units, denoted by D^{obs}; 2) the realized outcomes for all units, denoted by $Y^{obs} \equiv Y(D^{obs}) = (Y_1(D^{obs}), \ldots, Y_N(D^{obs}))$; 3) the proximity matrix G; 4) the covariates X; and 5) the probability distribution of the treatment assignment P. I adopt a design-based inference approach, where D is treated as random, while G, X, P, and the unknown potential outcome schedule $Y(\cdot)$ are considered fixed. For simplicity in notation, these elements will not be treated as arguments of functions in the rest of the paper. To illustrate these notations, consider the following running example.

Running Example. Consider four street segments, where two segments are adjacent if they are connected, as shown in Figure 1. Units i_1 and i_2 are connected, forming one area, while units i_3 and i_4 are connected, forming another. For simplicity, the distance between units in the same area is set to 1. In practice, the distance between units in different areas

could be up to infinity, but for the sake of this example, assume it is 2.

Figure 1: Example Network Structure and Distance Matrix



(a) Network Structure

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

(b) Distance Matrix

Notes: Panel (a) presents the network structure of four units, while panel (b) displays the corresponding distance matrix.

Suppose the outcome of interest, Y, is the total number of crimes over a year, and a random treatment D is applied to increase policing in one unit. Assume the treatment is randomly assigned with $P(d) = 1/4$ for each possible assignment. Let the observed treatment be $D^{obs} = (1, 0, 0, 0)$ and the observed outcomes $Y^{obs} = (2, 4, 3, 1)$.

Table 1 illustrates the potential outcome schedule under the design-based framework, where the first row corresponds to the observed dataset. Although all potential outcomes are fixed values, only the outcomes under the observed treatment are known. In general, since potential outcomes can depend on assignments across all units, there could theoretically be up to 2^N potential outcomes.

Table 1: Potential Outcome Schedule in the Example

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

Notes: The table shows the potential outcome schedule under the design-based view. The first row represents the observed assignment D^{obs}, while potential outcomes denoted by ? are unobserved values.

2.1 Partial Null Hypothesis

I begin by formally defining the partial null hypothesis.

Definition 1 (Partial null hypothesis)**.** *A partial null hypothesis holds if there exists a collection of subsets* $\{\mathcal{D}_i\}_{i=1}^N$*, where each* $\mathcal{D}_i \subsetneq \{0,1\}^N$*, such that*

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

In this definition, the set $\mathcal{D}_i$ can vary with each unit i and is always a strict subset of $\{0,1\}^N$. As discussed in Zhang and Zhao (2023), the partial null hypothesis implies that the missing potential outcomes are only partially known. Compared to the traditional sharp null hypothesis (see Appendix A.1), which requires the null to hold for all potential treatment assignments, the partial null in Definition 1 applies to only a subset of potential assignments. The flexibility of the set $\mathcal{D}_i$, which is self-chosen, makes this hypothesis useful in settings where the strict sharp null hypothesis may be overly restrictive, allowing potential outcomes to differ for certain assignment vectors.

Although the method introduced in this paper can apply to any partial null hypothesis, we will specifically focus on cases where $\mathcal{D}_i$ is defined based on a distance measure in a network. This leads to the following definition.

Definition 2 (Distance interval assignment set)**.** *For a unit* $i \in \{1,\ldots,N\}$ *and a given distance* ϵ_s*, the distance interval assignment set is defined as*

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

When $d \in \mathcal{D}_i(\epsilon_s)$*, unit* i *is said to be in the distance interval* (ϵ_s, ∞)*.*

This definition involves two key concepts: $\mathcal{D}_i(\epsilon_s)$ and the interval (ϵ_s, ∞), both of which are specific to each unit i. The distance interval assignment set $\mathcal{D}_i(\epsilon_s)$ maps a distance ϵ_s to a set of treatment assignments where unit i is at least a distance ϵ_s away from any treated units. For any $\epsilon_s \geq 0$, since $G_{i,i} = 0$, it follows that $1\{G_{i,i} \leq \epsilon_s\} = 1$, implying that unit i is untreated ($d_i = 0$) for any assignment $d \in \mathcal{D}_i(\epsilon_s)$. Specifically, when $\epsilon_s = 0$, all $G_{i,j}$ for $i \neq j$ are positive, which ensures that $1\{G_{i,j} \leq \epsilon_s\} = 0$. As a result, there is no restriction on the treatment status of other units d_j, and $\mathcal{D}_i(0)$ includes all treatment assignments d where $d_i = 0$, while allowing others to be treated.[9]

The distance interval assignment set $\mathcal{D}_i(a)/\mathcal{D}_i(b)$ corresponds to treatment assignments where unit i is within the distance interval $(a, b]$. For any treatment assignment d, the set

[9]For any $\epsilon_s < 0$, since $G_{i,j} \geq 0$ for all i, j, we have $1\{G_{i,j} \leq \epsilon_s\} = 0$, meaning that $\mathcal{D}_i(\epsilon_s) = \{0,1\}^N$, where all treatment assignments are included.

$\{i : d \in \mathcal{D}_i(a)/\mathcal{D}_i(b)\}$ contains all units that fall within the distance interval $(a, b]$ relative to treated units.

Using the concept of distance interval assignment sets, we now define the partial null hypothesis of interference based on distance.

Definition 3 (Partial null hypothesis of interference on distance $\epsilon_s \geq 0$)**.** *The partial null hypothesis of interference on distance $\epsilon_s \geq 0$ is defined as*

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

This hypothesis asserts that no interference occurs beyond distance ϵ_s, meaning the potential outcomes for unit i remain unchanged for any treatment assignment where unit i is at least a distance ϵ_s away from all treated units. Under this null hypothesis, the potential outcomes for unit i can be imputed for treatment assignment vectors that satisfy this distance condition, allowing for a partial imputation of outcomes. The interpretation of distance here is context-specific and depends on the nature of the interference in the particular application.

Example 1 (Spatial distance continued)**.** *In a setting where units represent street segments, for a given spatial distance ϵ_s (e.g., 500 meters), $\mathcal{D}_i(\epsilon_s)$ consists of all treatment assignments where unit i is at least 500 meters away from any treated street segments. The partial null hypothesis $H_0^{\epsilon_s}$ tests whether spillover effects occur on an untreated unit located 500 meters away from any treated units.*

Example 2 (Network distance continued)**.** *Consider two schools, each with 100 students, where the goal is to test for cluster interference within schools. We assume that students within the same school have a distance of 100 from each other and are infinitely distant from students in the other school. Setting $\epsilon_s = 0$, we test for interference within schools. Cluster interference is present if students' outcomes are affected by treatment assignments in their own school but not in the other school.*[10]

Example 3 (Link intensity continued)**.** *Consider a scenario where units represent individuals with cell phones, and the intensity of their connection is measured by the number of text messages exchanged, with a maximum of 50 messages per week. We define the "distance" between two individuals as 50 minus the number of messages exchanged. For $\epsilon_s = 40$, $\mathcal{D}_i(\epsilon_s)$ represents all treatment assignments where unit i has exchanged fewer than 10 messages with*

[10]Setting $\epsilon_s = 101$ would test for interference across schools, but such a test may lack power in practice, as noted by Puelz et al. (2021).

any treated units. The partial null hypothesis $H_0^{\epsilon_s}$ tests whether interference occurs for an untreated unit that has exchanged fewer than 10 messages with treated units.

The null hypothesis defined in Definition 3 is useful for assessing the existence or extent of interference within a network, as researchers are often interested in whether interference occurs beyond a certain distance ϵ_s. If $\epsilon_s > 0$, researchers can use this approach to identify the neighborhood of interference or to find a suitable comparison group for subsequent estimation.

Comparison to the Traditional T-Test. A common strategy for testing interference at a given distance ϵ_s is to compare units at varying distances from treated units, often referred to as the inner versus outer ring method when the number of treated units is limited (Pollmann, 2023). This method operates on the premise that outcome values for outer ring units remain unaffected by treatment interference and can approximate the potential control outcomes of inner ring units. For example, Blattman et al. (2021) employ a similar approach to estimate the range of spillover effects, calculating the average outcome across different units in both the inner and outer rings and testing for systematic differences between these groups.[11]

However, as Pollmann (2023) notes, this approach requires assumptions that extend beyond the randomization in the experiment. First, as discussed by Aronow (2012), even in a randomized experiment, each unit's distance from treated units is not random. Specifically, because treated units may not be uniformly distributed across space, units in some regions are more likely to be closer to treated units than those in others. Consequently, outer ring units may differ systematically from inner ring units across different treatment assignments. Second, as Pollmann (2023) highlights, even if each unit has an equal chance of being in the inner or outer rings, functional form assumptions about the potential outcomes are required to eliminate bias in these comparisons. In summary, without further assumptions, outer ring units may not serve as a valid control group for inner ring units, potentially leading to biased results even in randomized experiments.

The advantage of the partial null hypothesis in Definition 3 is that it tests interference by directly evaluating *the same unit's* potential outcome whenever it is at least distance ϵ_s from treated units. This unit-level approach circumvents the issues associated with averaging outcomes across different units, which may not be comparable, even under randomization.

[11] Blattman et al. (2021) use an F-test to compare mean differences in outcomes such as "perceived risk" and "crime incidence." For details, see Blattman et al. (2021), Section A.2 in the Online Appendix.

The key contribution of this paper is to show that interference can be tested with only the assumption of random treatment assignment.

If $\epsilon_s = 0$, we test the partial null hypothesis of no interference as $\mathcal{D}_i(0)$ consists of all treatment assignments where $d_i = 0$, meaning the unit is untreated. Treatment assignments where $d_i = 1$ are excluded, ensuring that the hypothesis solely focuses on spillover effects. In contrast, the traditional sharp null hypothesis includes potential outcomes where $d_i = 1$, which also involves direct treatment effects. We can simplify H_0^0 further, as illustrated in the following running example.

Running Example Continued. Suppose researchers want to test for the existence of spillover effects using the partial null hypothesis in Definition 3 with $\epsilon_s = 0$:

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

Throughout the paper, I use the above H_0^0 for illustration in the running example. This hypothesis implies that the potential outcome for any untreated unit i remains unchanged regardless of the treatment assignments of other units. The potential outcome schedule under H_0^0 is shown in Table 2.

As shown in Table 2, the null hypothesis H_0^0 allows us to impute many of the previously missing potential outcomes. For example, since we observe the outcome when unit i_2 is not treated, we can impute other outcomes as long as unit i_2 remains untreated. Consequently, the outcome for i_2 when either unit i_3 or i_4 is treated is also 4.

Table 2: Potential Outcome Schedule Under Partial Null H_0^0

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

Notes: The tables shows the potential outcome schedule with the partial null hypothesis under Definition 3 for the toy example. Assignment D includes all potential assignments, with the first row representing the observed assignment D^{obs}. Potential outcomes marked in ? are non-imputable values under the partial null.

2.2 Two Technical Challenges for Randomization Tests

Even under the partial null hypothesis, where fewer potential outcomes are missing, technical challenges remain. Table 2 illustrates the potential outcome schedule under the partial null hypothesis H_0^0, highlighting two specific challenges that persist in more general settings.

First, only a subset of potential outcomes can be observed or imputed. For example, under H_0^0, if unit i_2 is treated, the hypothesis provides no information about the potential outcomes of unit i_1. This leaves the potential outcomes for both i_1 and i_2 missing. Second, the set of units with imputable outcomes varies depending on the treatment assignment. For instance, if unit i_3 is treated instead, the missing values now belong to i_1 and i_3, which differs from other assignments. These challenges make traditional test statistics inapplicable because randomization requires knowledge of all $Y_i(d)$ values for the relevant assignment. Therefore, constructing a valid test statistic that accounts for the missing potential outcomes is the first key technical challenge.

Traditional Test Statistics. In practice, researchers often specify a distance ϵ_c such that units farther than ϵ_c from treated units are assumed to experience no interference. For instance, in a spatial setting, we might assume that no interference occurs for units more than $\epsilon_c = 1,000$ meters away. For cluster interference, we might assume that no spillover occurs once ϵ_c exceeds the maximum distance within a cluster, indicating no interference across clusters.

A natural test statistic compares units within the distance interval $(\epsilon_s, \epsilon_c]$ to the treated group, while using units in the distance interval (ϵ_c, ∞) as a pure control group. The idea behind ϵ_c is to identify a threshold beyond which the influence of the treatment is negligible, allowing researchers to separate units likely to be impacted by interference from those that serve as clean controls. If the researcher has no prior value for ϵ_c, Section 4.2 proposes a sequential testing procedure to help select an appropriate ϵ_c. Even if ϵ_c is misspecified and does not provide a clean control group, the proposed testing procedure remains valid, though it may reduce test power.

For example, consider the difference in means with control distance ϵ_c:

$$T(Y(D^{obs}), D) = \underbrace{\bar{Y}(D^{obs})_{\{i:D\in\mathcal{D}_i(\epsilon_s)/\mathcal{D}_i(\epsilon_c)\}}}_{\text{Mean of neighbor}} - \underbrace{\bar{Y}(D^{obs})_{\{i:D\in\mathcal{D}_i(\epsilon_c)\}}}_{\text{Mean of control}},$$

where

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

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

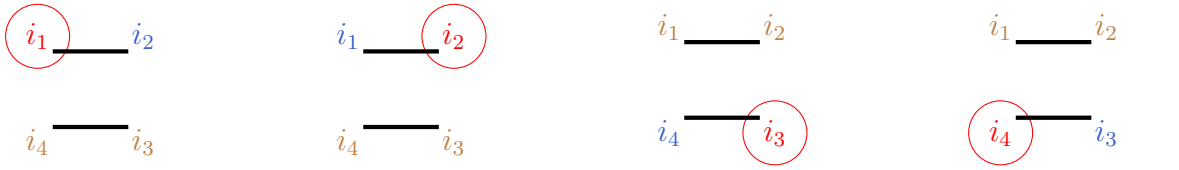
$$\bar{Y}(D^{obs})_{\{i:D\in\mathcal{D}_i(\epsilon_c)\}} = \frac{\sum_{i=1}^{N} 1\{D \in \mathcal{D}_i(\epsilon_c)\}Y_i(D^{obs})}{\sum_{i=1}^{N} 1\{D \in \mathcal{D}_i(\epsilon_c)\}},$$

which represents the mean value of units in the distance interval (ϵ_c, ∞). The difference-in-means estimator is widely used in the literature, including Basse et al. (2019) and Puelz et al. (2021). Given treatment assignment d, $\{i : d \in \mathcal{D}_i(a)/\mathcal{D}_i(b)\}$ denotes the set of units in the distance interval $(a, b]$.

Let us now consider an illustration using the running example.

Running Example Continued. For the rest of the discussion in the running example, I would use $\epsilon_c = 1$. Therefore, there are two relevant distance intervals for the difference-in-means estimator: $(0, 1]$ and $(1, \infty)$. Figure 2 shows how these intervals change with different treatment assignments.

Figure 2: Example Network Structure with Treated, Neighbor, and Control Units



(a) Treated Unit: i_1 (b) Treated Unit: i_2 (c) Treated Unit: i_3 (d) Treated Unit: i_4

Notes: Units with red circles are treated, units in blue are neighbors in the interval $(0, 1]$, and units in brown are control units in the interval $(1, \infty)$.

Applying traditional test statistics, like the difference-in-means estimator, can be problematic when some potential outcomes remain unknown under H_0^0. Table 3 shows that test statistics under non-observed treatment assignments still involve missing values, making randomization tests inapplicable under the partial null hypothesis.

Table 3: Traditional Test Statistics Under Partial Null H_0^0

Assignment D	Potential Outcome Y_i				$T(Y(D^{obs}), D)$
	i_1	i_2	i_3	i_4	
$(1,0,0,0)$	2	4	3	1	2
$(0,1,0,0)$	?	?	3	1	?
$(0,0,1,0)$	?	4	?	1	?
$(0,0,0,1)$	?	4	3	?	?

Notes: The table shows the potential outcome schedule under the partial null hypothesis for the example. Assignment D includes all potential assignments, with the first row representing the observed assignment D^{obs}. Potential outcomes marked in red question marks are non-imputable under the partial null.

3 Two Types of PNRTs

For notational simplicity, I fix ϵ_s and ϵ_c for the rest of this section. For each treatment assignment d, I focus on the units that are imputable under $H_0^{\epsilon_s}$ given the observed information.

Definition 4 (Imputable units set)**.** *Given a treatment assignment $d \in \{0,1\}^N$ and a partial null hypothesis $H_0^{\epsilon_s}$, the set of units*

$$\mathbb{I}(d) \equiv \{i \in \{1,\ldots,N\} : d \in \mathcal{D}_i(\epsilon_s)\} \subseteq \{1,\ldots,N\}$$

is called the set of imputable units under treatment assignment d.

The set of imputable units is a subset of units for which imputation is possible and corresponds to units in the distance interval (ϵ_s, ∞) for the partial null hypothesis $H_0^{\epsilon_s}$. Given the observed treatment D^{obs}, the set $\mathbb{I}(D^{obs})$ includes all the units we can use for testing. Units outside this set do not provide additional information as their potential outcomes are not imputable under the partial null. For instance, if $\epsilon_s = 0$, the set $\mathcal{D}_i(\epsilon_s)$ includes all assignments d where $d_i = 0$, meaning $\mathbb{I}(D^{obs})$ consists of all units not treated under D^{obs}.

Running Example Continued. Under H_0^0, the imputable units for a treatment assignment d can be expressed as $\mathbb{I}(d) \equiv \{i \in \{1,\ldots,N\} : d_i = 0\}$. That is, under the null hypothesis of no interference, all non-treated units are imputable. For example, as shown in Figure 3, when unit i_1 is treated, units i_2 to i_4 belong to the imputable set, and when unit i_2 is treated, units i_1, i_3, and i_4 are imputable. It is worth noting that this is a special case where all untreated units are imputable. In more general settings, this depends on the ϵ_s in the null hypothesis.

Figure 3: Example Network Structure with Imputable Units

(a) Treated Unit: i_1 (b) Treated Unit: i_2 (c) Treated Unit: i_3 (d) Treated Unit: i_4

Notes: Treated units are marked with red rectangles, while imputable units are shown in black.

As shown in Figure 3, generally, $\mathbb{I}(d) \neq \mathbb{I}(d')$ for different assignments d and d'. For example, when testing for spillover effects among friends, the set of friends affected will change with different treatment assignments due to varying social connections. In practice, $\mathbb{I}(D^{obs})$ could sometimes be empty, depending on the network structure and the specific partial null hypothesis. If no units meet the required criteria (i.e., $\mathbb{I}(D^{obs})$ is empty), one approach is to reject the null hypothesis α percent of the time, in line with the desired significance level. This ensures control of the test's size, even in cases where the imputable set is empty. However, to achieve power in such cases, additional data or a different study design may be necessary. See Appendix E for further discussion.

The set of imputable units can also be defined under the sharp null hypothesis, though in this case, $\mathbb{I}(d) = \{1, \ldots, N\}$ for any assignment d, meaning all units are imputable under the sharp null. Therefore, there has been less focus on the imputable units set in the randomization tests literature. To help define the test statistics later, I further define the following.

Definition 5 (Imputable outcome vector)**.** *For any treatment assignment $d \in \{0,1\}^N$ and a partial null hypothesis $H_0^{\epsilon_s}$, the vector*

$$Y_{\mathbb{I}(d)} \equiv \{Y_i\}_{i \in \mathbb{I}(d)}$$

is called the imputable outcome vector for the treatment assignment d, with each component representing the potential outcome for the units in $\mathbb{I}(d)$. When the value of Y is determined by an alternative treatment assignment d', we denote

$$Y_{\mathbb{I}(d)}(d') \equiv \{Y_i(d')\}_{i \in \mathbb{I}(d)}$$

as the imputable outcome vector for d, with each component representing the potential outcome under the alternative treatment assignment d' for the units in $\mathbb{I}(d)$.

Two factors influence the imputable outcome vector. First, the value of the potential outcome depends on d'. For example, when $d' = D^{obs}$, $Y(d') = Y^{obs}$ is the observed outcome in the dataset. Second, the set of units included in the vector is determined by assignment d. For the potential outcome vector $Y(d')$ under treatment assignment d', $Y_{\mathbb{I}(d)}(d')$ is a subvector of it, and different assignments d lead to different sets of units in the imputable outcome vector when testing the partial null hypothesis. In contrast, under the sharp null hypothesis, $\mathbb{I}(d) = \{1, \ldots, N\}$, so $Y_{\mathbb{I}(d)}(d') = Y(d')$.

3.1 Pairwise Imputable Statistics

The definitions above allow me to further define the core idea behind the test statistics.

Definition 6 (Pairwise imputable statistic)**.** *Let $T : \mathbb{R}^N \times \{0,1\}^N \times \{0,1\}^N \to \mathbb{R} \cup \{\infty\}$ be a measurable function, and let $Y_{\mathbb{I}(d)}$ be an imputable outcome vector. The function T is called a pairwise imputable statistic if*

$$T(Y_{\mathbb{I}(d)}, d') = T(Y'_{\mathbb{I}(d)}, d')$$

for any $d, d' \in \{0,1\}^N$ such that $Y_i = Y'_i$ for all $i \in \mathbb{I}(d) \cap \mathbb{I}(d')$.

This formal definition generalizes the concept introduced earlier in the introduction, where the pairwise imputable statistic was framed as a bivariate function $T(D_1, D_2)$. While the earlier framing emphasized the key role of imputability, this definition provides a more precise mathematical representation, capturing the interaction of imputable outcome vectors across different treatment assignments.

The set $\mathbb{I}(d) \cap \mathbb{I}(d')$ in Definition 6 is similar to the set H in Definition 1 of Zhang and Zhao (2023). Intuitively, it excludes units that are not imputable under the partial null hypothesis in the test statistics. At first glance, the pairwise imputable statistic may seem to restrict the form of the test statistics, but it is actually general enough to accommodate commonly used test statistics. For instance, the classic difference in means can be written as

$$T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), D) = \underbrace{\bar{Y}_{\mathbb{I}(D^{obs})}(D^{obs})_{\{i:D\in\mathcal{D}_i(\epsilon_s)/\mathcal{D}_i(\epsilon_c)\}}}_{\text{Mean of } \textit{imputable} \text{ neighbor}} - \underbrace{\bar{Y}_{\mathbb{I}(D^{obs})}(D^{obs})_{\{i:D\in\mathcal{D}_i(\epsilon_c)\}}}_{\text{Mean of } \textit{imputable} \text{ control}}.$$

Here,

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

which is the mean value of imputable 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 imputable units in the distance interval (ϵ_c, ∞).

This formula matches the classic difference in means when $\mathbb{I}(D^{obs}) = \{1, \ldots, N\}$, and whether unit i belongs to distance interval $(\epsilon_s, \epsilon_c]$ or (ϵ_c, ∞) depends on D. In practice, one of these mean values might be undefined if no unit in $\mathbb{I}(D^{obs})$ belongs to one of these two intervals. In that case, I define $T = \infty$ or $T = \max(Y^{obs})$ to ensure that the test remains conservative and valid. See Appendix E for further discussion on when it would be the case.[12]

Running Example Continued. Consider the test statistic

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

Table 4 presents the corresponding values for the first and second terms of the test statistic, while Figure 4 provides a visual representation of how we determine the imputable neighbor units and imputable control units.

Table 4: Constructing a Pairwise Imputable Statistic

Assignment D	Potential Outcome Y_i				$\bar{Y}_{\mathbb{I}(D^{obs})}(D^{obs})$		$T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), D)$
	i_1	i_2	i_3	i_4	$\{i : D \in \mathcal{D}_i(0)/\mathcal{D}_i(1)\}$	$\{i : D \in \mathcal{D}_i(1)\}$	
$(1,0,0,0)$	2	4	3	1	4	2	2
$(0,1,0,0)$		?	3	1	?	2	4
$(0,0,1,0)$		4	?	1	1	4	-3
$(0,0,0,1)$		4	3	?	3	4	-1

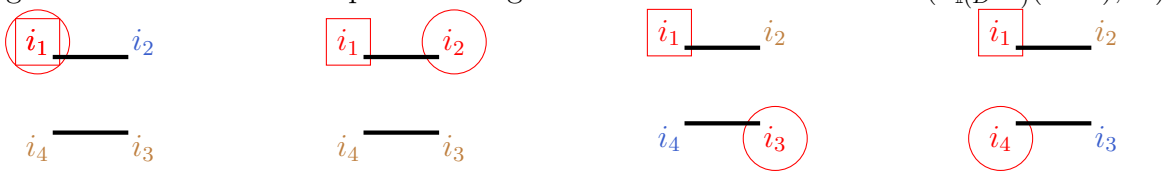
Notes: Assignment D includes all potential assignments, with the first row corresponding to the observed assignment D^{obs}. Potential Outcome Y_i is the potential outcome of each unit under the null H_0^0, with red question marks representing missing values. Unit i_1 does not belong to set $\mathbb{I}(D^{obs})$, so the entire column is marked in red. $\bar{Y}_{\mathbb{I}(D^{obs})}(D^{obs})$ with $\{i : D \in \mathcal{D}_i(0)/\mathcal{D}_i(1)\}$ is the mean potential outcome for units in the distance interval $(0, 1]$, marked in blue. $\bar{Y}_{\mathbb{I}(D^{obs})}(D^{obs})$ with $\{i : D \in \mathcal{D}_i(1)\}$ is the mean potential outcome for units in the distance interval $(1, \infty)$. $T = \max(Y^{obs})$, marked in red when one of the mean values is undefined.

As illustrated in Figure 4, D^{obs} refers to the scenario where unit i_1 is treated, so the set of imputable units remains the same across different potential assignments D. However, the po-

[12]To increase the test's power, it might be helpful to combine it with conditional randomization testing, which truncates the set of treatment assignments to avoid undefined cases (Zhang and Zhao, 2023).

tential assignment D can change, altering which units belong to the neighborhood set and the control set. When $D = D^{obs}$ and unit i_1 is treated, the first term $\bar{Y}_{\mathbb{I}(D^{obs})}(D^{obs})_{\{i:D\in\mathcal{D}_i(0)/\mathcal{D}_i(1)\}}$ corresponds to the outcome of i_2, while the second term $\bar{Y}_{\mathbb{I}(D^{obs})}(D^{obs})_{\{i:D\in\mathcal{D}_i(1)\}}$ is the mean outcome of i_3 and i_4. When unit i_2 is treated, there are no imputable units in the neighborhood set, so I define $T = \max(Y^{obs}) = 4$ to ensure the test's validity.

Figure 4: Illustration of Imputable Neighbor and Control Units for $T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), D)$



(a) D: i_1; D^{obs}: i_1 (b) D: i_2; D^{obs}: i_1 (c) D: i_3; D^{obs}: i_1 (d) D: i_4; D^{obs}: i_1

Notes: Red circles indicate treated units in D, which determine neighbor units in the interval $(0, 1]$ and control units in $(1, \infty)$. Red rectangles indicate treated units in D^{obs}, which determine imputable units.

Additionally, I can incorporate rank statistics by excluding non-imputable units and reranking the remaining units. Following Imbens and Rubin (2015), I define the rank as

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

Thus, the test statistic becomes

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

When $Y_i(D^{obs}) = Y_i(D)$ for all $i \in \mathbb{I}(D^{obs})\cap\mathbb{I}(D)$, $R_i(Y_{\mathbb{I}(D^{obs})\cap\mathbb{I}(D)}(D^{obs})) = R_i(Y_{\mathbb{I}(D^{obs})\cap\mathbb{I}(D)}(D))$, meaning the ranks remain unchanged. Therefore, $T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), D) = T(Y_{\mathbb{I}(D^{obs})}(D), D)$, satisfying Definition 6. Both test statistics discussed above can also be adapted to use the absolute value version for two-sided testing.

For further details on the choice of statistics in randomization testing, refer to Section 5 of Imbens and Rubin (2015). For other choices of T in network settings, see Section 5 of Athey et al. (2018). Additionally, one could use the regression coefficient of interest as illustrated in Hoshino and Yanagi (2023). While the method is valid without covariate adjustments, incorporating them may increase the test's power in practice (Wu and Ding,

2021) (see Appendix D for further discussion). Under the sharp null hypothesis, where all units are imputable regardless of the treatment assignment d, $\mathbb{I}(d) \cap \mathbb{I}(d') = \{1, \ldots, N\}$ for any d and d', and all formulas reduce to the classic form as defined in Imbens and Rubin (2015).

Following Definition 6 of pairwise imputable statistics, I can derive a property to calculate test statistics using only the observed information:

Proposition 1. *Suppose the partial null hypothesis $H_0^{\epsilon_s}$ is true. Suppose $T(Y_{\mathbb{I}(d)}(d), d')$ is a pairwise imputable statistic. Then,*

$$T(Y_{\mathbb{I}(d)}(d), d') = T(Y_{\mathbb{I}(d)}(d'), d')$$

for any $d, d' \in \{0,1\}^N$.

The proof is provided in Appendix C.

Let $d = D^{obs}$ and $d' = D$. By Proposition 1, $T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), D) = T(Y_{\mathbb{I}(D^{obs})}(D), D)$ under the null $H_0^{\epsilon_s}$, ensuring I observe a counterfactual test statistic for comparison.

3.2 Unconditional Randomization Test

To differentiate from the existing literature on CRTs, I introduce the following definition for the unconditional randomization test.

Definition 7 (Unconditional randomization test)**.** *An unconditional randomization test $\phi : \{0,1\}^N \to [0,1]$ is defined such that for any $D^{obs} \in \{0,1\}^N$,*

$$\phi(D^{obs}) = Q(\tilde{p}(D^{obs}), \alpha),$$

where $Q : [0,1] \times [0,1] \to [0,1]$ is a measurable function, α is the nominal level, and $\tilde{p}(D^{obs})$ can be written as

$$\tilde{p}(D^{obs}) = \sum_{d \in \{0,1\}^N} g(D^{obs}, d) P(D = d),$$

with P being the pre-specified probability distribution on the treatment assignment, and $g : \{0,1\}^N \times \{0,1\}^N \to \{0,1\}$ a measurable function.

The key feature of the unconditional randomization test is that the probability of rejection, $\phi(D^{obs})$, is computed by randomizing the treatment assignment according to the same probability distribution P that governs the original treatment assignment. This contrasts

with methods in the existing literature, such as Athey et al. (2018), where the rejection function is based on randomizing the treatment assignment within a conditional probability space, conditioned on certain events. A more detailed discussion is provided in Appendix A.2.

One example is the simple randomization test, which uses pairwise imputable statistics, with p-values constructed similarly to the classic FRT.

Definition 8 (Simple randomization test)**.** *A simple randomization test is an unconditional randomization test defined by* $\phi(D^{obs}) = 1\{pval(D^{obs}) \leq \alpha\}$*, where* $pval(D^{obs}) : \{0,1\}^N \to [0,1]$ *is the p-value function given by*

$$pval(D^{obs}) = P(T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), D) \geq T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), D^{obs})) \text{ for } D \sim P(D),$$

and $T(Y_{\mathbb{I}(d)}(d), d')$ *denotes a pairwise imputable statistic.*

Running Example Continued. Using pairwise imputable statistics $T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), D)$ and following Table 4, we can construct Table 5 with the test statistics for each assignment.

Table 5: Simple Randomization Test in the Example

Assignment D	Potential Outcome Y_i				$T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), D)$
	i_1	i_2	i_3	i_4	
$(1,0,0,0)$	2	4	3	1	2
$(0,1,0,0)$		?	3	1	4
$(0,0,1,0)$		4	?	1	-3
$(0,0,0,1)$		4	3	?	-1

Notes: Assignment D includes all potential assignments, with the first row representing the observed assignment D^{obs}. Potential Outcome Y_i is the potential outcome of each unit under the null H_0^0, while red question marks denote missing values. Unit i_1 does not belong to the set $\mathbb{I}(D^{obs})$, so the column is marked in red. Blue cells represent the units used to calculate the mean value in the first term of the test statistics. $T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), D)$ are test statistics for different D, fixing D^{obs} where unit i_1 is treated.

Following Definition 8, the p-value is given by

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

with respect to $D \sim P(D)$, where D is drawn independently from D^{obs}. Based on Table 5, this results in a p-value of $2/4$. However, one might question whether this procedure guarantees finite-sample validity—specifically, whether it satisfies the condition $E_P(\phi(D^{obs})) \leq \alpha$ under the null hypothesis.

Investigating Finite-Sample Validity. Although pairwise imputable statistics are used, naively constructing the p-value as defined in the classic FRT does not guarantee the test's validity. For the test to be valid, the following condition must hold under the partial null hypothesis:

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

where $\stackrel{d}{=}$ indicates equality in distribution. The distribution on the left-hand side (LHS) is with respect to D, while the distribution on the right-hand side (RHS) is with respect to D^{obs}.

By Proposition 1, under the null hypothesis, we also have:

$$T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), D) \stackrel{H_0}{=} T(Y_{\mathbb{I}(D^{obs})}(D), D).$$

Here, $\stackrel{H_0}{=}$ denotes equality under the null hypothesis. However, the term $T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), D^{obs})$, being induced by the randomness of D^{obs}, satisfies:

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

Thus, for the test to maintain validity, we require:

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

This condition is not guaranteed under the partial null hypothesis because $\mathbb{I}(D^{obs}) \neq \mathbb{I}(D)$ in general. Different treatment assignments D result in different sets of imputable units, leading to variability in $\mathbb{I}(D)$. This is a key technical challenge. In the special case of testing the sharp null hypothesis, where $\mathbb{I}(D^{obs}) = \{1, \ldots, N\} = \mathbb{I}(D)$, the validity trivially holds.

To address the challenges posed by varying imputable unit sets, previous literature suggests a remedy through the design of a conditioning event that consists of a fixed subset of imputable units, known as *focal units*, and a fixed subset of assignments, known as *focal assignments*. CRTs are then performed by conducting FRTs within this conditioning event (see Appendix A.2 for a detailed discussion). However, using conditioning events in practice introduces two key 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 typically corresponds to a smaller subset of experimental units. This inevitably results in a loss of information,

with fewer units and assignments within the conditioning events, potentially affecting the test's power. Second, constructing the conditioning event adds a layer of computational complexity. This raises the question: can unconditional randomization testing still be valid in finite samples?

While previous approaches rely on 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 achieves valid testing through a carefully designed p-value calculation, ensuring finite-sample validity without the need for conditioning events.

3.3 The Pairwise Comparison-Based PNRT

Building on the selective inference literature (Wen et al., 2023; Guan, 2023), the key idea is to compute p-values 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, $T(Y_{\mathbb{I}(d^r)}(D^{obs}), D^{obs})$ remains relatively large 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, and rejection of the null is still possible when the units in the neighborhood set tend to have high outcome values. As a result, we would expect a small p-value, as the probability that $T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), d^r)$ exceeds $T(Y_{\mathbb{I}(d^r)}(D^{obs}), D^{obs})$ is low.

I refer to any randomization test with p-values constructed through this pairwise comparison method as a "PNRT." Formally, this procedure is called the "pairwise comparison-based PNRT," with the p-value defined below.

Definition 9 (Pairwise comparison-based PNRT)**.** *A pairwise comparison-based PNRT is an unconditional randomization test defined by* $\phi^{pair}(D^{obs}) = 1\{pval^{pair}(D^{obs}) \leq \alpha/2\}$*, where* $pval^{pair}(D^{obs}) : \{0,1\}^N \to [0,1]$ *is the p-value function given by*

$$pval^{pair}(D^{obs}) = P(T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), D) \geq T(Y_{\mathbb{I}(D)}(D^{obs}), D^{obs})) \text{ for } D \sim P(D),$$

and $T(Y_{\mathbb{I}(d)}(d), d')$ *denotes a pairwise imputable statistic.*

In practice, this p-value can be computed using Algorithm 1, where the p-value is calculated as the mean value of $1 + R$ draws, with $r = 0$ corresponding to $d = D^{obs}$.

Running Example Continued. Using the difference-in-mean estimator as before,

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

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 α.

for $r = 1$ *to* R **do**

Randomly sample $d^r \sim P(D)$, and 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 $\hat{pval}^{pair} \leq \alpha/2$.

Table 6: PNRT in the Example

Assignment D	Potential Outcome Y_i				$\bar{Y}_{\mathbb{I}(D)}(D^{obs})$		$T(Y_{\mathbb{I}(D)}(D^{obs}), D^{obs})$
	i_1	i_2	i_3	i_4	$\{i : D^{obs} \in \mathcal{D}_i(0)/\mathcal{D}_i(1)\}$	$\{i : D^{obs} \in \mathcal{D}_i(1)\}$	
$(1,0,0,0)$	2	4	3	1	4	2	2
$(0,1,0,0)$		?	3	1	?	2	4
$(0,0,1,0)$		4	?	1	4	1	3
$(0,0,0,1)$		4	3	?	4	3	1

Notes: Assignment D includes all potential assignments, with the first row representing the observed assignment D^{obs}. Potential Outcome Y_i is the potential outcome of each unit under the null H_0^0, with red question marks indicating missing values. Unit i_1 does not belong to either the neighborhood set or the control set under D^{obs}, so the column is marked red. Unit i_2 is in the distance interval $(0, 1]$ under D^{obs}, so the column is marked blue. Units i_3 and i_4 are in the distance interval $(1, \infty)$ under D^{obs}, so those columns are marked brown. $T(Y_{\mathbb{I}(D)}(D^{obs}), D^{obs})$ is calculated as the mean of non-missing potential outcomes in the blue columns minus the mean of non-missing potential outcomes in the brown columns.

As shown in Figure 5, for each treatment assignment D, the test statistic is calculated as the mean value of i_2 (excluding missing values) minus the mean value of i_3 and i_4 (excluding missing values).

Based on Tables 6 and 4, I can construct Table 7, where each row represents the values used to compare and construct the p-value for each (D^{obs}, D) pair.

Only when D involves treating units i_1 or i_2 does $T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), D) \geq T(Y_{\mathbb{I}(D)}(D^{obs}), D^{obs})$. Hence, $pval^{pair} = 2/4$. In practice, similar to Guan (2023), using $1/2$ to discount the number of equalities can reduce the p-value without compromising test validity. Additionally, in simulation experiments, using a uniform random number multiplied by the number of equalities also maintains test validity.

The validity of Algorithm 1 follows from the symmetry between $T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), d^r)$ and $T(Y_{\mathbb{I}(d^r)}(D^{obs}), D^{obs})$ under the null hypothesis $H_0^{\epsilon_s}$. Intuitively, for each pair of assignments D^{obs} and d^r, both terms are restricted to units $i \in \mathbb{I}(D^{obs}) \cap \mathbb{I}(d^r)$ by Definition 6. Moreover, by Proposition 1, under the null, with $d = D$ and $d' = D^{obs}$, we have $T(Y_{\mathbb{I}(D)}(D^{obs}), D^{obs}) =$

Figure 5: Illustration of Imputable Neighbor and Control Units for $T(Y_{\mathbb{I}(D)}(D^{obs}), D^{obs})$

(a) D^{obs}: i_1, D: i_1 (b) D^{obs}: i_1, D: i_2 (c) D^{obs}: i_1, D: i_3 (d) D^{obs}: i_1, D: i_4

Notes: Treated units in D^{obs} are marked with red circles and determine the neighbor units in the interval $(0, 1]$ and control units in $(1, \infty)$. Treated units in D are marked with red rectangles and determine the imputable units.

$T(Y_{\mathbb{I}(D)}(D), D^{obs})$, which is the counterfactual value of $T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), D)$ when flipping the observed assignment and randomized assignment between D and D^{obs}. Thus, the pairwise comparison is symmetric, and its validity follows from the conformal lemma in the conformal prediction literature (Guan, 2023).

Table 7: Pairwise Comparison for PNRTs

Assignment D	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		
$(1, 0, 0, 0)$	2	4	3	1	2	2
$(0, 1, 0, 0)$	?	?	3	1	4	4
$(0, 0, 1, 0)$	?	4	?	1	-1	1
$(0, 0, 0, 1)$	?	4	3	?	-3	3

Notes: Assignment D includes all potential assignments, with the first row representing the observed assignment D^{obs}. Potential Outcome Y_i is the potential outcome of each unit under the null H_0^0, with red question marks indicating missing values. $T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), D)$ are test statistics under different D while fixing D^{obs} for imputable units, with the same values as in Table 5. $T(Y_{\mathbb{I}(D)}(D^{obs}), D^{obs})$ are test statistics under different D for imputable units, with the same values as in Table 6.

Theorem 1. *Suppose the partial null hypothesis H_0^{es} holds. Then, the pairwise comparison-based PNRT, as defined in Definition 9, satisfies $\mathbb{E}_P[\phi^{pair}(D^{obs})] < \alpha$ for any $\alpha \in (0, 1)$, where the expectation is taken with respect to $D^{obs} \sim P$.*

Proofs and discussions on cases with many potential treatment assignments are provided in Appendix C. Similar to Guan (2023), for non-directional tests, the absolute value of the test statistic can be used. For directional tests, the statistic can be applied to test for positive effects, or negations of the statistic can be used to test for negative effects.

3.4 The Minimization-Based PNRT

The main limitation of the pairwise comparison-based PNRT is that when rejecting the null hypothesis at significance level α, the probability of a false rejection can be as high as 2α instead of α. While one way to address this is to reject the null hypothesis when the p-value is below $\alpha/2$, this adjustment might result in a loss of power. As an alternative, a more conservative testing procedure inspired by Wen et al. (2023) can be considered. The core idea behind this minimization-based PNRT is to compute a test statistic that reflects the worst-case scenario across all possible treatment assignments. Specifically, I define the test statistic as

$$\tilde{T}(D^{obs}) = \min_{d \in \{0,1\}^N} T(Y_{\mathbb{I}(d)}(D^{obs}), D^{obs}),$$

where the test statistic T is evaluated for each potential treatment assignment d. Based on this, I define the p-value as follows.

Definition 10 (Minimization-based PNRT)**.** *The minimization-based PNRT is an unconditional randomization test defined by* $\phi^{min}(D^{obs}) = 1\{pval^{min}(D^{obs}) \leq \alpha\}$*, where* $pval^{min}(D^{obs}) : \{0,1\}^N \to [0,1]$ *is the p-value function:*

$$pval^{min}(D^{obs}) = P(T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), D) \geq \tilde{T}(D^{obs})) \text{ for } D \sim P(D).$$

Here, $T(Y_{\mathbb{I}(d)}(d), d')$ *represents the pairwise imputable statistic used to evaluate the hypothesis.*

To calculate this p-value in practice, Algorithm 2 is applied. It computes the mean of $1 + R$ draws, where $r = 0$ corresponds to $d = D^{obs}$.

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 α.

for $r = 1$ *to* R **do**

  Randomly sample $d^r \sim P(D)$, and 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: $\tilde{T}^{\star}(D^{obs}) = \min_{r=1,\ldots,R}(T_r^{obs})$

Output : p-value: $\hat{pval}^{min} = \frac{1+\sum_{r=1}^{R} 1\{T_r \geq \tilde{T}^{\star}(D^{obs})\}}{1+R}$.

Reject if $\hat{pval}^{min} \leq \alpha$.

In the toy example, shown in Table 6, $\tilde{T}(D^{obs}) = 1$; thus, $pval^{min} = 1/2$. The crucial distinction between minimization-based PNRT and pairwise comparison-based PNRT is that minimization ensures size control, as demonstrated by Theorem 2.

Theorem 2. *Suppose the partial null hypothesis $H_0^{\epsilon_s}$ holds. Then, the minimization-based PNRT, as defined in Definition 10, satisfies $\mathbb{E}_P[\phi^{min}(D^{obs})] \leq \alpha$ for any $\alpha \in (0, 1)$, where the expectation is taken with respect to $D^{obs} \sim P$.*

A detailed proof is provided in Appendix C.

Handling a Large Number of Potential Treatment Assignments. When N is large, finding the minimum $\tilde{T}(D^{obs})$ across all possible treatment assignments can be computationally intensive. Nonetheless, simulations in Section 5.1 show that running Algorithm 2 with $R = 1,000$, though it may not identify the true minimum, remains conservative and maintains the test's validity.

To ensure the validity of Algorithm 2 when dealing with a large number of units, optimization methods can be used to approximate $\tilde{T}^R(D^{obs})$ such that $\tilde{T}(D^{obs}) \geq \tilde{T}^R(D^{obs}) - \eta_R$ with probability $1 - \eta$. The rejection level can then be adjusted to $\tilde{\alpha}$ such that $\alpha = \tilde{\alpha}(1 - \eta) + \eta$, thereby ensuring validity. However, this approach introduces additional computational complexity.

An alternative strategy is to combine CRTs with PNRTs to reduce the space of potential treatment assignments. As discussed by Athey et al. (2018) and Zhang and Zhao (2023), researchers often limit the assignment space to only those assignments with the same number of treated units as in the observed assignment. This two-stage approach—first defining the number of treated units and then performing testing within the reduced assignment space—remains valid. Using PNRTs in this context increases the set of focal units, potentially improving test power.

Conservatism Versus Power. To balance computational efficiency with validity, researchers may prefer the pairwise comparison-based PNRT from Algorithm 1, applied at a conservative rejection level of $\alpha/2$. Simulations in Section 5.1 demonstrate that this approach achieves higher power than the minimization-based PNRT while maintaining an element of conservatism. This adjustment serves as a safeguard in worst-case scenarios, though in some cases, a standard rejection level of α may also be appropriate (see Section 5.1 for further details). In practice, combining the p-values from both PNRT methods, as suggested

in DiCiccio et al. (2020), could enhance robustness. Developing refined strategies to integrate information from both methods, potentially increasing power, represents a promising direction for future research.

3.5 Comparison to Previous Literature

As previously discussed, the key distinction in PNRTs is that the set of units included in $\mathbb{I}(d)$ varies across different assignments d, allowing me to use all imputable units for testing. The procedure in Owusu (2023) shares a similar characteristic, but their approach is more complex to implement, involving tuning parameters, and is only valid asymptotically. In contrast, PNRTs are straightforward to implement, require no tuning parameters, and are valid in finite samples.

Under the sharp null hypothesis, $\mathbb{I}(d) = \{1, \ldots, N\}$ for all $d \in \{0,1\}^N$. As a result,

$$\tilde{T}(D^{obs}) = \min_{d \in \{0,1\}^N} T(Y_{\mathbb{I}(d)}(D^{obs}), D^{obs}) = T(Y(D^{obs}), D^{obs}) \quad \text{for any } D^{obs} \in \{0,1\}^N.$$

Thus, both the pairwise comparison-based PNRT and the minimization-based PNRT reduce to the classical FRT. The proposed method generalizes the FRT framework, ensuring validity under the partial null hypothesis by allowing the set of units in the test statistic to vary across different assignments.

Additionally, since the value of $T(Y_{\mathbb{I}(d)}(D^{obs}), D^{obs})$ only depends on $\mathbb{I}(d)$ for a fixed D^{obs}, when N is large, the variation in $\mathbb{I}(d)$ may become small. In such cases, the pairwise comparison-based PNRT could behave similarly to the minimization-based PNRT, achieving asymptotic validity at level α, as opposed to the more conservative 2α bound observed in finite samples. A formal proof might be worth exploring in future work.

Comparison to the CRTs When testing under a partial null hypothesis, the p-values constructed in Definitions 9 and 10 align closely with those from CRTs if we interpret $\mathbb{I}(D^{obs})$ as a focal unit set and $\{0,1\}^N$ as a focal assignment set. The pair $(\mathbb{I}(D^{obs}), \{0,1\}^N)$ represents a broader conditioning event than the traditional events in CRTs, which may or may not yield higher statistical power depending on whether the additional potential units and assignments contribute meaningful information.

In scenarios where a conditioning event can be specified over all imputable units in $\mathbb{I}(D^{obs})$, as demonstrated in Basse et al. (2024), CRTs with a well-defined focal assignment set may facilitate more targeted comparisons, potentially leading to a higher power. However,

in cases where designing a suitable conditioning event is either infeasible or would produce only a limited number of focal units and assignments, PNRTs can serve as a more practical alternative. Broadly, when including all assignments from $\{0,1\}^N$ is suboptimal, combining PNRTs with CRTs could improve power by focusing on more pertinent test statistics and selected assignments (Lehmann and Romano, 2005; Hennessy et al., 2015). This approach highlights a promising avenue for future research on optimizing power through the flexibility of PNRTs and CRTs.

Having discussed the main methods of pairwise comparison-based and minimization-based PNRTs, I now extend this framework to address practical concerns about determining the boundary of interference in real-world applications.

4 Extension: Framework to Determine the Boundary of Interference

Building on the PNRT framework introduced in the previous section, this extension focuses on determining the boundary of interference by estimating a sequence of partial null hypotheses at varying distances ϵ_s. This approach is useful for selecting a pure control distance or assessing the extent of interference based on distance. To this end, I 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 settings introduced in previous sections. For instance, if the goal is to test for the existence of interference, one could set $K = 1$ with $\epsilon_0 = \epsilon_s = 0$ and $\epsilon_1 = \epsilon_c$.

Using this sequence of distances, I can test a series of null hypotheses as defined in Definition 3, where $\epsilon_s \in \{\epsilon_0, \ldots, \epsilon_K\}$. However, it is important to note that not all distance levels will yield non-trivial power. First, there is a trade-off between the number of thresholds tested and the power of each test. While testing more thresholds provides a richer understanding of how interference varies with distance, it can reduce the power to detect interference, especially if certain threshold groups lack sufficient units. Based on simulation results, I recommend ensuring that each exposure level includes at least 20 units to maintain sufficient power at a significance level of $\alpha = 0.05$.

Second, in some cases, ϵ_K may represent the maximum distance in the network, leaving no further room for ϵ_c. Although it remains possible to test $H_0^{\epsilon_K}$, alternative approaches—such as adjusting for the number of nearby treated units, as suggested by Hoshino and Yanagi

(2023)—may be needed to construct a test statistic with non-trivial power. For simplicity, this section will focus on testing $H_0^{\epsilon_k}$ for $k \leq K-1$.

Following Definition 3 of $H_0^{\epsilon_s}$, the multiple hypotheses under consideration exhibit a nested structure:

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

The proof is provided in Appendix C.

Proposition 2 implies that interference is bounded within a certain distance. Given this nested structure, I aim to develop an inference method that determines such boundaries by rejecting the null hypothesis up to a certain distance and failing to reject it beyond that point. However, in practice, situations may arise where $H_0^{\epsilon_k}$ cannot be rejected but $H_0^{\epsilon_{k+1}}$ is rejected. This could happen either because the test lacks power to reject the false null $H_0^{\epsilon_k}$ or due to multiple hypothesis testing errors, which lead to an erroneous rejection of the true null $H_0^{\epsilon_{k+1}}$. To mitigate the risk of over-rejecting true null hypotheses, I propose controlling the FWER.

Definition 11 (FWER over all $H_0^{\epsilon_k}$ for $k = 0, \ldots, K-1$)**.** *Given a test $\varphi : \{0,1\}^N \to \{0,1\}^K$, which maps the data to decisions for each hypothesis $H_0^{\epsilon_k}$, the family-wise error rate (FWER) is defined as*

$$FWER = P\left(\exists k \geq \bar{K} \text{ such that } \varphi_k(D^{obs}) = 1, \text{ meaning that } H_0^{\epsilon_k} \text{ is rejected}\right),$$

where $\bar{K} \geq 0$ is such that for any $k \leq \bar{K}-1$, $H_0^{\epsilon_k}$ is false, and for $k \geq \bar{K}$, $H_0^{\epsilon_k}$ is true.

The definition of the FWER in Definition 11 is motivated by the nested structure of $H_0^{\epsilon_k}$, where the null hypothesis is true for any $k \geq \bar{K}$. The critical issue is determining how to reject all the $H_0^{\epsilon_k}$ hypotheses when identifying the boundary of interference, while still ensuring control over the FWER.

4.1 A Valid Procedure to Determine the Neighborhood of Interference

A major challenge in testing the extent of interference with respect to 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 arising from multiple

tests, and drawing inspiration from Meinshausen (2008) and Section 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}, and treatment assignment mechanism $P(D)$.
Set : $\hat{K} = 0$.
for $k = 0$ *to* $K - 1$ **do**
  Test $H_0^{\epsilon_k}$ using the PNRT procedure and 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 leveraging the nested structure of sequential hypothesis testing. Unlike traditional multiple hypothesis testing procedures, such as the Bonferroni-Holm method, which require rejecting at a smaller level than α, this algorithm maintains the significance level without adjustment, potentially increasing power compared to conventional methods (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 adjusting for multiple hypothesis testing, as we would naturally stop rejecting 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 FWER at α.*

The proof is provided in Appendix C.

For example, suppose $K = 2$ with $(\epsilon_0, \epsilon_1, \epsilon_2) = (0, 1, 2)$. 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 Rationale for Using the FWER

In practice, controlling the FWER is not the only criterion for managing error rates in multiple hypothesis testing. As Anderson (2008) suggests, a false discovery rate (FDR) control may be particularly useful in exploratory analyses, as it allows a small number of type I errors in exchange for greater statistical power compared to a FWER control. A more flexible adjustment algorithm, such as an FDR control, might be worth exploring in future work. However, in cases where policymakers aim to implement a policy in a distant region while anticipating a positive far-distance interference effect, the more restrictive FWER control can help prevent overly optimistic conclusions about the interference boundary. Thus, the FWER remains valuable by providing a conservative distance threshold that better accounts for interference when estimating expected welfare changes.

Moreover, this procedure helps in selecting the pure control group. As discussed in Section 3.1, it is often necessary to define a "safe distance" ϵ_c to construct a pure control group. But how should this distance be chosen? A natural candidate is ϵ_K, which represents the furthest distance where non-trivial power for testing is retained. That said, there may be a temptation to reduce this distance to include more units in the control group and thereby boost the test's power. Algorithm 3 offers a principled approach to determining ϵ_c. However, post-model selection issues may arise, as the selected ϵ_c could be smaller than the true interference boundary due to the algorithm's conservative nature. Researchers should weigh the benefits of implementing a pre-testing step in their analysis. A special case where post-model selection concerns are avoided is when testing the same sequence of null hypotheses. For a more detailed discussion, see Appendix B.

5 Empirical Application: Blattman et al. (2021)

In 2016, a large-scale experiment was conducted in Bogotá, Colombia, as described by Blattman et al. (2021). The study involved 136,984 street segments, with 1,919 identified as crime hotspots. Among these, 756 were randomly assigned to a treatment involving increased daily police patrolling from 92 to 169 minutes over eight months. It also included a secondary intervention aimed at enhancing municipal services, though this is peripheral to the primary focus of my empirical application. The key outcome of interest was the number of crimes per street segment, encompassing both property crimes and violent crimes (e.g., assault, rape, and murder).

Figure 6a shows the distribution of hotspots, with many located close to each other.

While only 756 street segments received the treatment, every segment potentially experienced spillover effects, creating a dense network. This complicates the application of cluster-robust standard errors for addressing unit correlation. The study estimated a negative treatment effect and used FRTs with a sharp null hypothesis of no effect for inference.

Figure 6: Map of Experimental Sample and Treatment Conditions

(a) Experimental Sample Map

(b) Assignment to Experimental Conditions

Notes: Panel (a) displays a map of the experimental sample, with hotspot street segments marked in red. Panel (b) shows an example of the assignment to the four experimental conditions. Source: Blattman et al. (2021).

In evaluating the total welfare impact of the policy, it is essential to assess whether interference occurred following treatment assignment, such as crime displacement or deterrence in neighboring areas. I aim to address three key questions: 1) Does interference exist? 2) If so, is it displacement or deterrence? 3) At what distance is this interference effective? Given the complexity of modeling correlation across units in 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 interval $(500, \infty)$ represents a pure control group with no treated units within 500 meters. Figure 6b provides an example of the distance intervals identified in

Blattman et al. (2021).

Table 8 presents descriptive statistics for the number of crimes during the intervention period. The t-statistics for t-tests between each of the two columns reveal two key findings: treated hotspots experience significantly fewer crimes, and non-hotspot areas report even fewer crimes when located farther from any treated units. However, the extremely high values of the t-statistics raise concerns about interpreting these t-test results as evidence of a displacement effect. As noted earlier, standard errors may be under-estimated, and units at varying distances from treated areas may not be directly comparable. Both factors could contribute to the high t-statistics observed in the table.

Table 8: Descriptive Statistics During the Intervention

Stats	**Crime hotspots**		**Non-hotspots (distance to treated units)**			
	Treated	Non-treated	$(0m, 125m]$	$(125m, 250m]$	$(250m, 500m]$	$(500m, \infty)$
Obs.	756	1,163	24,571	32,034	45,147	33,313
# of total crimes						
Mean	0.935	1.311	0.378	0.294	0.242	0.180
SD	1.519	2.332	1.006	0.921	0.736	0.602
Max	12	43	33	40	25	31
% of > 0	44.84	53.22	23.17	19.01	16.69	13.13
t-stat of t-test		-3.93		10.34	8.60	12.62
# of property crimes						
Mean	0.712	1.035	0.262	0.195	0.158	0.111
SD	1.269	2.099	0.778	0.683	0.555	0.441
Max	12	40	32	36	21	27
% of > 0	38.36	45.66	17.44	13.96	11.90	8.86
t-stat of t-test		-3.81		10.93	8.26	12.83
# of violent crimes						
Mean	0.224	0.276	0.115	0.099	0.084	0.069
SD	0.593	0.650	0.467	0.473	0.376	0.334
Max	5	6	17	40	13	11
% of > 0	16.40	20.29	8.59	7.29	6.51	5.47
t-stat of t-test		-1.79		4.23	4.75	5.79

Notes: This table presents descriptive statistics for crime data during the intervention, divided into two categories: crime hotspots (treated and non-treated) and non-hotspot areas, which are further grouped by their distance from the treated units. The statistics cover three types of crimes: total crimes, property crimes, and violent crimes. For each group, the table provides the mean, standard deviation (SD), maximum (Max), and the percentage of units with positive crimes (% of > 0). The t-statistic values (t-stat of t-test) represent the results from t-tests comparing the difference in means between treated versus non-treated units within crime hotspots, non-hotspot units within $(0m, 125m]$ versus $(125m, 250m]$, non-hotspot units within $(125m, 250m]$ versus $(250m, 500m]$, and non-hotspot units within $(250m, 500m]$ versus $(500m, \infty)$.

In Blattman et al. (2021), the authors reported no significant displacement effect for violent crimes and a marginally significant displacement effect for property crimes.[13] However, as previously illustrated, p-values for the t-test may not adequately capture the extent of interference. Moreover, using FRTs to test partial null hypotheses may not be valid. Thus, how might these conclusions change if a valid testing approach is implemented?

5.1 Power Comparison of Spatial Interference: A Simulation Study

To comprehensively test the methodology in a large-scale experiment, I conduct a simulation study to preselect the preferred method. Specifically, I generate $N = 1,000$ points from a bivariate Gaussian distribution with non-diagonal covariance to simulate the network on a $[0,1] \times [0,1]$ space, including 20 hotspots and 7 randomly treated units, reflecting proportions similar to the original Bogotá study.

Figure 7: Unit Distribution

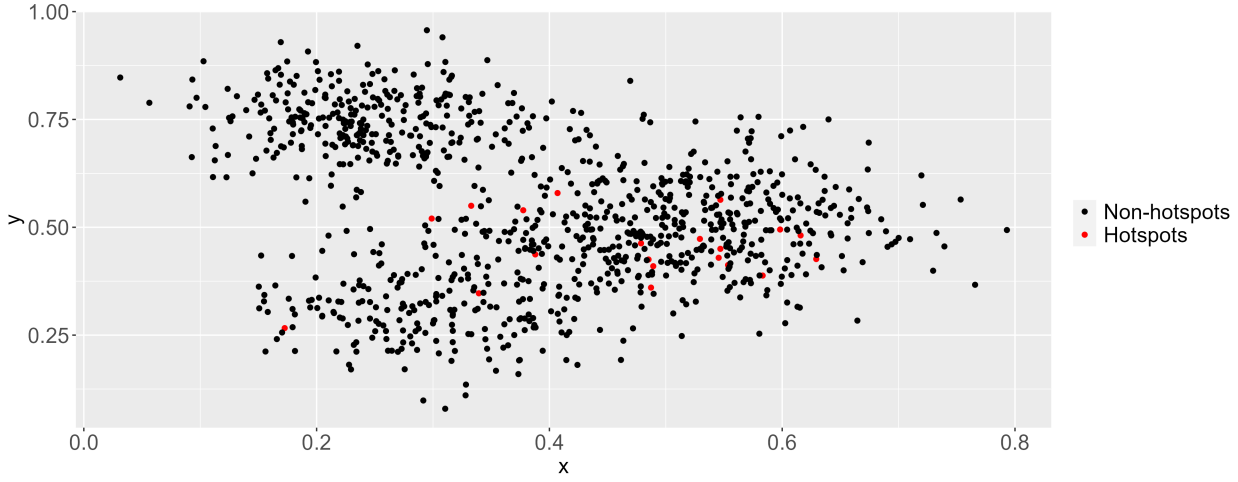


Figure 7 shows the unit distribution within this space. To simplify the analysis, I focus on two distance thresholds, with $(\epsilon_0, \epsilon_1, \epsilon_2) = (0, 0.1, 0.2)$. Across different treatment assignments, the distance interval $(0, 0.1]$ comprises approximately 420 units, $(0.1, 0.2]$ around 250 units, and the pure control group $(0.2, \infty)$ around 320 units.

The partial null hypothesis for $k = 0$ and 1 is given by

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

The potential outcome schedule is calibrated to match the Bogotá street network using

[13] The treatment effect for violent crime was significant, but property crime effects were insignificant.

gamma distributions, ensuring alignment with the observed mean and variance of total crimes, as outlined in Table 9. I set a negative treatment effect of 1 while maintaining non-negative crime numbers for all treated units. Additionally, I incorporate a decreasing displacement effect relative to distance using a positive τ. The primary focus of this analysis is on the spillover effect τ.

Table 9: 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:	$Y_i(d) = Y_i^C + 0.5\tau \quad \forall d \in \mathcal{D}_i(0.1)/\mathcal{D}_i(0.2)$

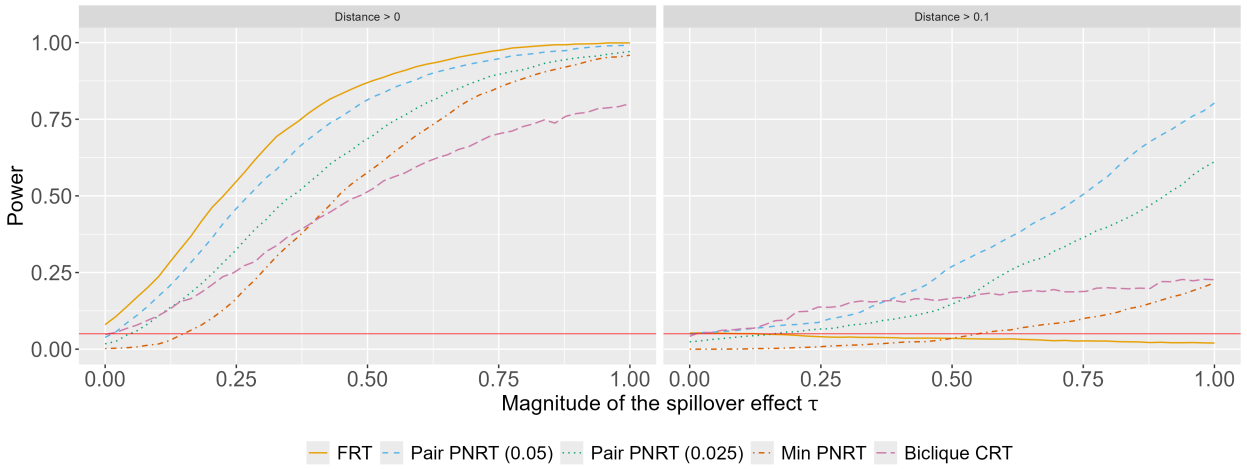
Notes: For $Gamma(k, \theta)$, k is the shape parameter and θ is the scale parameter. Y_i^C represents the pure control potential outcome for unit i, and Y_i^T represents the potential outcome for unit i when treated.

In the analysis, I compare five methods: 1) the classic FRT, using the sharp null hypothesis of no effect, as used in Blattman et al. (2021) for spillover effect inference; 2) the biclique CRT proposed by Puelz et al. (2021), a benchmark for CRTs due to their strong power properties in simulations involving general interference; 3) the minimization-based PNRT following Algorithm 2, which uses the minimum of $T(Y_{\mathbb{I}(d^r)}(D^{obs}), D^{obs})$ across R random assignments; 4) the pairwise comparison-based PNRT with rejection based on $\alpha/2$, ensuring validity in the worst-case scenario; and 5) the pairwise comparison-based PNRT with rejection based on α.

Two main criteria guide the testing procedure selection. First, in the absence of a spillover effect ($\tau = 0$), the partial null hypothesis should be rejected no more than 5% of the time, maintaining type I error control. Second, when a spillover effect exists ($\tau > 0$), the partial null hypothesis should be rejected as frequently as possible to maximize power. To assess power, I consider 50 equally spaced τ values between 0 and 1, conducting 2,000 simulations for each τ to compute the average rejection rate for each method. The algorithm is detailed in Appendix B, with a focus on displacement effects and one-sided testing using the non-absolute difference in mean.

Figure 8 (left panel) shows that the FRT over-rejects the true partial null hypothesis when $\tau = 0$, consistent with Athey et al. (2018)'s observation that testing the sharp null of no effect is invalid for partial null hypotheses. In my simulation, with only 7 treated units (0.7% of the total), the rejection rate is around 10%. Surprisingly, the pairwise comparison-based PNRT without adjustment at the α level maintains good size control, indicating that the 2α control is a worst-case guarantee. Other PNRT algorithms are also valid but more

Figure 8: Power Comparison of Testing Methods for Different Hypotheses



Notes: The left panel shows the power comparison for testing H_0^0, while the right panel illustrates the power comparison for $H_0^{0.1}$. The red line indicates the size level $\alpha = 0.05$. "Min PNRT" represents the minimization-based PNRT from Algorithm 2, "Pair PNRT (0.025)" indicates the pairwise comparison-based PNRT with a rejection threshold of $\alpha/2$, and "Pair PNRT (0.05)" denotes the pairwise comparison-based PNRT with rejection based on α.

conservative, with rejection rates below 5%. The biclique CRT remains valid with a rejection rate near 5%.

Regarding power, the FRT is excluded from the comparison due to its invalidity. The unadjusted pairwise comparison-based PNRT shows the best performance, dominating other methods across all effect magnitudes τ. Among methods with theoretical size control, the pairwise comparison-based PNRT with $\alpha/2$ rejection is optimal, though it slightly lags behind the biclique CRT for small τ values. The minimization-based PNRT, though dominated by the pairwise comparison-based PNRT, outperforms the biclique CRT for larger τ values. Despite its validity, the biclique CRT lacks power, with a rejection rate below 90% even when $\tau = 1$.

The right panel of Figure 8 provides a contrast to the results shown in the left panel. First, all methods, including the FRT, maintain validity under the null hypothesis. This may be attributed to the fact that hotspots rarely belong to the exposure levels $(0.1, 0.2]$ or $(0.2, \infty)$. As a result, despite the negative treatment effect, the impact on the test statistics used for this analysis remains minimal. Similar to the case of H_0^0, both the pairwise comparison-based PNRT and biclique CRT methods exhibit 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 demonstrate significantly lower power compared to the results from H_0^0. This is largely because only 60% of the units are relevant to the partial null hypothesis in this scenario, and the spillover effect magnitude is halved (0.5τ). Nevertheless, the pairwise comparison-based PNRT method still shows sufficient power when the spillover magnitude τ is large enough, outperforming other methods, particularly when using the unadjusted rejection level α. The minimization-based PNRT appears overly conservative, as it involves minimization over a large number of treatment assignments. Surprisingly, the FRT exhibits under-rejection, with almost no power across all values of τ. This can be explained by the nature of the FRT: the p-value remains large unless the observed test statistic exceeds most test statistics generated from randomized treatment assignments. However, since units within the group $(0, 0.1]$ under the observed assignment contribute to test statistics for other randomized assignments d—and these units are subject to a spillover effect of τ—the observed test statistics constructed from units in $(0.1, 0.2]$ and $(0.2, \infty)$ fail to display high values, even when τ is large. This results in large p-values for the FRT, explaining its under-rejection. Combined with the earlier discussion for H_0^0, this illustrates how the FRT, when testing the partial null hypotheses, may lead to either over- or under-rejection depending on the scenario. Finally, although the biclique CRT method demonstrates power, its increase is slower than that of the PNRT methods, which may be attributed to the complexity of finding an optimal conditioning event in the presence of spatial interference.

Overall, the results favor PNRT methods, especially the unadjusted pairwise comparison-based PNRT. Thus, I apply PNRTs to replicate the results of Blattman et al. (2021), using the non-absolute difference-in-mean estimator.

5.2 Implementation of PNRTs for Testing the Existence of a Displacement Effect

Consider the experimental setting described in Blattman et al. (2021), where the observed treatment assignment is denoted by D^{obs}. Similar to Blattman et al. (2021), we can regress the number of crimes, Y, on a spillover proximity indicator $S(D^{obs})$, which indicates whether units are within 125 meters of any treated unit. This proximity indicator is directly determined by the treatment assignment D^{obs} and would change with a different treatment assignment D. While additional covariates can be included in the regression, the key test statistic remains the coefficient from regressing Y on the proximity indicator $S(D)$.

To test the partial null hypothesis using the two proposed PNRTs—the pairwise comparison-based PNRT and the minimization-based PNRT—both methods involve the following steps:

1. Randomly reassign treatments D to the units.

2. For each reassignment D, identify the subsample of units that are not treated under both D^{obs} and D. These units form the set of imputable units, as their potential outcomes remain unaffected by the treatment under both assignments.

3. Collect the coefficient β from the regression of Y on $S(D^{obs})$ within the subsample of imputable units.

4. Collect the coefficient β' from the regression of Y on $S(D)$ within the same subsample of imputable units.

Pairwise Imputable Statistics. Both steps 3 and 4 rely on the concept of pairwise imputable statistics, which ensure that only units whose outcomes are not affected by either the observed treatment assignment D^{obs} or the randomized assignment D are included in the analysis. This guarantees that the test statistic, whether derived from D^{obs} or D, is constructed using the same set of units. The key idea here is to isolate the comparison to units whose outcomes remain unchanged across both assignments, thereby ensuring the validity of the test under the partial null hypothesis.

For the rejection decision when testing for displacement effects, the pairwise comparison-based PNRT computes the p-value as the fraction of reassignments D such that $\beta' \geq \beta$. The null hypothesis of no displacement effect is rejected if the p-value is less than or equal to $\alpha/2$. The simulation results also suggest that using α as the rejection threshold may be empirically valid.

For the minimization-based PNRT, the approach first determines the minimum value of β across all reassignments D—denoted by $\tilde{\beta}$—since each reassignment involves a regression on different subsamples. The p-value is then calculated as the fraction of reassignments D such that $\beta' \geq \tilde{\beta}$, and the null hypothesis is rejected if the p-value is less than or equal to α.

Both methods can accommodate two-sided tests by comparing $\|\beta'\| \geq \|\beta\|$ or by testing the deterrence effect with the negative of the coefficient, comparing $-\beta' \geq -\beta$.

5.3 PNRTs on Actual Data

I conduct the analysis using the publicly available dataset from Blattman et al. (2021), which includes street-level treatment assignments and distance intervals with thresholds at 125, 250, and 500 meters. The dataset also contains $1,000$ pseudo-randomized treatments and their respective distance intervals, used in the original paper for randomization inference. However, the dataset lacks precise longitude and latitude data for the street segments, preventing me

from extending randomization testing beyond the available $1,000$ random treatments.

Due to the large fraction of zero outcomes, I use both the indicator of any crime and the number of crimes as the outcome variable. Table 10 shows the results for the indicator of any crime, and Table 11 presents the results for the number of crimes as the outcome variable.

Table 10: Indicator of Any Crime: p-Values for Testing the Displacement Effect at Different Distances

	Unadjusted p-values		
	$(0m, \infty)$	$(125m, \infty)$	$(250m, \infty)$
Violent crime			
Pair PNRT	*0.027*	0.812	*0.060*
Pair PNRT + reg	0.110	0.887	0.195
Min PNRT	*0.043*	0.955	0.651
Property crime			
Pair PNRT	0.390	0.466	0.486
Pair PNRT + reg	0.563	0.510	0.668
Min PNRT	0.478	0.819	0.876

Notes: The table shows the impact of intensive policing on violent and property crime, using the indicator for any crime as the outcome variable. Pair PNRT denotes the pairwise comparison-based PNRT with the difference-in-mean estimator as the test statistic. Min PNRT is the minimization-based PNRT with the difference-in-mean estimator as the test statistic. Pair PNRT + reg is the 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.

In the main specification, I use the pairwise comparison-based PNRT with the difference-in-mean estimator as the test statistic. To assess robustness, I also consider two alternative approaches: the minimization-based PNRT using the difference-in-mean estimator and the pairwise comparison-based PNRT using the coefficient from a covariate-adjusted regression. These methods help evaluate how the results change when incorporating additional factors, following the framework outlined in Blattman et al. (2021), with slight modifications.

First, the regression includes the same covariates used in Blattman et al. (2021), such as police station fixed effects, but excludes those related to the municipal services intervention.[14] In the original study, randomization testing was conducted jointly for both the policing and

[14]The covariates include the following: the number of crimes (2012–2015); average patrol time per day; square meters built (100 meters around) per meter of longitude; distance to the nearest shopping center, educational center, religious/cultural center, health center, and additional services office (e.g., justice); transport infrastructure (e.g., bus/BRT station); indicators for industry/commerce zones and service sector zones; income level; eligibility for municipal services; and interactions with the crime hotspot indicator.

Table 11: Number of Crimes: p-Values for Testing the Displacement 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

Notes: The table shows the impact of intensive policing on violent and property crime, using the number of crimes as the outcome variable. Pair PNRT denotes the pairwise comparison-based PNRT using the difference-in-mean estimator. Min PNRT is the minimization-based PNRT using the difference-in-mean estimator. Pair PNRT + reg is the pairwise comparison-based PNRT using the coefficient from a regression that includes covariates (e.g., police station fixed effects) and inverse propensity weighting as the test statistic.

municipal services interventions, which makes interpretation more difficult when interaction effects are present. In this analysis, I hold the municipal services intervention fixed to isolate the effect of intensive policing. For robustness checks on different approaches to incorporating covariates, see Appendix D.

Second, Blattman et al. (2021) employ inverse propensity weighting in the weighted regression, using weights that account for both the hotspot policing and municipal services interventions.[15] In this replication, I use weights that only consider the hotspot policing intervention to focus on the impact of intensive policing.

Both Tables 10 and 11 reveal a significant displacement effect for violent crimes but not for property crimes. After adjusting for multiple hypothesis testing using Algorithm 3, both the pairwise comparison-based PNRT and minimization-based PNRT methods detect a significant short-range displacement effect within 125 meters at the 10% level using the difference-in-mean estimator. If we omit the $\alpha/2$ adjustment for the pairwise comparison-based PNRT, as suggested by the simulation study, the short-range spillover within 125 meters becomes significant at the 5% level using the difference-in-mean estimator and remains marginally significant at the 10% level with the regression coefficient. Additionally, there is no evidence of additional spillover effects beyond 125 meters for violent crimes and no

[15] Although this method does not fully eliminate bias, as discussed in Aronow et al. (2020), it helps address imbalance in the spillover group.

evidence of spillover effects at any distance for property crimes. While the unadjusted p-value of 0.045 for the $(250m, \infty)$ interval in the pairwise comparison-based PNRT might suggest a potential spillover effect within this range, it could also be a false discovery due to multiple hypothesis testing.[16]

In line with Puelz et al. (2021), the p-values tend to increase when covariates are included in the model, likely due to the heterogeneous nature of spillover effects. This suggests that geographic distance alone may not fully capture the intensity or pattern of these effects. Future work could improve the distance measure by incorporating additional factors, such as socioeconomic differences between street segments, as discussed in Puelz et al. (2021).

These results not only show the general applicability of the PNRT method but also provide suggestive evidence for policy implications and potential criminal motives in Bogotá, following the insights of Blattman et al. (2021). In terms of policy, it is unclear whether reallocating state resources to these hotspots has led to an overall reduction in crime, and thus further investigation is warranted. Regarding criminal motives in Bogotá, a potential explanation, consistent with standard economic models of crime, is that violent crime in the city's hotspots may not be solely expressive in nature, as implied by Blattman et al. (2021). Instead, some violent crimes—such as contract killings—might be 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 relocate rather than be entirely prevented. For property crimes, which are often instrumental and linked to immobile criminal rents, hotspot policing seems to deter crime without causing further spillover effects. As Blattman et al. (2021) pointed out, violent crimes are often considered more severe than property crimes, making the potential displacement effect a crucial consideration when evaluating the policy intervention's overall welfare impact.

6 Conclusion

This paper introduces a practical testing framework for detecting interference in network settings. The proposed tests offer computational simplicity over previous methods while retaining strong power and size properties, making them highly applicable for empirical research.

Theoretically, I formalize unconditional randomization testing and PNRTs, addressing two primary challenges in testing partial null hypotheses: only a subset of potential outcomes

[16]This might suggests that the distance interval $(500m, \infty)$ could serve as a more appropriate control group than the $(250m, \infty)$ interval used by Blattman et al. (2021).

is imputable, and the set of units with imputable potential outcomes varies across treatment assignments. PNRTs address the first challenge by employing pairwise imputable statistics and the second by constructing p-values through pairwise comparisons. I prove that both the pairwise comparison-based PNRT and the minimization-based PNRT maintain size control. Additionally, I propose a sequential testing procedure to estimate the "neighborhood" of interference, ensuring control over the FWER.

Beyond network settings, PNRTs hold broader applicability. For instance, Zhang and Zhao (2021) show that partial null hypotheses are relevant in time-staggered designs. This opens promising avenues for future research, including extending the framework to quasi-experimental settings and observational studies. In quasi-experimental designs, developing a unified framework that can be applied to time-staggered adoption, regression discontinuity, and network settings would be highly valuable (Borusyak and Hull, 2023; Kelly, 2021). For observational studies, incorporating propensity score weighting to create pseudo-random treatments and conducting sensitivity analyses would be crucial, as noted by Rosenbaum (2020).

One limitation of the method is its conservatism when ensuring finite-sample validity at a rejection threshold of $\alpha/2$. However, simulations indicate that the test remains empirically valid with a rejection level of α, showing stronger power properties than CRTs. This observation suggests several future directions. First, why does the method empirically achieve size control at level α? Is the method asymptotically valid with a rejection level of α, or does size control hold for smaller values of α? Second, while simulations suggest that PNRTs perform favorably compared to CRTs, their power properties remain unexplored. Insights from studies such as Basse et al. (2019) on CRT power properties and Wen et al. (2023) on the near-minimax optimality of minimization-based p-values suggest that further exploration of PNRTs' power could yield valuable insights. Additionally, power may increase when PNRTs are combined with CRTs in specific settings, making the construction of an optimal testing framework for interference an important question for future research.

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.

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.

Cai, J., A. De Janvry, and E. Sadoulet (2015): "Social Networks and the Decision to Insure," *American Economic Journal: Applied Economics*, 7, 81–108.

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.

DiCiccio, C. J., T. J. DiCiccio, and J. P. Romano (2020): "Exact tests via multiple data splitting," *Statistics & Probability Letters*, 166, 108865.

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 Inter-

ference 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.

Jayachandran, S., J. de Laat, E. F. Lambin, C. Y. Stanton, R. Audy, and N. E. Thomas (2017): "Cash for carbon: A randomized trial of payments for ecosystem services to reduce deforestation," *Science*, 357, 267–273.

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.

Leung, M. P. (2020): "Treatment and Spillover Effects Under Network Interference," *The Review of Economics and Statistics*, 102, 368–380.

——— (2022): "Causal Inference Under Approximate Neighborhood Interference," *Econometrica*, 90, 267–293.

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," .

Paluck, E., H. Shepherd, and P. Aronow (2016): "Changing climates of conflict: A social network experiment in 56 schools," *Proceedings of the National Academy of Sciences of the United States of America*, 113, 566–571.

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.

Shirani, S. and M. Bayati (2024): "Causal message-passing for experiments with unknown and general network interference," *Proceedings of the National Academy of Sciences*, 121.

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 Fisher Randomization Tests and Conditional Randomization Tests

A.1 Review of Fisher Randomization Tests

In Fisher (1925), Fisher introduced randomization testing under the sharp null hypothesis, which can be formally defined as follows.

Definition A.1 (Sharp null hypothesis)**.** *The sharp null hypothesis holds if*

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

This hypothesis is often referred to as the null hypothesis of no treatment effect whatsoever. Under this hypothesis, all missing potential outcomes are known, allowing the classical Fisher randomization test (FRT) to be applied since all potential outcomes can be imputed under the null (Zhang and Zhao, 2023).

The original 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: $Y_i(1)$ (the potential outcome when treated) and $Y_i(0)$ (the potential outcome when not treated) for every unit i. The typical null hypothesis tested in this setting is

$$H_0 : Y_i(0) = Y_i(1) \quad \text{for all } i = 1, 2, \ldots, N,$$

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

Let $T(Y, D) : \mathbb{R}^N \times \mathbb{D} \to \mathbb{R}$ denote a test statistic as a function of Y and D. A common example is the absolute difference in means between treated and control units:

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

where

$$\bar{Y}^{obs}_{\{i:D_i=1\}} = \frac{\sum_{i=1}^N 1\{D_i = 1\}Y_i}{\sum_{i=1}^N 1\{D_i = 1\}}, \quad \bar{Y}^{obs}_{\{i:D_i=0\}} = \frac{\sum_{i=1}^N 1\{D_i = 0\}Y_i}{\sum_{i=1}^N 1\{D_i = 0\}}.$$

For two-sided testing, the absolute value version of the test statistic can also be used.

Denote $T_{obs} = T(Y^{obs}, D^{obs})$. The p-value is then defined as

$$pval(D^{obs}) = P(T(Y^{obs}, D) \geq T_{obs}),$$

where the probability is taken with respect to $D \sim P(D)$. This p-value reflects a stochastic version of the "proof by contradiction" approach, as discussed by Imbens and Rubin (2015): if very few potential assignments D yield $T(Y^{obs}, D) \geq T_{obs}$, then observing T_{obs} under the null hypothesis is highly improbable, leading to a smaller p-value and a greater likelihood of rejecting the null hypothesis.

The formal testing procedure is outlined below, where the p-value is calculated as the mean value of $1 + R$ draws due to using $d = D^{obs}$ for $r = 0$. Thus, there are $R + 1$ draws.

Algorithm A.1 FRTs

Inputs : Test statistic $T = T(Y(d), d)$, observed assignment D^{obs}, observed outcome Y^{obs}, treatment assignment mechanism $P(D)$, and size α.

Compute: The observed test statistic, $T_{obs} = T(Y^{obs}, D^{obs})$.

for $r = 1$ *to* R **do**

  Randomly sample $d^r \sim P(D)$, and store $T_r \equiv T(Y^{obs}, d^r)$.

end

Output : p-value: $\hat{pval} = \frac{1+\sum_{r=1}^{R} 1\{T_r \geq T_{obs}\}}{1+R}$. Reject if p-value $\leq \alpha$.

If H_0 is true, $T_r = T(Y^{obs}, D') = T(Y(D'), D')$ for any $D' \sim P(D)$. Since D' is a random draw from $P(D)$, we have $T_r = T(Y(D'), D') \sim T(Y^{obs}, D^{obs}) = T_{obs}$. This ensures a 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).

Table A.1: Illustration of FRTs in the Example

Assignment D	Potential Outcome Y_i				$T(Y(D^{obs}), D)$
	i_1	i_2	i_3	i_4	
$(1,0,0,0)$	2	4	3	1	-2/3
$(0,1,0,0)$	2	4	3	1	2
$(0,0,1,0)$	2	4	3	1	2/3
$(0,0,0,1)$	2	4	3	1	-2

Notes: Assignment D includes all potential assignments, with the first row representing the observed assignment D^{obs}. Potential outcomes are the same for each potential assignment under the sharp null hypothesis of no effect. $T(Y^{obs}, D)$ is the value of the difference in means.

Running Example Continued. In the absence of interference (under SUTVA), Table A.1 illustrates the potential outcome schedule for FRTs in the example. According to Algorithm A.1, the observed test statistic is $-2/3$, which is greater than the statistic when unit i_4 is treated. Therefore, the p-value is $3/4$, and a uniform random variable can be used for

tiebreaking without affecting the test's validity (Lehmann and Romano, 2005). The absolute difference version of the test statistic can also be used for two-sided testing.

Using FRTs for the Partial Null Hypothesis. As discussed by Athey et al. (2018), using FRTs to test a partial null hypothesis can lead to over-rejection. The key issue is that implementing them for a partial null hypothesis mistakenly treats missing potential outcomes as if they were imputed under the sharp null hypothesis of no effect. Consequently, the test may ignore the variation caused by treatment effects. For example, Bond et al. (2012) use a permutation test to examine spillover effects from a randomly assigned encouragement to vote in the 2010 U.S. elections. Although Bond et al. (2012) reject the sharp null hypothesis, Athey et al. (2018) show that such tests could significantly inflate type I error rates.

A.2 Review of Conditional Randomization Tests

Pioneered by Aronow (2012) and Athey et al. (2018), conditional randomization tests (CRTs) have become a key tool in addressing interference in network settings. Building on this foundational work, recent studies such as Basse et al. (2019) and Puelz et al. (2021) have extended CRT methods to tackle various interference structures. The main idea behind CRTs is that while the null hypothesis $H_0^{\epsilon_s}$ is not generally sharp, it can be "made sharp" by focusing on 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 consists of a subset of units, referred to as "focal units" ($\mathbb{N}_\mathbb{U} \subseteq \{1, \ldots, N\}$), and a subset of treatment assignments, referred to as "focal assignments" ($\mathbb{D}_\mathbb{U} \subseteq \{0,1\}^N$), satisfying the following condition.

Definition A.2 (Conditions for conditioning event)**.** *Let $H_0^{\epsilon_s}$ be the partial null hypothesis. A conditioning event $\mathbb{U} = (\mathbb{N}_\mathbb{U}, \mathbb{D}_\mathbb{U})$, where $\mathbb{N}_\mathbb{U} \subseteq \{1, \ldots, N\}$ is a subset of units and $\mathbb{D}_\mathbb{U} \subseteq \{0,1\}^N$ is a subset of treatment assignments, satisfies the following condition: For each unit $i \in \mathbb{N}_\mathbb{U}$ and all treatment assignments $d, d' \in \mathbb{D}_\mathbb{U}$, the potential outcomes of unit i are equal under $H_0^{\epsilon_s}$, i.e.,*

$$Y_i(d) = Y_i(d').$$

Definition A.2 aligns with the formulations in Athey et al. (2018) and Basse et al. (2019), emphasizing the restrictive nature of the conditioning event. Basse et al. (2019) reinterpret this condition in the context of exposure mapping, which is a lower-dimensional summary of the treatment assignments but can be misspecified in practice. Both Athey et al. (2018) and Basse et al. (2019) allow $\mathbb{N}_\mathbb{U} \nsubseteq \mathbb{I}(D^{obs})$, though they restrict the focal assignments to control

the exposure levels of units not included in $\mathbb{I}(D^{obs})$. This leads to a final implementation that primarily relies on focal units within $\mathbb{I}(D^{obs})$.

Different studies propose various methods to select $\mathbb{U}$ via $P(\mathbb{U}|D^{obs})$. For example, Aronow (2012) and Athey et al. (2018) consider conditioning mechanisms of the form $P(\mathbb{U}|D^{obs}) = P(\mathbb{U})$, where conditioning is either random or based on auxiliary information but not directly related to the observed assignment. This omission of observed information can result in a loss of power. To address this, Basse et al. (2019) propose a two-step conditioning mechanism tailored to cluster interference, where they sample from a carefully constructed distribution $P(\mathbb{U}|D^{obs})$ and then run the test conditional on $\mathbb{U}$. Extending this framework, Puelz et al. (2021) introduce the biclique decomposition method to handle both clustered and spatial interference by constructing conditioning events for general interference.

The restricted test statistic under the conditioning event, $T^{\mathbb{U}}(Y, d)$, is defined as follows.

Definition A.3 (Conditioning event restricted test statistic)**.** *Let* $T^{\mathbb{U}}(Y, d) : \mathbb{R}^N \times \{0, 1\}^N \to \mathbb{R}$ *be a measurable function. The statistic* $T^{\mathbb{U}}$ *is called a conditioning event restricted test statistic if, for any* $(Y, Y', d, d') \in \mathbb{R}^{2N} \times \{0, 1\}^{2N}$*, we have*

$$T^{\mathbb{U}}(Y, d) = T^{\mathbb{U}}(Y', d')$$

whenever $Y_i = Y'_i$ *and* $d_i = d'_i$ *for all* $i \in \mathbb{N}_{\mathbb{U}}$.

The test statistic in Definition A.3 is similar to the pairwise imputable statistic defined earlier. The p-value is computed using the same procedure as in FRTs, except all computations are restricted to the conditioning event $\mathbb{U}$. The following algorithm outlines this process.

Algorithm A.2 CRTs

Inputs : Test statistic $T = T(Y(d), d)$, observed assignment D^{obs}, observed outcome Y^{obs}, treatment assignment mechanism $P(D)$, size α, and the conditioning event design $P(\mathbb{U}|D^{obs})$.

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 $d^r \sim P(D|\mathbb{U}) \propto P(\mathbb{U}|D)P(D)$, and store $T^{\mathbb{U}}_r \equiv T^{\mathbb{U}}(Y^{obs}, d^r)$.

end

Output : p-value: $\hat{pval} = \frac{1+\sum_{r=1}^{R} 1\{T^{\mathbb{U}}_r \geq T^{\mathbb{U}}_{obs}\}}{1+R}$.
Reject if p-value $\leq \alpha$.

Running Example Continued. As per Definition A.2, we require that for all $d \in \mathbb{D}_{\mathbb{U}}$ and $i \in \mathbb{N}_{\mathbb{U}}$, $d_i = 0$ for H_0^0. If there exists any i with $d_i = 1$, for any other $d' \in \mathbb{D}_{\mathbb{U}}$, we would also need $d'_i = 1$. However, in the experimental design, only one unit is treated, resulting in identical assignments across all $d \in \mathbb{D}_{\mathbb{U}}$. This essentially limits power due to only one effective treatment.

Following Definition A.3, we can use the difference-in-means estimator as follows:

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

For example, if unit i_1 is treated in the observed D^{obs}, one valid $\mathbb{U}$ might include $\mathbb{N}_{\mathbb{U}} = \{i_2, i_4\}$ and $\mathbb{D}_{\mathbb{U}} = \{(1,0,0,0), (0,0,1,0)\}$. The potential outcomes are shown in Table A.2.

Table A.2: CRTs in the Example

Assignment D	Potential Outcome Y_i		$T^{\mathbb{U}}(Y^{obs}, D)$
	i_2	i_4	
$(1,0,0,0)$	4	1	3
$(0,0,1,0)$	4	1	-3

Notes: Assignment D includes all potential assignments, with the first row representing the observed assignment D^{obs}. Potential outcomes are shown for each unit under the null H_0^0. Blue cells represent the units used to calculate the mean value in the first term of the test statistics. $T^{\mathbb{U}}(Y^{obs}, D)$ represents test statistics under different D while holding D^{obs} fixed.

Following Algorithm A.2, the p-value is $1/2$ since the observed test statistic is the highest among the sampled test statistics. However, in practice, designing $\mathbb{U}$ and $P(\mathbb{U}|D^{obs})$ in more complex settings to ensure meaningful power can be challenging.

Appendix B Other Algorithms in the Main Text

Procedure for Selecting a Pure Control Group. Theorem B.1 guarantees the validity of $H_0^{\epsilon_k}$ as long as $H_0^{\epsilon_k}$ holds, even after selecting ϵ_c using Algorithm 3 as an initial step. This approach potentially addresses the post-model selection inference issue highlighted by Leeb and Pötscher (2005), ensuring post-inference validity even after the pre-testing step. Consequently, researchers may use the distance $\epsilon_{\hat{K}}$ obtained from Algorithm 3 as the threshold ϵ_c for further analysis.

Definition B.1 (Two-step pre-testing procedure)**.** *Given an observed assignment D^{obs} and a partial null hypothesis $H_0^{\epsilon_k}$ for $k \in \{0, \ldots, K-1\}$, the two-step pre-testing procedure is defined as follows:*

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

Step 2: Use $\epsilon_c = \epsilon_{\hat{K}}$ to test $H_0^{\epsilon_k}$ and compute the p-value, $pval(D^{obs})$, using any valid inference method.

Theorem B.1. *Suppose the partial null hypothesis $H_0^{\epsilon_k}$ is true. Then, the two-step pre-testing procedure in Definition B.1 satisfies*

$$P(pval(D^{obs}) \leq \alpha) \leq \alpha.$$

The proof is provided in Appendix C.

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 those true nulls, thus preventing any false rejections. Conversely, if pre-testing rejects some true nulls, the chance of over-rejection in the second-step inference is minimized—less than α, as ensured by the design of Algorithm 3. Therefore, the probability of a false rejection remains below the significance level α, addressing post-model selection inference concerns.

Algorithm in Blattman et al. (2021) for the Pure Control Group. Other approaches, such as the one employed by Blattman et al. (2021), often involve using a pre-specified rule, starting with the null hypothesis $H_0^{\epsilon_{K-1}}$ and collapsing any unrejected nulls into a single control condition. However, this method may encounter issues with post-model selection inference, leading to over-rejection under the null.

In Blattman et al. (2021), they implement Algorithm B.1 with $K = 2$ and $(\epsilon_0, \epsilon_1, \epsilon_2) = (0, 250m, 500m)$.

Algorithm B.1 Procedure for Pure Control Group

Inputs : Test statistic $T = T(Y(d), d)$, observed assignment D^{obs}, observed outcome Y^{obs}, and treatment assignment mechanism $P(D)$.

Set : $\hat{K} = K$.

for $k = K - 1$ *to* 0 **do**

  Test $H_0^{\epsilon_k}$ using the PNRT procedure, and collect $pval^k$.

  If $pval^k \leq \alpha$, reject $H_0^{\epsilon_k}$ and terminate.

  If $pval^k > \alpha$, set $\hat{K} = k$.

end

Output: Set pure control group with $\epsilon_c = \epsilon_{\hat{K}}$.

Following Algorithm B.1, the procedure involves two steps. First, 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

$\epsilon_c = 500$. 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 rejected, we report no significant interference.

As illustrated in Algorithm B.1, this procedure does not account for multiple hypothesis testing, which can lead to over-rejection of the partial null hypothesis and 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 $\tilde{k} = 0$, $p(\hat{K} \geq 1) > \alpha$. In extreme cases, $P(pval(D^{obs}) \leq \alpha | \hat{K} \geq \tilde{k} + 1)$ can be close to 1, leading to over-rejection using this pre-selection procedure.

Algorithm for Simulation Exercise in Section 5.1. The algorithm for the simulation exercise in Section 5.1 is outlined in Algorithm B.2.

Algorithm B.2 Simulation Study Procedure

Inputs : 5,000 randomly chosen assignments as the potential assignments set, $\mathbb{D}_S$.
The biclique decomposition of $\mathbb{D}_S$ from Puelz et al. (2021).
Set : Spillover effect τ 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 $R = 1,000$.
  Average the number of rejections to obtain the power for that fixed τ.
end
Output: Power plot of each algorithm.

Appendix C Proof of the Theorems

Proof of Proposition 1. For any $d, d' \in \{0,1\}^N$. Consider any $i \in \mathbb{I}(d) \cap \mathbb{I}(d')$. By Definition 4 of imputable units, under $H_0^{\epsilon_s}$, we have $Y_i(d) = Y_i(d')$. Hence, by Definition 6 of pairwise imputable statistics, $T(Y_{\mathbb{I}(d)}(d), d') = T(Y_{\mathbb{I}(d)}(d'), d')$. □

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 $P(pval^{pair}(D^{obs}) \leq \alpha/2) = \sum_{D^{obs} \in \mathbb{D}} P(D^{obs}) = w$. Since $E_P(\phi(D^{obs})) = P(pval^{pair}(D^{obs}) \leq \alpha/2)$, to prove the theorem, we want to show $w < \alpha$.

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 Definition 9 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

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

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

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

Hence, $w^2/2 < w\alpha/2$, implying $w < \alpha$. As previously mentioned, using $1/2$ to discount the number of equalities does not affect the test's validity 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 $\tilde{T}(D)$, induced by $P(D)$. Denote its cumulative distribution function by F_U. We then 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 $H_0^{\epsilon_s}$. Thus, for any $\alpha \in [0,1]$,

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

□

Proof of Proposition 2. By Definition 3, 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 2,

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

Thus, for any $k \geq \bar{K}$ and any $d, d' \in \mathcal{D}_i(\epsilon_k) \subseteq \mathcal{D}_i(\epsilon_{\bar{K}})$, it follows that $Y_i(d) = Y_i(d')$ for all $i \in \{1, \ldots, N\}$. By Definition 3, $H_0^{\epsilon_k}$ is true for any $k \geq \bar{K}$. □

Proof of Theorem 3. Without loss of generality, consider the 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 $k \geq \bar{K}$ such that $H_0^{\epsilon_k}$ is rejected, it must be the case that $H_0^{\epsilon_{\bar{K}}}$ is rejected. Thus, by Definition 11,

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

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

Proof of Theorem B.1. 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 fail to reject any $H_0^{\epsilon_k}$ with $k \geq \hat{K}$. Thus, 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, $P(pval(D^{obs}) \leq \alpha | \hat{K} < \tilde{k} + 1) = 0$. Therefore,

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

□

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 regressions. As illustrated in the main text, this method involves conducting the 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 residual outcomes. 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. I 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 FRTs is proposed by Rosenbaum (2020), with detailed discussion in Sections 7 and 9.2 of Basse and Feller (2018).

The third approach is PNRT using pairwise residuals. In this method, for each pair of treatment assignments (D^{obs}, D), I 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.

D.1 Investigation on the Power of Incorporating Covariates

In this investigation, we extend the potential outcomes described in Table 9 by incorporating two covariates, X_1 and X_2. The new control potential outcomes, $Y_i^C(\text{new})$, are simulated based on the original control outcomes Y_i^C from Table 9 as follows:

$$Y_i^C(\text{new}) = 2 + 0.5 \times X_1 + 0.3 \times X_2 + Y_i^C,$$

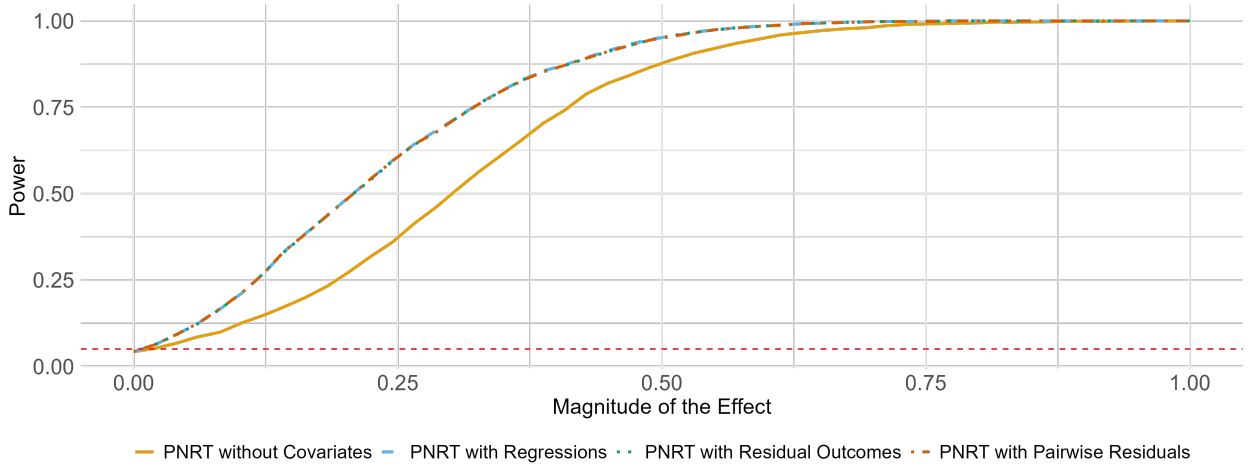
where X_1 is a binary covariate drawn from a Bernoulli distribution with parameter 0.5, and X_2 is a continuous covariate drawn from a standard normal distribution:

$$X_1 \sim \text{Bernoulli}(0.5), \quad X_2 \sim \mathcal{N}(0,1).$$

It is important to note that only the control potential outcomes, Y_i^C, are modified by these covariates. The remaining potential outcomes for treated units follow the same functional relationships as described in Table 9. By introducing X_1 and X_2, the control potential outcomes become more variable, reflecting the added noise from the covariates.

Next, I apply the three methods introduced earlier—PNRT with regressions, PNRT with residual outcomes, and PNRT using pairwise residuals—to construct the power curves. The simulation procedure remains consistent with that described in the main text, focusing on displacement effects and one-sided tests using non-absolute coefficients.

Figure D.1: Power Comparison of Testing Methods for Different Covariate Adjustments



Notes: The red line indicates the size level $\alpha = 0.05$. Power is based on the pairwise comparison-based PNRT with rejection at level α. I consider 50 equally spaced values of τ between 0 and 1, conducting 2,000 simulations for each τ to compute the average rejection rate for each method.

Figure D.1 illustrates the power gains achieved by incorporating covariate information. While all methods involving covariate adjustments demonstrate similar power performance, leveraging covariates consistently results in higher power. When $\tau = 0$, the rejection rates for all methods align with the nominal size of the test. As τ increases, the power also increases. For instance, when $\tau \approx 0.25$, the power of the test with covariate adjustment reaches approximately 0.65, compared to less than 0.4 for the test without covariate adjustments. Therefore, in practice, researchers should select the method that best suits their specific

context and data.

D.2 Robustness of Results to Adjustment Methods

The application of the above methods yields the results presented in Table D.1 below:

Table D.1: p-Values for the 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

Notes: The table shows p-values of pairwise comparison-based PNRT across different methods. Reg (WLS) is PNRT with regression, using the coefficient from the covariates-included regression with inverse propensity weighting as the test statistic. Reg (OLS) is the PNRT with regression, using the coefficient from the covariates-included regression without weighting as the test statistic. Pair residuals are 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 is PNRT with the 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.

As shown in Table D.1, 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 existence 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 non-zero spillover effect, leading to increased p-values under the same partial null hypothesis.

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, PNRTs 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 PNRTs differs from those in CRTs and FRTs, making classical approaches for weak nulls potentially inapplicable. Further investigation into these differences would be of interest to future research.

Appendix E Discussion on Some Extreme Cases

Emptiness of Imputable Units Set. The emptiness of the imputable units set depends on three factors: the distance being tested, the network structure, and the randomization design.

First, the target distance interacts with the network structure. If the distance $\epsilon_s > \max_{i,j} G_{i,j}$, meaning it exceeds any existing distance in the network, then there will be no units in the imputable units set. In this case, additional data may be required to gain sufficient power for the test, or the target distance ϵ_s could be reduced. For clarity in the following discussion, we focus on the case where $\epsilon_s = 0$.

Second, with $\epsilon_s = 0$, if all units in the sample are treated, the imputable units set will still be empty. To detect the existence of interference, a sufficient number of units beyond our target distance across various treatment assignments is necessary to achieve reasonable power.

Cases with an Undefined Comparison Group. The distance being tested, network structure, and randomization design also influence whether one of the comparison groups is undefined. To highlight the core intuitions, we focus on the case where $\epsilon_s = 0$, implying that we are testing for the existence of interference and not all units are treated. Thus, some untreated units remain to conduct the test.

A general example is a network of couples, where exactly one unit in each pair is treated. In the example from the main text, with treatment assignments rotating across couples, the neighborhood units set may be empty. In practice, the test statistic must then assume a very high value for implementation.

More generally, for each assignment d and given ϵ_c, let $\|\{i : d_i = 1\}\|$ denote the number of treated units, $\|\{i : d \in \mathcal{D}_i(0)/\mathcal{D}_i(\epsilon_c)\}\|$ the number of units in the neighborhood set, and $\|\{i : d \in \mathcal{D}_i(\epsilon_c)\}\|$ the number in the control set. Whenever the number of non-imputable units is equal to or exceeds the number of neighborhood units, there may exist a pair of assignments (D^{obs}, D) such that the neighborhood units set is empty. This principle also applies to the control units set. Therefore, to ensure that both neighborhood and control sets are defined, we impose Assumption E1.

Assumption E1 (Regularization when $\epsilon_s = 0$)**.**

$$\min\left\{\min_{d\in\{0,1\}^N}\|\{i : d \in \mathcal{D}_i(0)/\mathcal{D}_i(\epsilon_c)\}\|, \min_{d\in\{0,1\}^N}\|\{i : d \in \mathcal{D}_i(\epsilon_c)\}\|\right\} > \max_{d\in\{0,1\}^N}\|\{i : d_i = 1\}\|$$

It is worth noting that $\|\{i : d_i = 1\}\| = N - \|\mathbb{I}(d)\|$ when $\epsilon_s = 0$. Therefore, Assumption E1 implies that the groups of interest occupy a large proportion of the population across all treatment assignments. This condition depends on ϵ_c, the network structure, and the experimental design. With Assumption E1, we ensure all comparison groups remain non-empty across different potential assignments.

Proposition E3. *Suppose Assumption E1 holds. For any $D^{obs} \in \{0,1\}^N$, the pairwise imputable statistic $T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), D) \neq \infty$ across different D.*

Proof of Proposition E3. We proceed by contradiction. Assume there exists a $d \in \{0,1\}^N$ such that $T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), d) = \infty$. Without loss of generality, suppose for any $i \in \mathbb{I}(D^{obs})$, $d \notin \mathcal{D}_i(\epsilon_c)$. Then, for any unit $j \in \{i : d \in \mathcal{D}_i(\epsilon_c)\}$, it must be the case that $j \notin \mathbb{I}(D^{obs})$, and hence $j \in \{i : D_i^{obs} = 1\}$.

Thus, we have

$$\|\{i : d \in \mathcal{D}_i(\epsilon_c)\}\| \leq \|\{i : D_i^{obs} = 1\}\|$$

However, we have

$$\|\{i : d \in \mathcal{D}_i(\epsilon_c)\}\| \geq \min_{d\in\{0,1\}^N}\|\{i : d \in \mathcal{D}_i(\epsilon_c)\}\|$$

and

$$\|\{i : D_i^{obs} = 1\}\| \leq \max_{d \in \{0,1\}^N} \|\{i : d_i = 1\}\|.$$

This implies that

$$\min_{d \in \{0,1\}^N} \|\{i : d \in \mathcal{D}_i(\epsilon_c)\}\| \leq \max_{d \in \{0,1\}^N} \|\{i : d_i = 1\}\|,$$

which contradicts Assumption E1.

□