A Graph-Theoretic Approach to Randomization Tests of Causal Effects Under General Interference*

David Puelz† Guillaume Basse‡ Avi Feller§ Panos Toulis†

Abstract

Interference occurs when a unit's outcome depends on the treatment assignments of other units. For example, intensive policing on one street could have a spillover effect on neighboring streets. Classical randomization tests typically break down in this setting because many null hypotheses of interest are no longer sharp under interference. A promising alternative is to instead construct a conditional randomization tests on a subset of units and assignments for which a given null hypothesis is sharp. Finding these subsets is challenging, however, and existing methods either have low power or are limited to special cases. In this paper, we propose valid, powerful, and easy-to-implement randomization tests for a general class of null hypotheses that allow for arbitrary interference between units. Our key idea is to represent the hypothesis of interest as a bipartite graph between units and assignments, and to find a clique of this graph. Importantly, the null hypothesis is sharp for the units and assignments in this clique, enabling randomization-based tests. We can apply off-the-shelf graph clustering methods to find such cliques efficiently and at scale. We illustrate this approach in settings with clustered interference and show advantages over methods designed specifically for that setting. We then apply our method to a large-scale policing experiment in Medellín, Colombia, where interference has a spatial structure.

Keywords: randomization test, interference, causal inference, networks, clique.

*email: David.Puelz@chicagobooth.edu. We thank Peng Ding, Sam Pimental, and Santiago Tobón as well as seminar participants at the University of Chicago, Polmeth, and Advances with Field Experiments conference for helpful comments and discussion. AF gratefully acknowledges support from a National Academy of Education/Spencer Foundation postdoctoral fellowship. PT is grateful for the John E. Jeuck Fellowship at Booth.

†The University of Chicago, Booth School of Business

‡Stanford University

§The University of California, Berkeley

1. Introduction

The assumption of "no interference" between units (Cox, 1958) underlies most classical approaches to causal inference. The key premise is that a unit's treatment does not affect another unit's outcome, so that each unit's outcome depends only on its own treatment status. This is implausible in many settings, however, as it precludes peer effects, treatment spillovers, and other forms of treatment interference (Halloran and Hudgens, 2016).

Classical approaches often break down under interference. The canonical Fisher Randomization Test (FRT), for example, is valid for testing the global sharp null hypothesis of no treatment effect but fails when testing non-sharp null hypotheses, such as tests of treatment spillovers. Several recent proposals address this issue by restricting the randomization test to a subset of units, often called focal units, and a subset of assignments (Aronow, 2012; Athey et al., 2018; Basse et al., 2019b). The central idea is that, conditional on these subsets, the specified null hypothesis is sharp for every focal unit, and thus, the *conditional* randomization test is valid. These randomization-based approaches have many advantages over model-based alternatives (Manski, 2013; Graham, 2008; Jackson, 2010; Graham and Hahn, 2005; Brock and Durlauf, 2001; Blume et al., 2015; Toulis and Kao, 2013; Bowers et al., 2013), especially because they make minimal assumptions, and are thus more robust. However, the randomization-based methods can be difficult to apply to new settings (e.g., Basse et al., 2019b) or restrict the type of information used for constructing the test (Athey et al., 2018), limiting power and the widespread adoption of these methods.

In this paper, we propose a general procedure for identifying subsets of units and assignments for which the null hypothesis of interest is sharp, and use it to develop a method for randomization tests under arbitrary interference. Our key methodological contribution is the *null-exposure graph*. This is a bipartite graph with units and assignments as the sets of nodes, and an edge between any unit-assignment pair if a null exposure, i.e., a treatment exposure specified in the null hypothesis, is observed for that unit and assignment in the pair. Thus, a (bi)clique of this null-exposure graph is a subset of units and assignments for which the null hypothesis is sharp and a valid randomization test is possible. Importantly, we can apply advances in graph algorithms, especially biclustering methods (Zhang et al., 2014; Prelić et al., 2006), to find the necessary clique efficiently and at scale.

Our proposed method offers three main benefits over existing approaches. First, it allows conditioning on the observed assignment, which can increase testing power over methods that suggest random conditioning (Athey et al., 2018; Aronow, 2012). Second, since a null hypothesis uniquely determines a null-exposure graph, our method is constructive and defines a concrete randomization test under general forms of interference. This is an improvement over methods that are tailored to specific patterns of interference, but do not provide guidance on how to properly condition a randomization test under arbitrary interference (Basse et al., 2019b). Finally, our method translates questions of computation and statistical power into properties and operations on the null-exposure graph, separating these considerations from test validity. Our approach is therefore modular, and can benefit from separate advances in graph algorithms for clique computations.

To illustrate our method, we consider two natural structures of interference, namely clustered interference and spatial interference. In clustered interference, units can be separated into well-defined clusters, such as households or classrooms, where we assume that units interact within clusters but not between clusters. Our motivating example here is a two-stage randomized trial of a student attendance intervention in which households are assigned to treatment or control and then, within each treated household, one student is randomly treated; see Basse and Feller (2018). In spatial interference, we assume that interactions pass through "neighboring" units, but without the simpler structure of clustered interference. Here, we focus on re-analyzing a large-scale experiment in Medellín, Colombia studying the impact of "hotspot policing" on crime (Collazos et al., 2019). Our analysis of spatial interference considers hypotheses on any specified spillovers, which differs from other design-based methods that consider a marginal spillover effect over the design (Aronow et al., 2019).

Our paper is structured as follows. In Section 2, we introduce the problem setup and all necessary notation. In Section 3, we present our methodology, comprised of the null-exposure graph (Section 3.1) and clique decompositions of the graph (Section 3.2). Section 4 presents the proposed randomization test. We then illustrate our method in two applications. Section 5 considers settings with clustered interference, and Section 6 considers settings with spatial interference, specifically in the context of a large-scale policing experiment in Medellín, Colombia.

2. Overview of causal inference under interference and problem setup

2.1. Setup and notation

Consider a finite population of N units indexed by $i = 1, \ldots, N$. Let Z_i denote unit i's treatment, which we assume to be binary without loss of generality. Let $Z = (Z_1, Z_2, \ldots, Z_N) \in \{0,1\}^N$ denote the population treatment assignment, and $P(Z)$ its distribution according to the design. Let $Y_i(z) \in \mathbb{R}$ denote the potential outcome of unit i under population assignment $z \in \{0,1\}^N$. For the observed quantities we use the modifier "obs" as a superscript. Thus, $Z^{\text{obs}} \in \{0,1\}^N$ is the observed population treatment, and $Y^{\text{obs}} = (Y_1(Z^{\text{obs}}), \ldots, Y_N(Z^{\text{obs}})) \in \mathbb{R}^N$ is the vector of observed outcomes. In the randomization framework, Z^{obs} is random according to the design, $Z^{\text{obs}} \sim P(Z^{\text{obs}})$, whereas the potential outcomes are fixed. Let $\mathbb{U} = \{1, \ldots, N\}$ denote the set of all units, and $\mathbb{Z} = \{z \in \{0,1\}^N : P(z) > 0\}$ denote the set of all population treatment assignments supported by the design.

The main challenge is that estimation is infeasible without additional restrictions on the potential outcomes; unrestricted, each unit has 2^N possible potential outcomes. At one extreme, the common *no interference* assumption states that the outcome for each unit depends only on its own treatment assignment, so unit i has only two potential outcomes. When interference cannot be reasonably assumed away, one promising strategy is to define a low-dimensional summary of Z, known as an *exposure*; see Aronow et al. (2017). In particular, we assume that there is a finite set of possible treatment exposures, $\mathbb{F} = \{\mathsf{a}, \mathsf{b}, ...\}$, and exposure mapping functions, $f_i : \mathbb{Z} \to \mathbb{F}$, for each unit $i \in \mathbb{U}$. The definition of f_i is application-specific, as it needs to consider the possible structure of interference. To make this concrete, we briefly introduce two examples.

Example 1 (Clustered interference)**.** *Following the setup in Basse et al. (2019b), each unit belongs to a fixed cluster, such as individuals within households or students within classrooms. The key assumption is that units interact within each cluster, but not between clusters, also known as partial interference (Sobel, 2006). Specifically, we consider a randomized trial in which clusters are assigned to treatment or control, and, within treated clusters, one unit is randomly assigned to treatment. Here, the exposure function may be defined as a pair,* $f_i(Z) = (W_i, Z_i)$*, where* Z_i *denotes unit* i*'s treatment status, and* $W_i = \sum_{j\in[i]} Z_j$ *denotes the treatment status of unit* i*'s cluster defined as* $[i]$*. We explore this setting further in Section 5.*

Example 2 (Spatial interference)**.** *Units interact geographically with one another, like street segments in a city. A distance metric,* $d(i, j)$*, that denotes the distance between units* i *and* j*, affects this interaction. The goal is usually to estimate whether there are spillovers at certain distances. For example, let* $g_{ij} = \mathbb{I}\{d(i, j) < r\}$ *denote whether* i *and* j *are within distance* r *from each other. We define the exposure function as* $f_i(Z) = (W_i, Z_i)$*, where* $W_i = \mathbb{I}\{\sum_{j\neq i} g_{ij}Z_j > 0\}$ *indicates whether* i *receives short-range spillovers from all other units. We explore this setting further in Section 6.*

A common assumption is that the exposure functions are properly specified for potential outcomes in the sense that $f_i(Z) = f_i(Z')$ implies that $Y_i(Z) = Y_i(Z')$, for all units i and assignments Z, Z'. This ensures that the potential outcome function is well-defined on the set of exposures. See Aronow et al. (2017), Basse et al. (2019a), and Basse et al. (2019b) for a discussion of this assumption and its effect on testing interpretation. See also Sävje et al. (2017) for approaches when f_i is unknown.

2.2. Hypotheses on exposure mappings

Our primary goal is to test null hypotheses based on contrasts between exposure levels:

$$H_0^{\mathsf{a},\mathsf{b}} : Y_i(Z) = Y_i(Z'), \text{ for all } i = 1, \ldots, N, \text{ and any } Z, Z' \text{ such that } f_i(Z), f_i(Z') \in \{\mathsf{a}, \mathsf{b}\}, \tag{1}$$

where $\mathsf{a}, \mathsf{b} \in \mathbb{F}$ are two distinct treatment exposures. In words, $H_0^{\mathsf{a},\mathsf{b}}$ states that for each unit, the outcomes under exposure a and exposure b are identical. The null, however, does not constrain the outcomes under other exposures. For example, if unit i is exposed to exposure c, then we have no information about unit i's outcome under $H_0^{\mathsf{a},\mathsf{b}}$; in this case, we say that unit i's outcome is *not imputable* under $H_0^{\mathsf{a},\mathsf{b}}$.

The formulation in Equation (1) is quite general and can express hypotheses in many interference settings where the focus is on specific levels of treatment exposure (Toulis and Kao, 2013; Bowers et al., 2013; Rosenbaum, 2007; Aronow, 2012; Basse et al., 2019a,b; Athey et al., 2018). This is not universal, however and does not cover all possible hypotheses under interference (e.g., Athey et al., 2018, hypotheses 2 & 3). We return to this issue in Section 7.2.

Example 1 (Clustered interference (cont.))**.** *In this setting, we consider testing* $H_0^{\mathsf{a},\mathsf{b}}$ *with* $\mathsf{a} = (0, 0)$ *and* $\mathsf{b} = (1, 0)$*; that is, we test whether there is a spillover effect on control units from a treated unit in the same cluster.*

Example 2 (Spatial interference (cont.)). *In this setting, we consider testing* $H_0^{\mathsf{a},\mathsf{b}}$ *with* $\mathsf{a} = (0,0)$ *and* $\mathsf{b} = (1,0)$ *for some value of the distance threshold,* r*. That is, we test whether there is a spillover effect on an untreated unit from having one or more treated units within a distance* r*. For example,* $r = 125$ *meters in the crime application of Section 6.*

The main challenge in testing the null hypothesis of Equation (1) is that the treatment exposures, a and b, cannot be independently manipulated because individual unit exposures are functions of the treatment vector, Z, which precludes a simple randomization test. Without such manipulation, however, it is not generally possible to impute unit outcomes in the randomization test since $H_0^{\mathsf{a},\mathsf{b}}$ is not sharp.

2.3. Conditional randomization tests under interference: A review

We briefly review the general framework proposed by Basse et al. (2019b) for valid randomization tests under interference. This framework builds on the key insight first formulated by Aronow (2012) and developed by Athey et al. (2018) that, although the null hypothesis $H_0^{\mathsf{a},\mathsf{b}}$ is not sharp in general, it can be "made sharp" if we restrict our attention to a well chosen subset of units (U) and subset of assignments ($\mathcal{Z}$). Basse et al. (2019b) formalized the idea as sampling a *conditioning event*, $C = (U, \mathcal{Z})$, from a carefully constructed distribution $P(C \mid Z^{\text{obs}})$, called a *conditioning mechanism*, and then running a test conditional on C.

Theorem 1 (Basse et al. (2019b)). *Let* H_0 *be a null hypothesis. Let* $T_C(z, y)$ *be a test statistic indexed by* $C = (U, \mathcal{Z})$*, defined only on the units in* U*, and such that under* H_0*, it holds that* $T_C(z, Y(z)) = T_C(z, Y(z'))$ *for all* $z, z' \in \mathbb{Z}$*, and* C *for which* $P(C \mid z) > 0$ *and* $P(C \mid z') > 0$*. The p-value obtained from the following procedure:*

1. *Draw* $Z^{\text{obs}} \sim P(Z^{\text{obs}})$*, observe* $Y^{\text{obs}} = Y(Z^{\text{obs}})$*;*
2. *Draw* $C \sim P(C \mid Z^{\text{obs}})$ *and compute* $T^{\text{obs}} = T_C(Z^{\text{obs}}, Y^{\text{obs}})$*;*
3. *Compute p-value* $= E_{Z \sim P(Z|C)}\left[\mathbb{I}\{T_C(Z, Y^{\text{obs}}) > T^{\text{obs}}\} \mid C\right]$*, where the expectation is with respect to the correct randomization distribution,* $P(Z \mid C) \propto P(C \mid Z)P(Z)$*,*

is valid conditionally and marginally.

We can see from Theorem 1 that there are two main challenges in constructing a conditioning mechanism that leads to valid conditional randomization tests. First, the test statistic T_C should be imputable under the null hypothesis H_0. In words, this means that based only on the observed value of the outcomes, Y^{obs}, we can compute the null distribution of $T_C(z, Y(z)) = T(z, Y^{\text{obs}})$ that is induced by the randomization distribution $P(Z \mid C)$. Second, we must be able to draw samples from this distribution, given by its conditional-marginal decomposition $P(Z \mid C) \propto P(C \mid Z)P(Z)$ in the third step. Ensuring that this distribution is computationally tractable can be challenging (Basse et al., 2019a,b).

We can also describe the approaches of Basse et al. (2019b), Athey et al. (2018) and Aronow (2012) within this framework, each corresponding to different choices of the conditioning mechanism. In particular, Basse et al. (2019b) propose a conditioning mechanism under clustered interference, such that the implied randomization distribution $P(Z \mid C)$ leads to a permutation test. However, their approach does not readily generalize to other settings, such as spatial interference. The methods of Athey et al. (2018) and Aronow (2012) correspond to conditioning mechanisms of the form $P(C \mid Z) = P(C)$, where conditioning is either random, or guided by auxiliary information, such as a social network between units. In contrast to the approach of Basse et al. (2019b), these methods can be applied in general interference settings. However, they are usually underpowered because they do not use the observed assignment to do the conditioning; as an extreme example, these approaches cannot test $H_0^{\mathsf{a},\mathsf{b}}$ if C has no units exposed to a or b, as the exposure information is contained only in the observed treatment assignment. We discuss this issue further in Section 4.2.

In this paper, we propose a method that is both general and powerful. In the next sections, we develop the key concepts of our approach, following the framework of conditioning mechanisms presented here. Our proposed randomization test for $H_0^{\mathsf{a},\mathsf{b}}$ is presented later in Section 4, and in Section 4.2 we follow up on this discussion of related methods and describe the benefits of our approach.

3. The null-exposure graph and cliques

We now introduce some preliminary concepts underlying our method for testing $H_0^{\mathsf{a},\mathsf{b}}$ in Equation (1). The key idea is to represent the imputability of outcomes under the null hypothesis through a graph between units and assignments, which we term *the null-exposure graph*. The conditioning event C will then be taken to be a clique in that graph, and the conditioning mechanism $P(C \mid Z)$ simply determines the clique that contains Z. This transforms the analytical task of defining $P(C \mid Z)$ analytically into a computational task: decomposing the graph into dense and balanced cliques.

3.1. The null-exposure graph

The first component of our method is a graph that encodes the units' treatment exposures under different treatment assignments, which we call the null-exposure graph. This graph is a key conceptual contribution of our paper: as we will show in the next section, it determines the appropriate conditioning for the randomization test of $H_0^{\mathsf{a},\mathsf{b}}$.

Definition 1 (Null-exposure graph). *Let $\mathbb{U} = \{1, \ldots, N\}$ and $\mathbb{Z} = \{z_1, \ldots, z_J\}$ denote the sets of units and assignments, respectively. Define the vertex set as $V = \mathbb{U} \cup \mathbb{Z}$, and the edge set as*

$$E = \{(i, z) \in \mathbb{U} \times \mathbb{Z} : f_i(z) \in \{\mathsf{a}, \mathsf{b}\}\}. \tag{2}$$

That is, an edge between unit i and assignment z exists if and only if i is exposed to either a or b under assignment z. There are no edges between units or between assignments. Then, $G_f^{\mathsf{a},\mathsf{b}} = (V, E)$ is the null-exposure graph of $H_0^{\mathsf{a},\mathsf{b}}$ with respect to exposure mapping f.

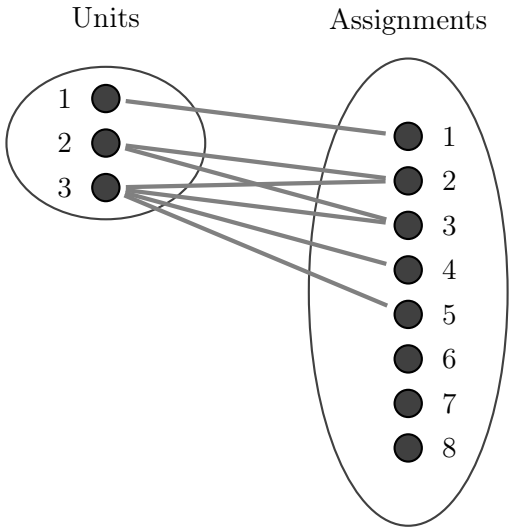

Figure 1. Depiction of null-exposure graph of some $H_0^{\mathsf{a},\mathsf{b}}$. The left nodes represent the experimental units, and the right nodes represent the population treatment assignments. The graph is *bipartite* because no units and assignments are connected with other like nodes.

The key property of a null-exposure graph $G_f^{\mathsf{a},\mathsf{b}}$ is that if some unit i is exposed to either a or b, i.e., $f_i(Z^{\text{obs}}) \in \{\mathsf{a}, \mathsf{b}\}$, then the outcomes of unit i can be imputed for any assignment connected to i in $G_f^{\mathsf{a},\mathsf{b}}$ assuming $H_0^{\mathsf{a},\mathsf{b}}$ is true. In particular, a test statistic defined only on this subset of units and assignments is imputable under $H_0^{\mathsf{a},\mathsf{b}}$.

An illustration of an example null-exposure graph is shown in Figure 1, where there are three units and eight assignments. Unit 1 is exposed to either a or b only under assignment 1; unit 2 is exposed to a or b under assignments 2 and 3; and unit 3 is exposed to a or b under assignments 2, 3, 4, and 5. If we restrict attention to units 2 and 3 and assignments 2 and 3, the missing potential outcomes are imputable under $H_0^{\mathsf{a},\mathsf{b}}$. Specifically, under the null hypothesis, we are able to infer all potential outcomes of units within a clique of the null-exposure graph. However, this no longer holds if we also include either unit 1 or assignment 1.

3.2. Cliques and clique decompositions

The previous section described the null-exposure graph and its complete subgraphs, or cliques, which are a crucial piece of our testing procedure. The "completeness" of these cliques means that all units are connected to all assignments. As mentioned earlier, this implies that for the subset of units and assignments that comprise the clique, we can impute all potential outcomes for exposures a and b under $H_0^{\mathsf{a},\mathsf{b}}$. In the paragraphs below, we give definitions for these important objects as well as discuss how to partition the assignments using the structure of the null-exposure graph. This decomposition will be crucial for the proposed randomization test. We describe algorithms for finding cliques in Section 4.3.

Definition 2 (Clique). *A clique in the null-exposure graph, $G_f^{\mathsf{a},\mathsf{b}} = (V, E)$, where $V = \mathbb{U} \cup \mathbb{Z}$ and E is defined in Equation (2), is a set-pair $C = (U, \mathcal{Z})$, with $U \subseteq \mathbb{U}$ and $\mathcal{Z} \subseteq \mathbb{Z}$, such that $(i, f_i(z)) \in E$, for every $i \in U$ and every $z \in \mathcal{Z}$.*

As an example, $C = (\{2, 3\}, \{z_2, z_3\})$ is a clique in Figure 1 since it is a fully connected subgraph. We now formalize the intuition that cliques in the null-exposure graph allow imputation of the missing potential outcomes.

Proposition 1. *Consider a null-exposure graph, $G_f^{\mathsf{a},\mathsf{b}}$, with some clique $C = (U, \mathcal{Z})$. If $Z^{\text{obs}} \in \mathcal{Z}$, then $Y_i(z) = Y_i(Z^{\text{obs}})$ under $H_0^{\mathsf{a},\mathsf{b}}$, for all $i \in U$ and all $z \in \mathcal{Z}$.*

Proof. For any unit i it holds $f_i(Z^{\text{obs}}) \in \{\mathsf{a}, \mathsf{b}\}$ since $Z^{\text{obs}} \in \mathcal{Z}$ is in the clique by assumption. Fix any unit $i \in U$ and assignment $z \in \mathcal{Z}$. Then, there is an edge between i and z by definition of the clique. This implies that $f_i(z) \in \{\mathsf{a}, \mathsf{b}\}$ as well, by construction of the null-exposure graph in Definition 1. Hence, both $f_i(Z^{\text{obs}})$ and $f_i(z) \in \{\mathsf{a}, \mathsf{b}\}$, and so $Y_i(z) = Y_i(Z^{\text{obs}})$ under $H_0^{\mathsf{a},\mathsf{b}}$ in Equation (1). □

Proposition 1 shows that we can impute the missing potential outcomes in a clique that contains the observed treatment assignment Z^{obs}. Finally, we need the additional technical constraint that we choose a unique clique containing Z^{obs}.

Definition 3 (Clique Decomposition). *A clique decomposition, $\mathcal{C} = \{C_1, \ldots, C_K\}$, of the null-exposure graph in Definition 1 is a finite set of cliques, $C_k = (U_k, \mathcal{Z}_k)$, $k = 1, \ldots, K$, such that*

$$\bigcup_k \mathcal{Z}_k = \mathbb{Z}, \text{ and } \mathcal{Z}_k \bigcap \mathcal{Z}_{k'} = \emptyset, \text{ for any } k \neq k'.$$

A clique decomposition defines a partition of $\mathbb{Z}$, the set of possible treatment assignments. This guarantees that there will be a unique clique C that will contain Z^{obs}. On the other hand, U_k need not form a partition of the set of units, $\mathbb{U}$. This is crucial because partitioning both $\mathbb{U}$ and $\mathbb{Z}$ may not be possible in general, or could lead to low-powered or even empty randomization tests.

4. Clique-based randomization tests

4.1. Main method and test validity

We can now describe our proposed conditional FRT for $H_0^{\mathsf{a},\mathsf{b}}$ in Equation (1), which is the key methodological contribution of this paper. Throughout, let $\mathcal{C}$ be a clique decomposition of the null-exposure graph $G_f^{\mathsf{a},\mathsf{b}}$, and for $C = (U, \mathcal{Z}) \in \mathcal{C}$, let T_C be a test statistic defined only on units in U and assignments in $\mathcal{Z}$. Consider the following procedure:

Procedure 1. *For observed assignment $Z^{obs} \sim P(Z^{obs})$:*

1. *Find the unique clique, $C = (U, \mathcal{Z}) \in \mathcal{C}$, such that $Z^{\text{obs}} \in \mathcal{Z}$.*
2. *Calculate the observed value of the test statistic, $T^{\text{obs}} = T_C(Z^{\text{obs}}, Y^{\text{obs}})$.*

3. *Define the randomization distribution:* $r(Z) \propto \mathbb{I}\{Z \in \mathcal{Z}\} \cdot P(Z)$.

4. *Define the randomization p-value as follows:*

$$\text{pval}(Z^{\text{obs}} \mid \mathcal{C}) = \mathbb{E}_{Z \sim r}\left[\mathbb{I}\{T_C(Z, Y^{\text{obs}}) > T^{\text{obs}}\}\right]. \tag{3}$$

The following theorem shows that this procedure is valid; the proof is in the Appendix.

Theorem 2. *Consider the null hypothesis,* $H_0^{\mathsf{a},\mathsf{b}}$ *in Equation* (1)*. Construct the corresponding null-exposure graph,* $G_f^{\mathsf{a},\mathsf{b}}$*, and compute a clique decomposition* $\mathcal{C}$*. Let* $C \in \mathcal{C}$ *be the unique clique such that* $Z^{\text{obs}} \in C$*. Then, the randomization test described in Procedure 1 is valid at any level, i.e., the p-value defined in Equation* (3) *satisfies:*

$$E\left(\mathbb{I}\{\text{pval}(Z^{\text{obs}} \mid \mathcal{C}) \leq \alpha\} \mid \mathcal{C}, H_0^{\mathsf{a},\mathsf{b}}\right) \leq \alpha,$$

where the expectation is with respect to the design, $P(Z^{\text{obs}})$*.*

This theorem follows from Theorem 1 by recognizing that Procedure 1 describes a conditional randomization test in which the conditioning event is a clique $C \in \mathcal{C}$, and the conditioning mechanism is $P(C = (U, \mathcal{Z}) \mid Z) = \mathbb{I}\{C \in \mathcal{C}\}\mathbb{I}\{Z \in \mathcal{Z}\}$. The proof first verifies that any test statistic T_C defined only on C is imputable: this follows from the construction of the clique and Proposition 1. It then shows that the randomization distribution $r(Z)$ defined in Step 3 of the procedure is the correct conditional distribution $P(Z \mid C)$ implied by the design and the conditioning mechanism.

Remark 4.1. The computational tractability of the randomization distribution, $P(Z \mid C) \propto r(Z) \propto \mathbb{I}\{Z \in \mathcal{Z}\} \cdot P(Z)$ in Step 3 of Procedure 1 is immediate provided that we can compute $P(Z)$ and enumerate the assignments $\mathcal{Z}$ in any clique $C \in \mathcal{C}$. This last condition may be prohibitive if the support of the design $P(Z)$ is too large. We show in Section 7.1 how a simple modification of our method can address this issue without compromising validity.

Remark 4.2. While Procedure 1 automates the construction of a conditioning mechanism, it still allows flexibility in the choice of the test statistic. The simplest choice is for T_C to denote the difference-in-means between outcomes of units in C exposed to a and b. However, we may improve power by finding test statistics that explain more variation. For example, we could use a network regression model with spillovers (Jackson, 2010; Graham and Hahn, 2005; Brock and Durlauf, 2001; Blume et al., 2015; Toulis and Kao, 2013) to derive a test statistic. Such models may explain the data better than standard linear regression, leading to sharper inference. See also Athey et al. (2018, Section 5.3) for an excellent related discussion.

4.2. Comparison to related work

In Section 2.3, we discussed how our method, along with those of Aronow (2012) and Athey et al. (2018), could be described within the general framework of Basse et al. (2019b). In fact, we can also

describe these approaches using the framework in this paper, in which each method corresponds to a different approach for selecting cliques from the null-exposure graph.

The method of Basse et al. (2019b), for instance, can be viewed as implicitly considering cliques of the null-exposure graph with (possibly) overlapping assignments; that is, assignments in the cliques form a covering — not a partition — of $\mathbb{Z}$ and an assignment may belong to more than one clique. The conditioning mechanism is then uniform on the set of all cliques containing the observed assignment. This approach works in their particular setting and results in powerful tests because the conditioning is guided by the observed assignment, Z^{obs}. The drawback of this approach, however, is that there is no general or automated way to construct good clique coverings, instead requiring case-by-case derivations. Specifically, the covering that is implied by the conditioning mechanism constructed in Basse et al. (2019b) works only in two-stage experiments with stratified and clustered interference.

On the other hand, the approach of Athey et al. (2018) applies, in principle, to more general settings but typically leads to underpowered tests. To understand their approach, we need some additional notation. For any $U \subseteq \mathbb{U}$, let $C(U; z)$ denote the largest clique of the null-exposure graph that contains only units from U and also contains assignment z. Then, the conditioning mechanism implicitly considered by Athey et al. (2018) is of the form:

$$P(C \mid Z) = P(U)\mathbb{I}\{C = C(U; Z)\}, \tag{4}$$

where $P(U)$ is specified by the analyst. In other words, the approach of Athey et al. (2018) implicitly suggests, first, to sample units from one side of the bipartite null-exposure graph, and then to calculate the induced clique that contains the observed assignment. This approach is general, but the random choice of U may lead to underpowered or even ill-defined tests. This is the case when, for example, $C(U; Z)$ is empty due to a poor initial choice of U.

Our method combines the benefits of both approaches. Unlike the clique covering strategy implicit in Basse et al. (2019b), our clique decomposition approach is not problem-specific and completely automates the construction of condition mechanisms. Also, in contrast to the method in Athey et al. (2018), our proposed approach gives concrete guidance on how to properly condition the test to achieve higher power. We discuss these advantages in the context of clustered and spatial interference in Sections 5 and 6.

4.3. Clique decomposition algorithm

We end this section by presenting an algorithm for decomposing the null-exposure graph into cliques, which is assumed in Step 1 of our proposed Procedure 1.

For a given clique, C, in the null-exposure graph, let $\mathrm{E}(C)$, $\mathrm{U}(C)$, and $\mathrm{Z}(C)$ denote the set of edges, the set of units and the set of assignments in the clique, respectively. Thus, $|\mathrm{E}(C)| = |\mathrm{U}(C)||\mathrm{Z}(C)|$ since the clique is bipartite. Also, let $C \in G$ indicate that C includes nodes and edges from null-exposure graph G. Our proposed clique decomposition algorithm can be described as follows:

1. Start with an empty clique set: $\mathcal{C} = \{\}$, and the original null-exposure graph $G = G_f^{\mathsf{a},\mathsf{b}}$ that

corresponds to the null hypothesis, $H_0^{\mathsf{a},\mathsf{b}}$.

2. Solve the "largest clique problem":

$$C^* = \arg\max_{C \in G} |E(C)|. \tag{5}$$

3. Remove clique edges: $E(G) \leftarrow E(G) \setminus E(C^*)$.

4. Remove clique assignments: $Z(G) \leftarrow Z(G) \setminus Z(C^*)$.

5. Update clique set, $\mathcal{C} \leftarrow \mathcal{C} \cup \{C^*\}$, and repeat from Step 2 if $|E(G)| > 0$.

The output of this procedure is a clique decomposition, $\mathcal{C}$. The main computational challenge is in Equation (5), where we calculate cliques possessing the largest possible number of edges in the remaining null-exposure graph. This is a variant of the maximal edge clique problem, and is computationally challenging — in fact, Peeters (2003) show that finding such cliques is NP-hard. See Zhang et al. (2014) for a review.

In this paper, we use a related method, known as the "binary inclusion-maximal biclustering" (`Bimax`) method Prelić et al. (2006) to solve Equation (5). This is a fast divide-and-conquer method to find sub-blocks of ones in a binary matrix, known as biclusters, mainly used in the bioinformatics and gene expression literature. The `Bimax` algorithm finds all biclusters that are not contained in any other bicluster. The algorithm is implemented in the R package `biclust` and can incorporate constraints on the solution space to speed up computation. The constraints are placed on the number of unit and assignment nodes so that `Bimax` returns all cliques C^* where, for example, $|\mathrm{U}(C^*)| \geq n_0, |\mathrm{Z}(C^*)| \geq n_1$ for specified, integer-valued parameters n_0 and n_1. Constraining the search space allows for relatively fast computation of cliques. By choosing n_0 and n_1 so that they are similar to each other and their product is large, we are able to quickly approximate the output of Step 2 of our FRT. The `Bimax` algorithm's runtime is on the order of minutes to complete a clique decomposition in our application with the Medellín street network.

5. Application to clustered interference: A simulation study

In this section, we illustrate the proposed FRT of Section 4 in settings with clustered interference, as described in Example 1. In these settings, we assume that interactions between units occur within, but not between, well-defined clusters of units, such as households, classrooms, or firms. For concreteness, we illustrate our method for a two-stage randomized trial in which clusters are assigned to treatment or control and then, within each treated cluster, one unit is randomly assigned to treatment; see Basse and Feller (2018, Section 8).

5.1. Problem setup and comparison of available methods

Following Example 1, we have N units divided equally into K clusters. In this example, the exposure function is defined as a pair $f_i(Z) = (W_i, Z_i)$, where Z_i denotes the treatment status of unit

i, and $W_i = \sum_{j \in [i], j \neq i} Z_j$ gives the treatment status of unit i's cluster, denoted $[i]$. Thus, each unit i has three possible exposures, namely $(0,0), (1,0)$ and $(1,1)$, which correspond to "control," "spillover", and "treated"; and three potential outcomes, denoted, respectively, by $Y_i(0,0), Y_i(1,0), Y_i(1,1)$. For illustration, we assume that we first treat $K/2$ clusters at random, and then randomly treat one unit in each treated cluster.

We focus on the following spillover effect hypothesis,

$$H_0: \quad Y_i(1,0) = Y_i(0,0) + \tau, \text{ for all } i, \tag{6}$$

and vary τ. We assess validity when $\tau = 0$ and power when $\tau \neq 0$. The null hypothesis in Equation (6) is therefore an instance of $H_0^{\mathsf{a},\mathsf{b}}$ in Equation (1), with $\mathsf{a} = (1,0)$ and $\mathsf{b} = (0,0)$. Following our discussion in Section 4.2, we compare three methods for testing H_0:

(i) The clique test proposed in this paper (Procedure 1);

(ii) The method of Basse et al. (2019b), which samples one focal unit per cluster among those who are not treated ("conditional focals");

(iii) The method of Athey et al. (2018), which samples one focal unit per cluster at random ("random focals").

Figure 2 illustrates how these tests differ in a hypothetical example with six units arranged in two clusters, namely, $\{1,2,3\}$ and $\{4,5,6\}$, and where unit 6 is assigned to treatment (colored green). For unit 6, we observe the treated potential outcome, $Y_6(1,1)$; thus, unit 6 is not connected to any other assignment in the null-exposure graph, since we have no information about $Y_6(0,0)$ or $Y_6(1,0)$ under H_0. For units 4 and 5, we observe the spillover potential outcomes, $Y_4(1,0)$ and $Y_5(1,0)$, respectively; and for units 1, 2, and 3 in the first cluster, we observe the control potential outcomes $Y_1(0,0), Y_2(0,0)$, and $Y_3(0,0)$, respectively. Under H_0, we can impute the control potential outcomes for units 4 and 5 and the spillover potential outcomes for units 1, 2, and 3.

Figure 2 also depicts the conditioning events for every method, i.e., the cliques implied by each method, with the selected focal units in each clique marked with an asterisk. The leftmost subfigure shows that the clique test in Procedure 1 conditions on the clique that includes $\{2,3,4,5\}$ as focal units. Unit 6 is not included since it does not add any edges in the objective function of Equation (5). This clique is balanced and dense in the sense that there are two units in each exposure of interest, and there are many edges connecting those units to the assignments on the other side. More density means more support for the randomization test, which leads, intuitively, to more power. The middle subfigure shows the "conditional focals" method of Basse et al. (2019b). We see that only one focal unit is selected per cluster by construction of this test. This leads to a sparser clique, and a lower power compared to the clique test. The rightmost subfigure shows the "random focals" method of Athey et al. (2018). This differs from "conditional focals" because it can select treated units as focal units. For illustration, if the method selects unit 6, which is assigned to treatment, this unit is effectively dropped from the conditioning clique. We return to the issue of power in Section 7.4.

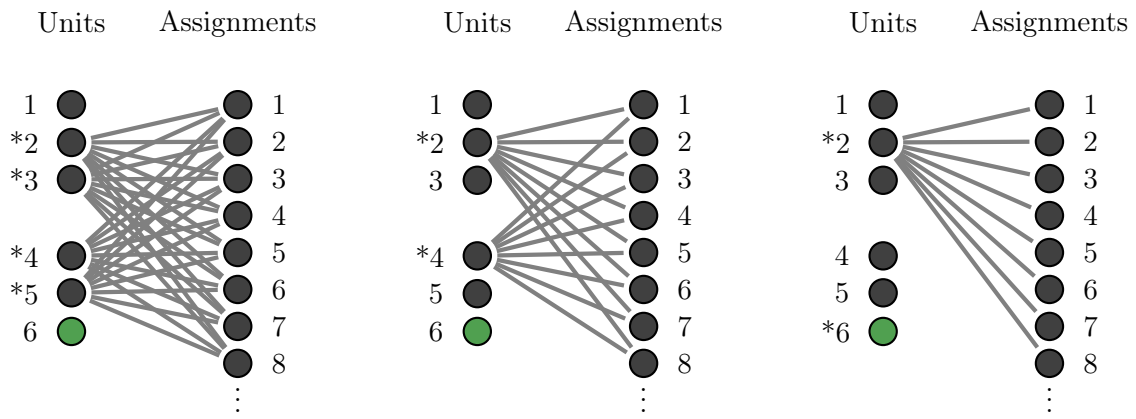

Figure 2. Example cliques used by the three methods of Section 5.1. Interference is clustered with six units divided equally into two clusters. Z^{obs} is Assignment 1 and unit 6 is treated (colored green). The units of each clique (focal units) are marked by an asterisk, "*". An edge denotes that the unit is exposed to either $(1, 0)$ or $(0, 0)$, the exposures in H_0 of Equation (6).
Left: conditioning event of the clique test: two focals per cluster; *Middle:* conditioning event of "conditional focals" (Basse et al., 2019b): one untreated focal per cluster is randomly chosen (units 2 and 4); *Right:* conditioning event of "random focals" (Athey et al., 2018): only one focal per cluster is randomly chosen (units 2 and 6). However, unit 6 is effectively removed since unit 6 is treated and, there are no edges between this unit and any assignment.

5.2. Simulation study: Two-stage experiment

We follow Basse and Feller (2018) to generate data for the setting with clustered interference:

$$\begin{aligned} y_{i,00} &\sim N(\mu_{00}, \sigma_\mu^2), \\ \tau_i^P &\sim N(\tau^P, \sigma_\tau^2), \\ \tau_i^S &\sim N(\tau^S, \sigma_\tau^2), \\ y_{i,11} &= y_{i,00} + \tau_i^P, \\ y_{i,10} &= y_{i,00} + \tau_i^S. \end{aligned}$$

To generate the individual potential outcomes, we sample $Y_i(c, z) \sim N(y_{i,cz}, \sigma_y^2)$, where $c, z \in \{0, 1\}$. As such, τ_i^P and τ_i^S correspond to idiosyncratic primary and spillover effects, respectively. We consider the following specifications: $N = 300$, $\mu_{00} = 2$, $\sigma_\mu = \sigma_\tau = 0.1$, $\sigma_y = 0.5$, $\tau^S = 0.7$, $\tau^P = 1.5$, and $K \in \{20, 25, 30, 50, 60, 75, 100, 150\}$. We vary K among eight values to see how different methods perform with small and large-sized clusters. The different cluster sizes, N/K, are therefore contained in the set $\{15, 12, 10, 6, 5, 4, 3, 2\}$. For the simulations, we generate 5,000 different assignments and construct the null-exposure graph for the clique method. For each cluster size and a fixed additive effect, τ, we generate 2,000 data sets from the DGP given above. τ is varied among 300 equally spaced values from 0 to 1.

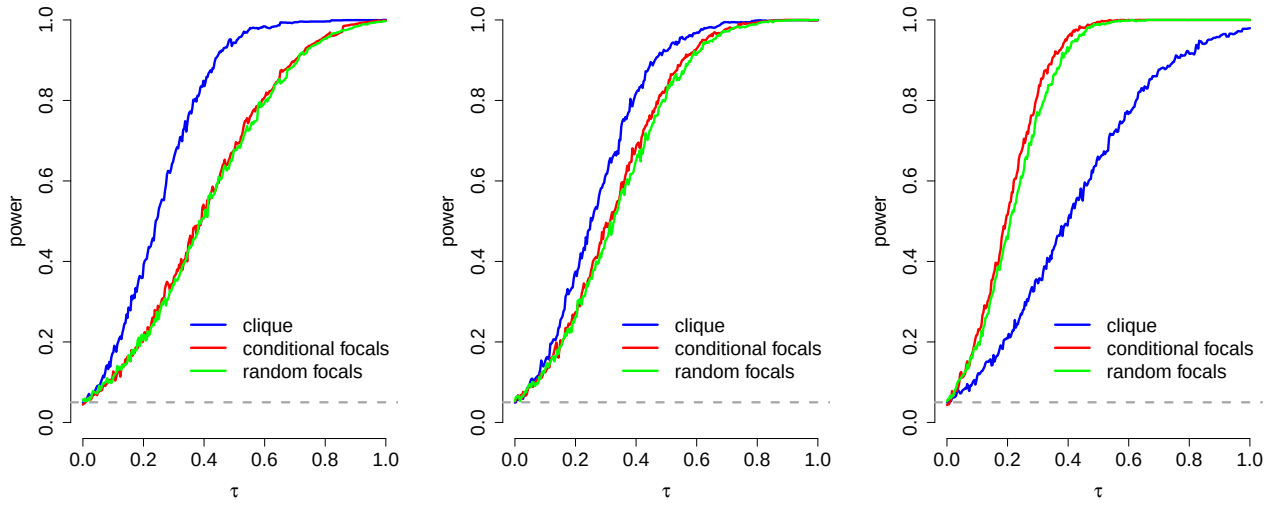

Figure 3. Power plots for three data configurations: *left:* $N = 300, K = 20$; *middle:* $N = 300, K = 30$; and *right:* $N = 300, K = 75$.

The power plots are shown in Figure 3. From left to right, we report results for $K = 20, 30, 75$ which corresponds to $15, 10, 4$ units per cluster, respectively. We see that the clique method performs substantially better with larger cluster sizes. Recalling the conclusions from Figure 2, this is explained by noting that the clique method can select multiple focal units per cluster in the clique decomposition. In contrast, the methods of Basse et al. (2019b) and Athey et al. (2018) both define one focal unit per cluster. To confirm, we calculated the average number of focal units per cluster for each method and data configuration with 15 units per cluster ($N = 300, K = 20$). The clique method had 5.24 focal units per cluster compared to 1 for method (ii) and 0.97 for method (iii).

To investigate what affects the power of the clique test, we gather data on the cliques generated by the clique decomposition. We obtain variability by altering the optimization constraints in the clique decomposition algorithm we employ (see Section 4.3 for details), and fix $N = 300, K = 20$ and $\tau = 0.3$. Figure 4 displays the results. We see that the average number of units and assignments in the decomposition are negatively correlated with each other, which suggests a power tradeoff. As the number of (focal) units per clique increases, power also increases until some threshold. Beyond this threshold, more units in the clique come at the expense of fewer treatment assignments, which on aggregate decreases power. While the actual threshold value depends on the context, researchers will generally face this tradeoff between units and assignments.

Not evident in Figures 3 and 4 is the relative ease of implementation for each method. The clique method is straightforward and essentially automated for a well-defined clique decomposition algorithm. On the other hand, methods (ii) and (iii) may be hard to implement. For example, Basse et al. (2019b) show that their test can be implemented as a permutation test only under the conditioning mechanism described in (ii). Selecting more units per cluster to increase power will generally break this property. Furthermore, the test described in (iii) is a permutation test only when all clusters have equal size. It

is not clear how to implement a valid test based on the method of Athey et al. (2018) with unequal cluster sizes.

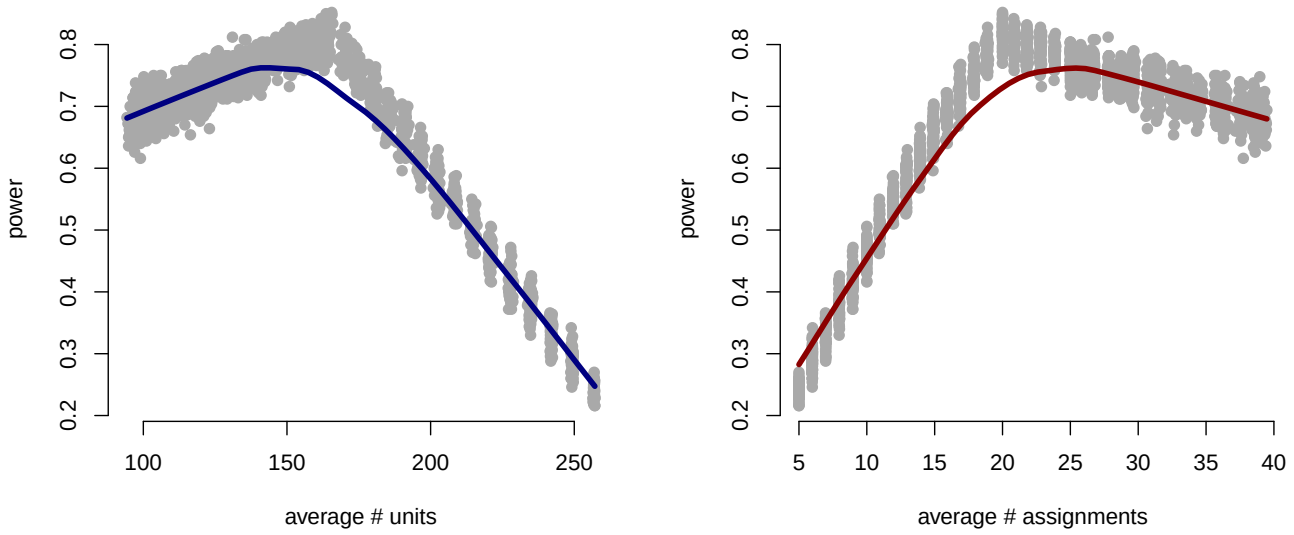

Figure 4. Power values for varying clique characteristics. The data structure is fixed at $\{N = 300, K = 20\}$ and $\tau = 0.3$. Each individual gray dot corresponds to a clique decomposition of the null exposure which we condition on for the clique method. The left graph shows power as a function of the average number of units (focals), and the right graph shows power as a function of the average number of units. A figure combining this information into a single plot is shown in Appendix B.

6. Application to spatial interference: Crime in Medellín

In this section, we illustrate our method in a spatial interference setting in which interactions occur between "neighboring" units, but without the simpler structure of clustered interference. We focus on re-analyzing a large-scale experiment in Medellín, Colombia studying the impact of "hotspot policing" on crime (Collazos et al., 2019).

6.1. Problem setup and comparison of available methods

Following Collazos et al. (2019), the units are $N = 37,055$ street segments, 967 of which were identified as *hotspots* using geo-located police data and further consultation with police. Of these hotspots, 384 were randomly assigned to treatment, a six-month increase in daily police presence, via a completely randomized design over a domain $\mathbb{Z}$ of roughly $10,000$ possible assignments. The outcome of interest is a *crime index*, a weighted sum of the crime counts on each street segment.[1]

[1] As discussed in Collazos et al. (2019), the index weights are chosen based on the length of sentence for a given crime. They are: 0.550 for homicides, 0.112 for assaults, 0.221 for car and motorbike theft, and 0.116 for personal robbery. Crime data is matched to street segment within 40-meter buffers. In other words, if a crime happened in an alley, it will

To define the spillover hypothesis, we need to define the exposure function, f. Following Collazos et al. (2019), we use geographic distance as a measure for spillover exposure. Let $d(i,j)$ be the distance (in meters) between unit i and unit j, then we define

$$f_i(Z) = \begin{cases} \text{“pure control”}, & \text{if } Z_i = 0 \text{ and } \sum_{j \neq i} \mathbb{I}\{d(i,j) \leq r\} Z_j = 0; \\ \text{“spillover}_r\text{”}, & \text{if } Z_i = 0 \text{ and } \sum_{j \neq i} \mathbb{I}\{d(i,j) \leq r\} Z_j > 0; \\ \text{“other”}. \end{cases} \tag{7}$$

This defines a “pure control” as a control unit that has no treated units closer than r meters. A control unit receives “spillovers” when there are treated units closer than the specified distance.

Our goal is to test contrast hypotheses of the form $H_0^{\mathsf{a},\mathsf{b}_r}$, with a = “pure control” and b_r = “spillover$_r$”, where r is a free variable in the set:

$$r = \{75, 100, 125, 150, 175, 225, 275, 325, 375, 425\}.$$

Each hypothesis for a given r will have its own null-exposure graph and clique decomposition, which are necessary to execute our clique test in Procedure 1. Following Collazos et al. (2019), we refer to “short-range spillovers” as the set of hypotheses $H_0^{\mathsf{a},\mathsf{b}_r}$ with $r \leq 125$m.

To illustrate, Figure 5 shows the induced exposures under the randomization realized in the experiment (left), and an alternative randomization that was not realized (right), for $r = 125$m. The figure highlights short-range spillover units (among all available units) in light blue and light green for the observed treatment and randomization, respectively. Figure 5 also depicts important geographic features of Medellín, with many hotspots near the city center, and where dark areas or holes correspond to physical barriers (e.g., mountains) or major infrastructure (e.g., airports). Even though only 967 street segments can receive the active treatment, every street segment can potentially receive spillovers, and so we use the entire street network in our analysis. A key challenge is that street segments near the city center have a much higher probability of being exposed to crime spillovers than segments on the outskirts of the city.

Finally, unlike our analysis of clustered interference, we will only apply the clique test in this case study. In principle, we could adapt the test of Athey et al. (2018) to this setting, but the implementation of their procedure would be underpowered due to the spatial structure. Specifically, we will see that (under the short-range spillover hypothesis) focal units are concentrated either at the center or outskirts of the city; see Section 6.3 and Figure 8. It is unlikely that this “center-outskirts” pattern could be generated through a random selection of focals. Furthermore, the ϵ-net method of focal selection (Athey et al., 2018, Section 5.4.2) would not work because it would generate patterns of focal units that are spatially uniform. Similarly, it is not clear how to apply the approach of Basse et al. (2019b), which is more narrowly tailored to the clustered interference setting.

be matched to the closest street segment within a 40-meter radius. If there is overlap, it is matched to street segment closest in terms of Euclidean distance.

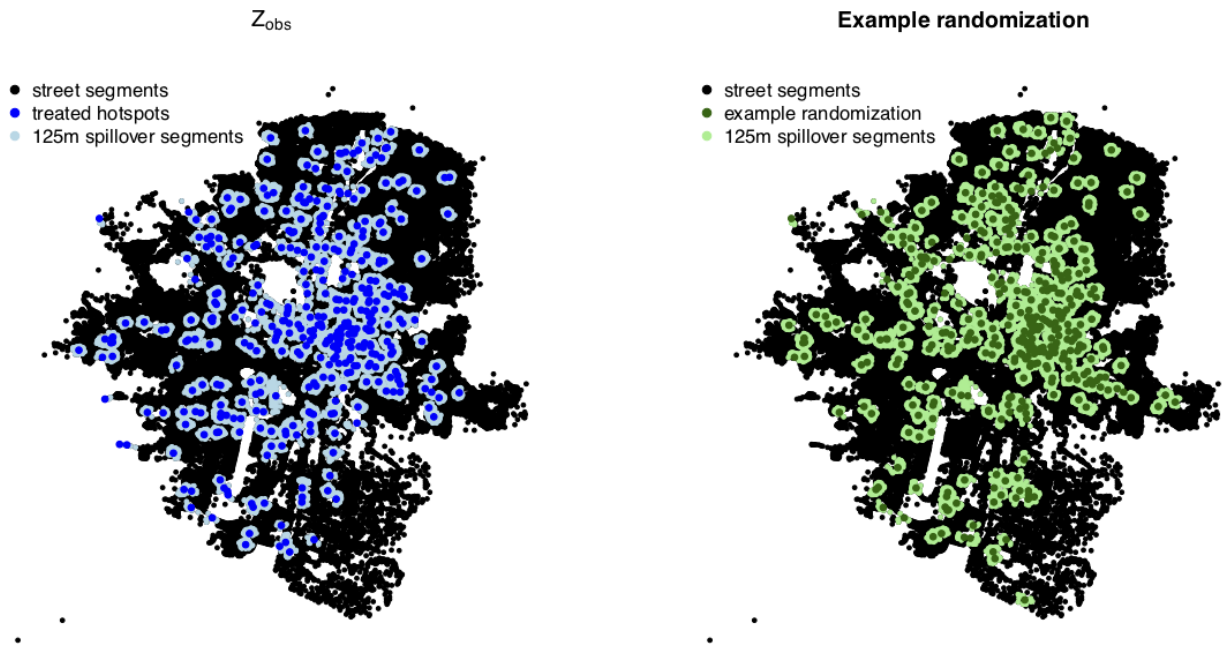

Figure 5. Street segments, hotspots, and treated hotspots for the data set. The left figure is the observed assignment for the experiment. The right figure is an example randomization of the assignment vector. The dots cover different hotspots, but they are still within the 967 segments representing the hotspots. Additionally, the light colored dots represent the 125m spillover units, i.e.: street segments that are within 125m of the treated hotspots for a given randomization.

6.2. Spillover hypotheses: A simulation study

We now assess our proposed method via a simulation study calibrated to the actual Medellín street network. For these simulations, each clique decomposition is constrained to include cliques with at least 100 focal units and 1000 assignments. We explore the effect of radius r on test power using the following simple model for the outcomes, Y:

$$\begin{aligned} Y_i(\text{“pure control”}) &\sim \text{Gamma}(\alpha, \beta), \\ Y_i(\text{“spillover}_r\text{”}) &= Y_i(\text{“pure control”}) + \tau_r. \end{aligned} \tag{8}$$

The shape and rate parameters (α, β, respectively) are selected to match the mean and variance of the observed outcome in the actual experiment, the crime index. The parameter τ_r determines an additive spillover effect at radius r. We set $\tau_r \propto 1/r^2$ to allow for heterogeneity of the spillover effect with respect to radius.[2] In the simulation, we sample assignments according to the true design for

[2]This is in line with existing literature (Thomas, 2013; Barr and Pease, 1990; Verbitsky-Savitz and Raudenbush, 2012), which has pointed out that “the likelihood that an offender will target an opportunity will be inversely related to the distance it is located from their routine activity spaces” (Eck, 1993; Johnson et al., 2014).

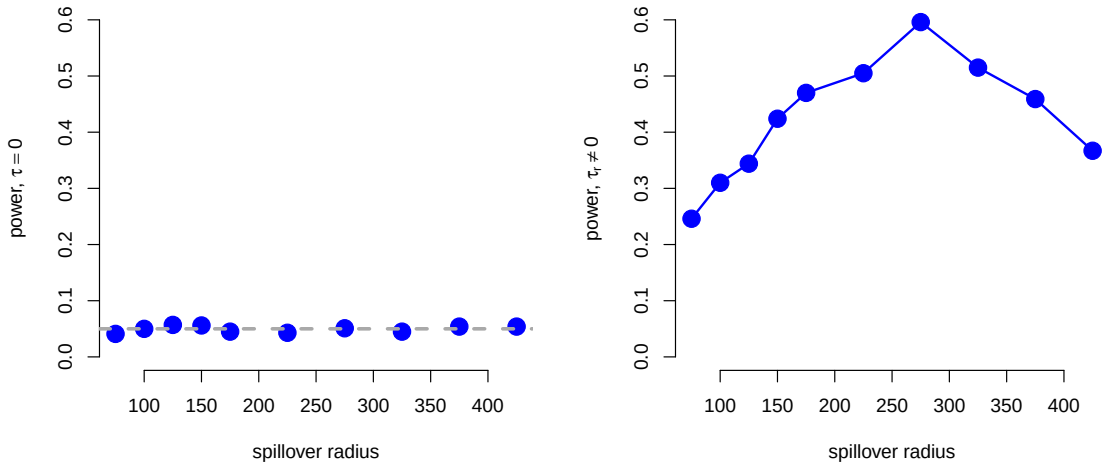

Figure 6. The left figure shows the power analysis across 1000 simulations when $\tau_r = 0$. We reject the null at the 0.05 level. Therefore, the left figure confirms the validity of the clique method. The right figure displays the power analysis for nonzero $\tau_r \propto 1/r^2$.

every value of r and outcomes according to Equation (8). We then test the null hypothesis $H_0^{\mathsf{a},\mathsf{b}_r}$ on spillovers at distance r, defined at the beginning of this section. The test statistic is the simple difference in means between focal units exposed to a and b_r.

The results are shown in Figure 6. In the left subfigure, we fix the additive treatment effect at zero ($\tau_r = 0$) to assess validity. Our rejection level is 0.05, and so we confirm the validity of the clique method since all power values gather around the 5% rejection rate. In the right subfigure, we consider nonzero additive spillovers effects ($\tau_r > 0$) that are calibrated based on the spillover radius. We observe that the power curve is generally concave and nonmonotonic since it increases until some radius and then decreases. At first, this may be counterintuitive since as r increases the spillover effect, τ_r, decreases (by definition), which should make it harder to detect. However, as shown in Appendix B, the number of focal units also increases sharply with respect to r. The net effect is an increase of power. At the same time, as we discussed in Section 5.2, an increase in the number of focals per clique generally results in a decrease in the number of assignments per clique (see also Figure 11 in Appendix B), and, eventually, in decreased testing power. Under our assumed outcome model, the maximum power is achieved approximately at a 275m spillover radius.

This kind of analysis gives a useful estimate for the power profile (Figure 6, right) of our proposed clique test for a given clique decomposition algorithm. In practice, we could compare between the power profiles of different clique decomposition algorithms, and choose the most favorable algorithm to apply on the real data. See Section 7.4 for additional discussion on testing power.

6.3. Spillover hypotheses: Clique test on real data

In this section, we demonstrate our proposed clique test using the real outcome data. We focus on the spillover hypothesis $H_0^{\mathsf{a},\mathsf{b}_r}$ with radius $r = 125$m, following Collazos et al. (2019) who define this type of exposure as "short-range spillover". Results for all radius values in the previous simulation are included in Appendix C. We therefore test whether there is a difference between outcomes of pure control units and units who receive short-range spillovers:

$$H_0 : Y_i(Z) = Y_i(Z'), \text{ for all } Z, Z' \text{ such that } f_i(Z), f_i(Z') \in \{\text{"pure control"}, \text{"spillover}_r\text{"}\}, \quad (9)$$

where $r = 125$m, and the exposures are defined in Equation (7).

The first step of the clique test is to construct the null-exposure graph. This graph has 37,055 nodes on one side (number of streets/units), and 10,000 nodes on the other (number of assignments). Figure 7 displays a visual summary of this graph. The left subfigure shows a block of the null-exposure graph that is reasonably discernible, where a dot corresponds to a unit-assignment pair, say (i, z). If the dot has white color, then there is no edge between unit i and assignment z in the graph. The light blue and navy blue colors indicate whether i receives short-range spillover or pure control exposure under z, respectively. We note that the null hypothesis is not sharp for units exposed to "white", therefore we need to condition on a (bi)clique where the white components are effectively removed. This is analogous to choosing a clique where all units are either exposed to "navy blue" or "light blue". The right subfigure of Figure 7 shows such clique. The colorings on this subfigure provide important information. We see, for example, that many units in the clique are always pure control (navy blue columns), and a handful of units are always short-range spillovers (light blue columns). Ideally, more variation in exposures across units (i.e., more random coloring in Figure 7, right) leads to more power. Currently, our clique decomposition algorithm cannot guarantee such variation; we leave this problem open for future work.

We now apply our proposed test in Procedure 1 to the spillover hypothesis of Equation (9). The left side of Figure 8 visualizes hotspots, treated hotspots, short-range spillover and pure control units, and focal units identified by the clique decomposition. As mentioned earlier, most focal units are at the center or outskirts of the city due to the particular spatial structure of spillovers. The right side of Figure 8 displays the randomization distribution of the test statistic measuring the difference in means between crime index values on short-range spillover units and pure control units:

$$T_C(z, y) = \frac{1}{N_{\mathsf{b}}} \sum_{i \in C} \mathbb{I}\{f_i(z) = \mathsf{b}\} Y_i - \frac{1}{N_{\mathsf{a}}} \sum_{i \in C} \mathbb{I}\{f_i(z) = \mathsf{a}\} Y_i, \quad (10)$$

where C denotes the clique we condition on, $i \in C$ denotes that unit i is a node in the clique, a = "pure control", b = "short-range spillover", $N_{\mathsf{a}} = \sum_{i=1}^{N} \mathbb{I}\{f_i(Z) = \mathsf{a}\}$ and $N_{\mathsf{b}} = \sum_{i=1}^{N} \mathbb{I}\{f_i(Z) = \mathsf{b}\}$ are the exposure counts. Thus, positive values of the test statistic indicate that the average crime outcome for units exposed to "short-range spillovers" is larger than the average outcome for units that are exposed to "pure control".

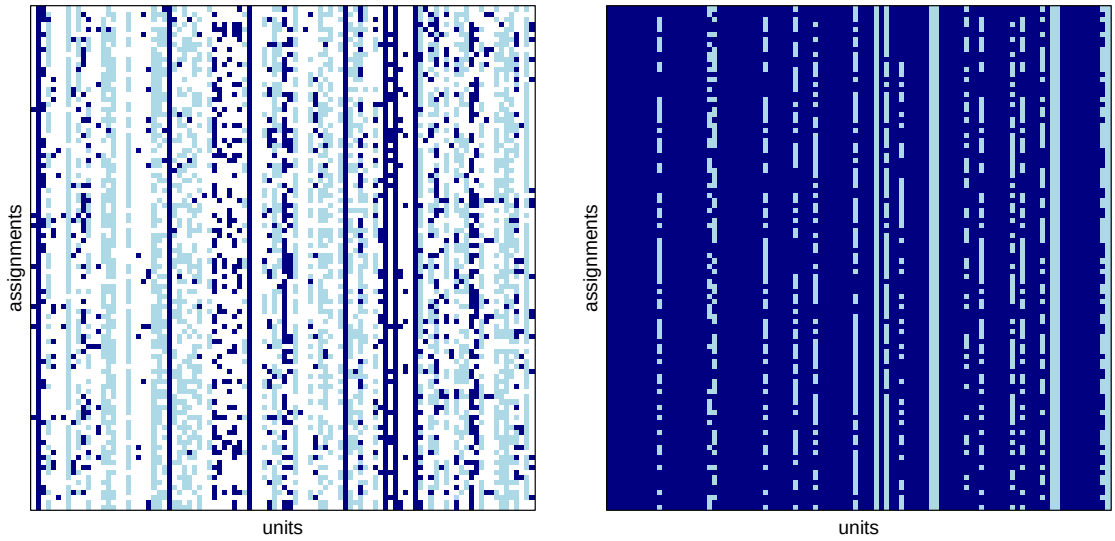

Figure 7. The left figure visually depicts the null-exposure graph. The vertical axis corresponds to the assignments from the randomization procedure, and the horizontal axis displays the units. Light blue denotes an untreated unit that is a spillover and close to a treated low crime hotspot, and navy denotes pure control. The right figure is a clique of the null-exposure graph containing the observed assignment. There is no white in the clique since it does not contain "unexposed" units by construction. To conserve space, we only display the first 100 assignments and units for both.

The randomization values of the test statistic are computed conditional on the clique depicted on the right of Figure 7. The observed test statistic is only larger than 7.7% of all the randomization values, which is not significant at the 0.05-level. Assuming an additive short-range spillover effect, inversion of the randomization test gives us $[-0.51, 0.03]$ as the 95% confidence interval, indicating that negative values for the spillover effect are more plausible under the additivity assumption. This observation suggests a decrease in crime of street segments surrounding an area with increased law enforcement/community policing, which is consistent with the literature (Collazos et al., 2019; Verbitsky-Savitz and Raudenbush, 2012).

One caveat with this result is that even though our randomization test is always valid, the power of the test may be affected by the inherent differences between the focal units. Specifically, due to the particular spatial arrangement in our application, the focal units that receive spillovers are mostly downtown streets, whereas units that are pure controls are mostly on the outskirts of the city (see Figure 8, left). The observed differences in crime outcomes of these focal units could depend, say, on differences in demographics between these two city areas. Failing to account for such differences could reduce the power of the randomization test. We discuss this issue in more detail, along with potential solutions using covariate adjustments in Section 7.3.

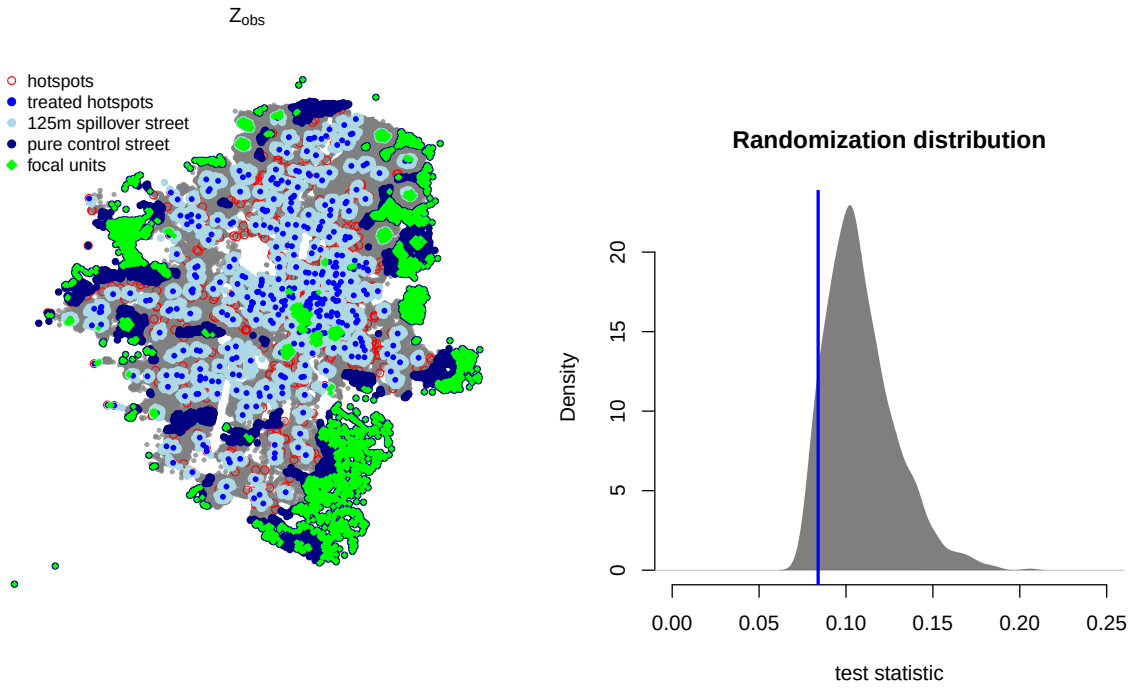

Figure 8. *Left:* Representation of unit exposures for Z^{obs}. Also shown in green are the focal units from the clique represented on the right in Figure 7. *Right:* Test of the $r = 125$m spillover radius hypothesis, where the test statistic is a difference in average outcomes between "short-range spillover" units and pure control units, defined in Equation (10). Shown is the distribution of the test statistic under the null, and the blue line is the observed test statistic. The p-value of the observed test statistic is approximately 0.077.

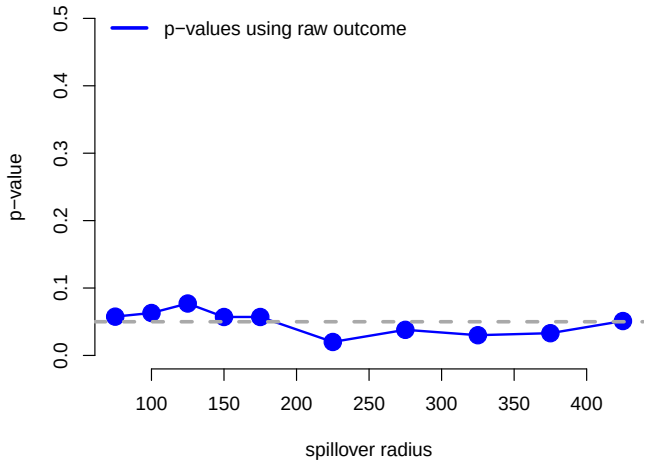

Figure 9. P-values (left vertical axis) for clique tests with varying spillover radii (horizontal axis). The blue line shows p-values for tests using the raw crime index.

Finally, we conduct clique randomization tests, as described in the previous section, while varying

the distance threshold r for the definition of spillovers. The results are shown in Figure 9, which also includes the results for the 125m-spillover presented in Figure 8 of the previous section. For outcomes, we consider the raw crime index (as in Section 7.3). We see that the p-values for the raw crime index are all small for varying radii; see the flat blue line in Figure 9. This suggests that some form of spillovers exists, where the distance does not seem to matter. Alternatively, the results could indicate that the spillover effects may be heterogeneous with respect to distance. We investigate this question further in Appendix C, where the clique tests on outcomes adjusted for known covariates show that spillovers are significant only at small distances.

7. Discussion and extensions

In this section, we discuss a few aspects of our methodology in more detail, including computation, composite hypotheses and heterogeneous spillover effects.

7.1. Implementation for arbitrary designs

Although our method works for arbitrary designs, it can become computationally intractable if the support, $\mathbb{Z} = \{z : P(z) > 0\}$, is too large (on the order of hundreds of thousands of nodes), since clique enumeration is NP-hard. Fortunately, a small modification of our test can address this issue. The idea is to add a step at the beginning of Procedure 1 that subsamples assignments to limit the size of $\mathbb{Z}$. In Appendix A we show that the following procedure is still valid:

1. Draw $Z^{obs} \sim P(Z^{obs})$.
2. Draw $M-1$ assignments uniformly at random from $\mathbb{Z} \setminus \{Z^{\text{obs}}\}$ and let $\mathbb{Z}_M$ be the set of size M formed as the union of $\{Z^{obs}\}$ and the set of size $M-1$ just constructed.
3. Run our clique test in Procedure 1, using the null-exposure graph of the new support set, $\mathbb{Z}_M$.

7.2. Composite hypotheses

As mentioned in Section 2.2, not all interesting spillover hypotheses can be expressed in the form of Equation (1). For example, Athey et al. (2018, hypothesis 2) consider a hypothesis of the form:

$$H_0 : Y_i(Z) = Y_i(Z'), \text{ if } Z_i = Z'_i. \tag{11}$$

This is essentially testing the "no interference" assumption, also known as "stable unit treatment value assumption" (Rubin, 1980). If we set $f_i(Z) = Z_i \in \{0, 1\}$, then H_0 is in fact a composite hypothesis comprised of the contrast hypotheses, $H_0^{0,0}$ and $H_0^{1,1}$, of Equation (1). The main difficulty in testing H_0 through our framework is that it is not possible to build the null-exposure graph, since Definition 1 requires only one set of exposures, whereas the composite hypothesis has two.

We can extend our proposed test to calculate the null-exposure graph based on the units' observed exposures. Specifically, define the vertex set as $V = \mathbb{U} \cup \mathbb{Z}$, the joint set of units and assignments. Second, define the edge set as follows:

$$\widetilde{E} = \left\{ (i, z) \in \mathbb{U} \times \mathbb{Z} : f_i(z) = Z_i^{\text{obs}} \right\}. \tag{12}$$

Then, the clique test in Procedure 1 can be applied on the updated null-exposure graph, $G = (V, \widetilde{E})$, to test the null in Equation (11). This construction can be generalized through the following theorem.

Theorem 3. *Consider a null hypothesis, H_0, that is a composite hypothesis comprised of individual contrast hypotheses, $H_0^{\mathsf{a}_j, \mathsf{b}_j}, j = 1, \ldots, J$, for some exposure functions $f_i, i = 1, \ldots, N$, where all $\mathsf{a}_j, \mathsf{b}_j$ are distinct, for any $j = 1, \ldots, J$, with J fixed. Consider the null-exposure graph, $G = (V, E)$, where $V = \mathbb{U} \cup \mathbb{Z}$ is the joint set of units and assignments, and*

$$E = \left\{ (i, z) \in \mathbb{U} \times \mathbb{Z} : f_i(z) \in A_i \right\}, \tag{13}$$

where $A_i = \{\mathsf{a}_{j'}, \mathsf{b}_{j'}\}$ is the unique exposure set for which $f_i(Z^{\text{obs}}) \in A_i$. Then, the clique test in Procedure 1 operating on the null-exposure graph G is a valid test for H_0.

Theorem 3 shows that our proposed method can in fact test more complex hypotheses than the contrast hypothesis of Equation (1). The key requirement is that the complex hypothesis can be expressed as a composite hypothesis of non-overlapping contrasts.

7.3. Covariate adjustment for heterogeneous focals

In the Medellín application, we discussed how the difference between short-range spillovers and pure control affects the power of the spillover hypothesis tests. The concern becomes evident in Figure 8, where we see that the focal units that receive spillovers are mostly downtown streets, whereas units that are pure controls are mostly on the outskirts of the city.

One straightforward way to address such possible heterogeneity in the focal units is to control for known covariates. For example, we could regress outcomes on observed covariates, and then perform our proposed clique test on the residuals (Rosenbaum et al., 2010). To illustrate, we used this approach for the test of Section 6.3 and adjusted the outcome by distance from important societal center points, such as school, police station, courthouse, church, park, as well as "comuna", which represents a neighborhood or district. The results are shown in Figure 10. The new p-value is 0.13, and suggests that short-range spillover effects are not statistically significant compared to pure control. In Appendix C, however, we include a randomization analysis for many different radii which suggests that spillovers may exist at distances larger than 125m. This result hints at the insufficiency of geographic distance to fully capture the intensity of spillovers. In future work, we could incorporate additional information in the distance function, such as socioeconomic differences between street segments. An advantage of the clique-based testing methodology is the ability to arbitrarily define the distance function (and thus exposures) of interest.

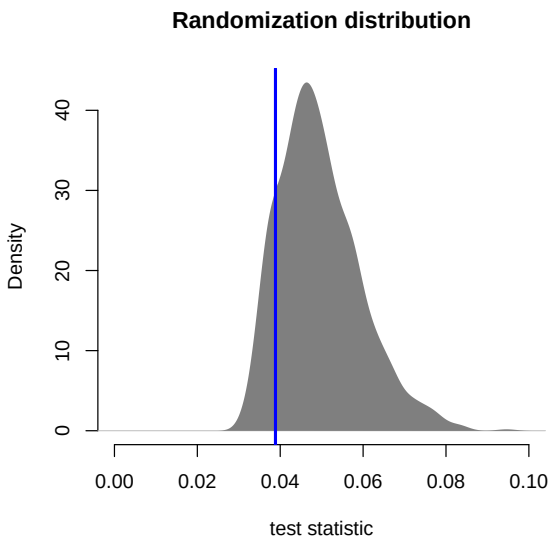

Figure 10. Randomization test for 125m spillover radius hypothesis, applied on the residuals of a regression of outcomes on known covariates. The p-value of the observed test statistic is approximately 0.13.

More broadly, this analysis suggests that adjusting for heterogeneity may be important in practice. The regression-based approach could be extended to adapt related randomization-based approaches that account for treatment effect heterogeneity (Ding et al., 2016). Another approach could be to balance the focal units while incorporating covariates. We leave these ideas open for future work.

7.4. Power

In this paper, our main theoretical results (Theorem 2 and Theorem 3) refer to the validity of our proposed clique test based on the null-exposure graph. The analysis of power is generally complicated as it depends on many factors, including the test statistic which is user-defined. In the context of our proposed method with the null-exposure graph, intuition and empirical evidence suggests that power depends mainly on the clique size and shape we condition on; this is backed by extensive empirical evaluations. It follows that a clique decomposition algorithm that produces a decomposition containing cliques with similar size and shape properties is preferable to an algorithm that has more variation, since the latter may lead to a suboptimal clique.

Thus, one way to optimize the power of our clique test is to consider multiple decomposition algorithms, and estimate their power profiles using an outcome model that is reasonable for the application. For example, the `Bimax` algorithm in Section 4.3 is parameterized by the minimum desired size of cliques. We used this feature to produce Figure 4 and Figure 11 in Appendix B. For a given algorithm, we can simulate outcomes, and calculate the power profile of the test as a function of, say, the spillover parameter, as we did in Section 6.2 and Figure 6. We could then select the algorithm with the most favorable summary of such power profile. Of course, this approach is limited in that it relies on specifying a reasonable outcome model.

8. Conclusion

In this paper, we extend the classical Fisher randomization test to settings with general interference. Our main contribution is the concept of the null-exposure graph, which represents the null hypothesis as a bipartite graph between units and assignments. For the subset of units and assignments that comprise a clique on this graph, the null hypothesis of interest is sharp and the corresponding (conditional) randomization test is valid. We illustrate the benefits of this approach in both clustered and spatial interference settings, showing clear advantages over existing methods.

There are a number of promising directions for future work. First, we could investigate the power of our randomization tests through graph-theoretic properties, such as density, of the null-exposure graph. Second, we could investigate how much data "is thrown away" by conditioning, and suggest clique decompositions of the null-exposure graph that minimize such data loss. Finally, it would be interesting to know under which conditions our proposed tests can be implemented more efficiently.

References

Aronow, P. M. (2012). A general method for detecting interference between units in randomized experiments. *Sociological Methods & Research*, 41(1):3–16.

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

Aronow, P. M., Samii, C., and Wang, Y. (2019). Design-based inference for spatial experiments with interference.

Athey, S., Eckles, D., and Imbens, G. W. (2018). Exact p-values for network interference. *Journal of the American Statistical Association*, 113(521):230–240.

Barr, R. and Pease, K. (1990). Crime placement, displacement, and deflection. *Crime and justice*, 12:277–318.

Basse, G., Ding, P., Feller, A., and Toulis, P. (2019a). Randomization tests for peer effects in group formation experiments. *arXiv preprint arXiv:1904.02308*.

Basse, G. and Feller, A. (2018). Analyzing two-stage experiments in the presence of interference. *Journal of the American Statistical Association*, 113(521):41–55.

Basse, G., Feller, A. M., and Toulis, P. (2019b). Randomization tests of causal effects with interference between units. *Biometrika*.

Blume, L. E., Brock, W. A., Durlauf, S. N., and Jayaraman, R. (2015). Linear social interactions models. *Journal of Political Economy*, 123(2):444–496.

Bowers, J., Fredrickson, M. M., and Panagopoulos, C. (2013). Reasoning about interference between units: A general framework. *Political Analysis*, 21(1):97–124.

Brock, W. A. and Durlauf, S. N. (2001). Interactions-based models. In *Handbook of econometrics*, volume 5, pages 3297–3380. Elsevier.

Collazos, D., García, E., Mejía, D., Ortega, D., and Tobón, S. (2019). Hot spots policing in a high crime environment: An experimental evaluation in medellín. *Documento CEDE*, (2019-01).

Cox, D. R. (1958). Planning of experiments.

Ding, P., Feller, A., and Miratrix, L. (2016). Randomization inference for treatment effect variation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(3):655–671.

Eck, J. E. (1993). The threat of crime displacement. In *Criminal Justice Abstracts*, volume 25, pages 527–546.

Graham, B. S. (2008). Identifying social interactions through conditional variance restrictions. *Econometrica*, 76(3):643–660.

Graham, B. S. and Hahn, J. (2005). Identification and estimation of the linear-in-means model of social interactions. *Economics Letters*, 88(1):1–6.

Halloran, M. E. and Hudgens, M. G. (2016). Dependent happenings: a recent methodological review. *Current epidemiology reports*, 3(4):297–305.

Jackson, M. O. (2010). *Social and economic networks*. Princeton university press.

Johnson, S. D., Guerette, R. T., and Bowers, K. (2014). Crime displacement: what we know, what we donâĂŹt know, and what it means for crime reduction. *Journal of Experimental Criminology*, 10(4):549–571.

Manski, C. F. (2013). Identification of treatment response with social interactions. *The Econometrics Journal*, 16(1):S1–S23.

Peeters, R. (2003). The maximum edge biclique problem is np-complete. *Discrete Applied Mathematics*, 131(3):651–654.

Prelić, A., Bleuler, S., Zimmermann, P., Wille, A., Bühlmann, P., Gruissem, W., Hennig, L., Thiele, L., and Zitzler, E. (2006). A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics*, 22(9):1122–1129.

Rosenbaum, P. R. (2007). Interference between units in randomized experiments. *Journal of the American Statistical Association*, 102(477):191–200.

Rosenbaum, P. R. et al. (2010). *Design of observational studies*, volume 10. Springer.

Rubin, D. B. (1980). Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American Statistical Association*, 75(371):591–593.

Sävje, F., Aronow, P. M., and Hudgens, M. G. (2017). Average treatment effects in the presence of unknown interference. *arXiv preprint arXiv:1711.06399*.

Sobel, M. E. (2006). What do randomized studies of housing mobility demonstrate? causal inference in the face of interference. *Journal of the American Statistical Association*, 101(476):1398–1407.

Thomas, T. A. (2013). *Quantifying crime displacement after a hot-spot intervention*. PhD thesis.

Toulis, P. and Kao, E. (2013). Estimation of causal peer influence effects. In *International conference on machine learning*, pages 1489–1497.

Verbitsky-Savitz, N. and Raudenbush, S. W. (2012). Causal inference under interference in spatial settings: A case study evaluating community policing program in chicago. *Epidemiologic Methods*, 1(1):107–130.

Zhang, Y., Phillips, C. A., Rogers, G. L., Baker, E. J., Chesler, E. J., and Langston, M. A. (2014). On finding bicliques in bipartite graphs: a novel algorithm and its application to the integration of diverse biological data types. *BMC bioinformatics*, 15(1):110.

A. Proofs

A.1. Main results

Theorem 1. *Consider the null hypothesis, $H_0^{\mathsf{a},\mathsf{b}}$ in Equation (1). Construct the corresponding null-exposure graph, $G_f^{\mathsf{a},\mathsf{b}}$, and compute a clique decomposition $\mathcal{C}$. Let $C \in \mathcal{C}$ be the unique clique such that $Z^{\text{obs}} \in C$. Then, the randomization test described in Procedure 1 is valid at any level, i.e., the p-value defined in Equation (3) satisfies:*

$$E\left(\mathbb{I}\{\text{pval}(Z^{\text{obs}} \mid \mathcal{C}) \leq \alpha\} \mid \mathcal{C}, H_0^{\mathsf{a},\mathsf{b}}\right) \leq \alpha,$$

where the expectation is with respect to the design, $P(Z^{\text{obs}})$.

Proof. Let Z^{obs} be the observed assignment and C be the clique that contains Z^{obs}. Let $\mathcal{Z}(C)$ denote the set of assignments in C, In the formalism of Basse et al. (2019b), C is the conditioning event of our test. We will use (Basse et al., 2019b, Theorem 1) to prove the validity of our proposed test. This requires to show that the following two conditions hold.

1. Imputability of test statistic: The potential outcomes are imputable within the clique C under $H_0^{\mathsf{a},\mathsf{b}}$, since, by definition, the clique units that are exposed to either a or b for any assignment in the clique. Our test statistic is using only outcomes from units and assignments within the clique, and so the condition in Equation (4) of Basse et al. (2019b, Theorem 1) holds.

2. Correct randomization distribution.: It remains to show that $r(Z) = p(C|Z)$; i.e., that the randomization distribution, $r(Z)$, of our test coincides with the actual conditional distribution, $P(Z \mid C) \propto P(C \mid Z)P(Z)$, induced by the conditioning mechanism $p(C|Z)$ (Basse et al., 2019b, Section 3.2), and the design $P(Z)$. The conditioning mechanism of our procedure is equal to:

$$p(C|Z^{\text{obs}} = z) = \mathbb{I}\{z \in \mathcal{Z}(C)\}, \tag{14}$$

since the test simply conditions on the clique that contains the assignment. The marginal probability of conditioning on clique C is therefore equal to:

$$p(C) = \sum_{z'} p(C|z')P(z') = \sum_{z'} \mathbb{I}\{z' \in \mathcal{Z}(C)\}P(z'). \tag{15}$$

The randomization distribution defined in Step 3 of the testing procedure of Section 4 is equal to:

$$\begin{aligned}
r(z) &= \mathbb{I}\{z \in \mathcal{Z}(C)\}\frac{P(z)}{\sum_{z'} \mathbb{I}\{z' \in \mathcal{Z}(C)\}P(z')} && \text{[By definition, in Step 3 of the clique test]} \\
&= \mathbb{I}\{z \in \mathcal{Z}(C)\}\frac{P(z)}{p(C)} && \text{[by Equation (15)]} \\
&= P(C \mid Z)\frac{P(z)}{p(C)} && \text{[by Equation (14)]} \\
&= P(Z \mid C).
\end{aligned}$$

The validity of our test now follows from (Basse et al., 2019b, Theorem 1). □

Theorem 2. *Consider a null hypothesis, H_0, that is a composite hypothesis comprised of individual contrast hypotheses, $H_0^{\mathsf{a}_j,\mathsf{b}_j}, j = 1, \ldots, J$, for some exposure functions $f_i, i = 1, \ldots, N$, where all $\mathsf{a}_j, \mathsf{b}_j$ are distinct, for any $j = 1, \ldots, J$, with J fixed. Consider the null-exposure graph, $G = (V, E)$, where $V = \mathbb{U} \cup \mathbb{Z}$ is the joint set of units and assignments, and*

$$E = \{(i, z) \in \mathbb{U} \times \mathbb{Z} : f_i(z) \in A_i\}, \tag{16}$$

where $A_i = \{\mathsf{a}_{j'}, \mathsf{b}_{j'}\}$ is the unique exposure set for which $f_i(Z^{\text{obs}}) \in A_i$. Then, the clique test in Procedure 1 operating on the null-exposure graph G is a valid test for H_0.

Proof. The main difference with Theorem 2 is that the test now also conditions on $A = \{A_i : i = 1, \ldots, N\}$, the exposure sets of each unit. In this setting, there is no single clique decomposition. Instead, for every possible value of A there corresponds one clique decomposition, say, $D(A)$. Thus, Equation (14) is updated to:

$$p(C|Z^{\text{obs}} = z) = \mathbb{I}\{C \in D(A_z)\}\mathbb{I}\{z \in \mathcal{Z}(C)\}, \tag{17}$$

where A_z denotes the value of A, which is uniquely determined by the observed assignment z. We will show that the equality

$$\mathbb{I}\{C \in D(A_z)\}\mathbb{I}\{z \in \mathcal{Z}(C)\} = \mathbb{I}\{z \in \mathcal{Z}(C)\}$$

is guaranteed by construction of our null-exposure graph. For that, it suffices to show that $\mathbb{I}\{C \in D(A_z)\} = 0$ implies that $\mathbb{I}\{z \in \mathcal{Z}(C)\} = 0$. We prove by contradiction. Suppose that $\mathbb{I}\{C \in D(A_z)\} = 0$ and $\mathbb{I}\{z \in \mathcal{Z}(C)\} = 1$ for some clique C and observed assignment z. Note that the construction of the null-exposure graph in Equation (16) implies that all units in the null-exposure graph receive exposures contained in A_z. Thus, $\mathbb{I}\{z \in \mathcal{Z}(C)\} = 1$ implies that all units in C receive exposures contained in A_z. However, since the exposures $\mathsf{a}_j, \mathsf{b}_j$ are all distinct, $\mathbb{I}\{C \in D(A_z)\} = 0$ implies that there exists at least one unit in C that receives an exposure that is not in A_z. This is a contradiction. We conclude that:

$$p(C|Z^{\text{obs}} = z) = \mathbb{I}\{z \in \mathcal{Z}(C)\},$$

and so the rest of the proof of Theorem 2 follows.

□

A.2. Proof of result in Section 7.1

The key for the proof is to consider the conditioning event as the pair $(C, \mathbb{Z}_M)$, where $\mathbb{Z}_M$ is the sampled support set, and C is the clique we condition on. The original test in Procedure 1 conditions only on a clique, and $\mathbb{Z}_M = \mathbb{Z}$. We obtain:

$$\begin{aligned} P(Z \mid C, \mathbb{Z}_M) &\propto P(C, \mathbb{Z}_M \mid Z)P(Z) \\ &\propto P(C \mid \mathbb{Z}_M, Z)P(\mathbb{Z}_M \mid Z)P(Z) \\ &\propto \mathbb{I}\{Z \in C\}\mathbb{I}\{C \in \mathbb{Z}_M\}P(Z), \end{aligned} \tag{18}$$

where we used the fact that $P(\mathbb{Z}_M \mid Z)$ is independent of Z by construction of $\mathbb{Z}_M$; also, $P(C \mid \mathbb{Z}_M, Z) = \mathbb{I}\{Z \in C\}\mathbb{I}\{C \in \mathbb{Z}_M\}$ because we simply condition on the clique in $\mathbb{Z}_M$ that assignment Z is contained in. Equation (18) ensures validity of this test if we simply make sure to make a clique decomposition on $\mathbb{Z}_M$, so that $\mathbb{I}\{C \in \mathbb{Z}_M\} = 1$. The remainder terms in the randomization distribution, $\mathbb{I}\{Z \in C\}P(Z)$, correspond to the sampling distribution of Procedure 1 (Step 3).

B. More on clustered interference

Here, we continue our discussion on testing power in Section 5.2. In Figure 11, we reexamine how clique characteristics affect the testing power. The data shown are the same as in Figure 4, now displayed on a single plot with color denoting power. Each dot corresponds to a different clique decomposition of the null exposure graph, and we compute the average number of assignments and units within each decomposition. Requiring more assignments in a clique will make including more units a challenge. Intuitively, this results from the graphical nature of cliques – they are fully connected subgraphs, and including more left nodes will dampen the size of the right node set. This inverse (nonlinear) relationship can be seen on the plot. Note also that there is a balancing of power for different sized cliques. The highest powered tests come from cliques with ~ 150 units and ~ 25 assignments. In practice, this is a tradeoff that can be navigated.

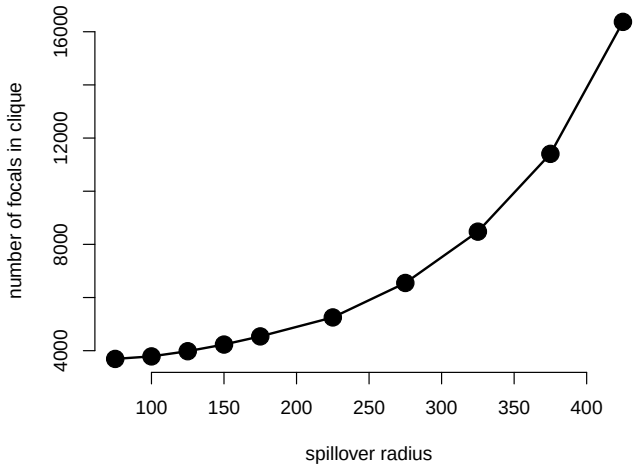

Figure 12. Number of focals versus spillover radius for cliques containing the observed assignment. The radii considered are in the set $\{75, 100, 125, 150, 175, 225, 275, 325, 375, 425\}$. Notice that the number of focals increases nonlinearly as the spillover radius increases.

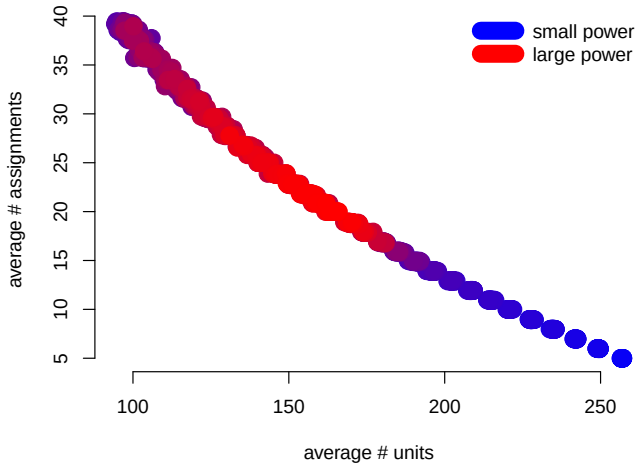

Figure 11. Average number of clique assignments and units versus the power. Each dot corresponds to clique decomposition of the null exposure graph, and the color denotes the power value from the simulation. Red (blue) correspond to large (small) power values.

C. More on spatial interference

Here, we show more information on how properties of the cliques (such as clique size) affect testing power in the context of spatial interference. Figure 12 displays the number of focals contained in the clique for each hypothesis, $H_0^{\mathbf{a},\mathbf{b}_r}$ as a function of r. We see that for larger radii, there are more focals per clique, on average, since more units are exposed to spillovers as the radius gets larger.

Next, we show an extended version of the randomization analysis of Figure 9 shown in Figure 12. First, we show the p-values of the clique randomization test (left vertical axis) with respect to distance radius r, as described earlier ("raw outcome" curve). Second, we show a version of the test where we first regress the crime outcomes on known covariates, including information about the neighborhood and social center points, and then perform the clique test on the residuals ("adjusted outcome" curve). Finally, as a baseline, we also show regression coefficients from a simple OLS model that includes a binary variable indicating whether a unit receives spillovers at distance r or not, and known covariates.

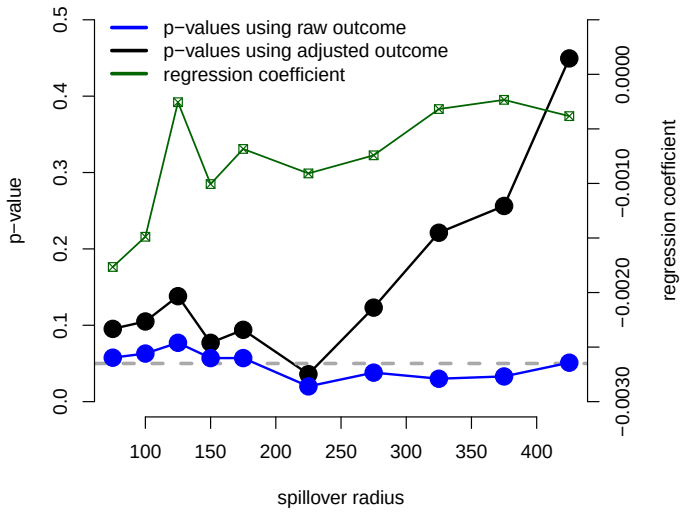

Figure 13. P-values (left vertical axis) for clique tests with varying spillover radii (horizontal axis). The blue line shows p-values for tests using the raw crime index, and the black line shows p-values for tests using the the crime index adjusted for known covariates. The right vertical axis displays regression coefficients on the binary variable defined by spillover or pure control statuses. For each radii, we restrict the OLS estimation to observations such that they are either exposed to spillover or pure control.

We see that the p-values for the raw outcome are all small for varying radii; see the flat blue line. This suggests that some form of spillovers exists, where the distance does not seem to matter. However, the clique test on the adjusted outcomes (black curve in Figure 13) points to the other direction, as it does not indicate significance of spillover effects at any distance. This result suggests that the covariate distributions of "pure control" and "spillover" units are very different, such that the significance of the raw outcome FRTs may be attributed to that difference. The regression coefficient (green curve) agrees with the adjusted outcome results: no regression coefficient is significant at the 0.05 level, and there is a similar, though concave, trend for increasing radii.

Finally, we show here additional information on the Medellín policing experiment. Figure 14 shows the randomization distribution of the test statistics for various radii, r, and for both raw outcomes and adjusted outcomes. We see that the tests with the adjusted outcomes are sharper than the tests with the raw outcomes, indicating heterogeneity in the spillover effects.

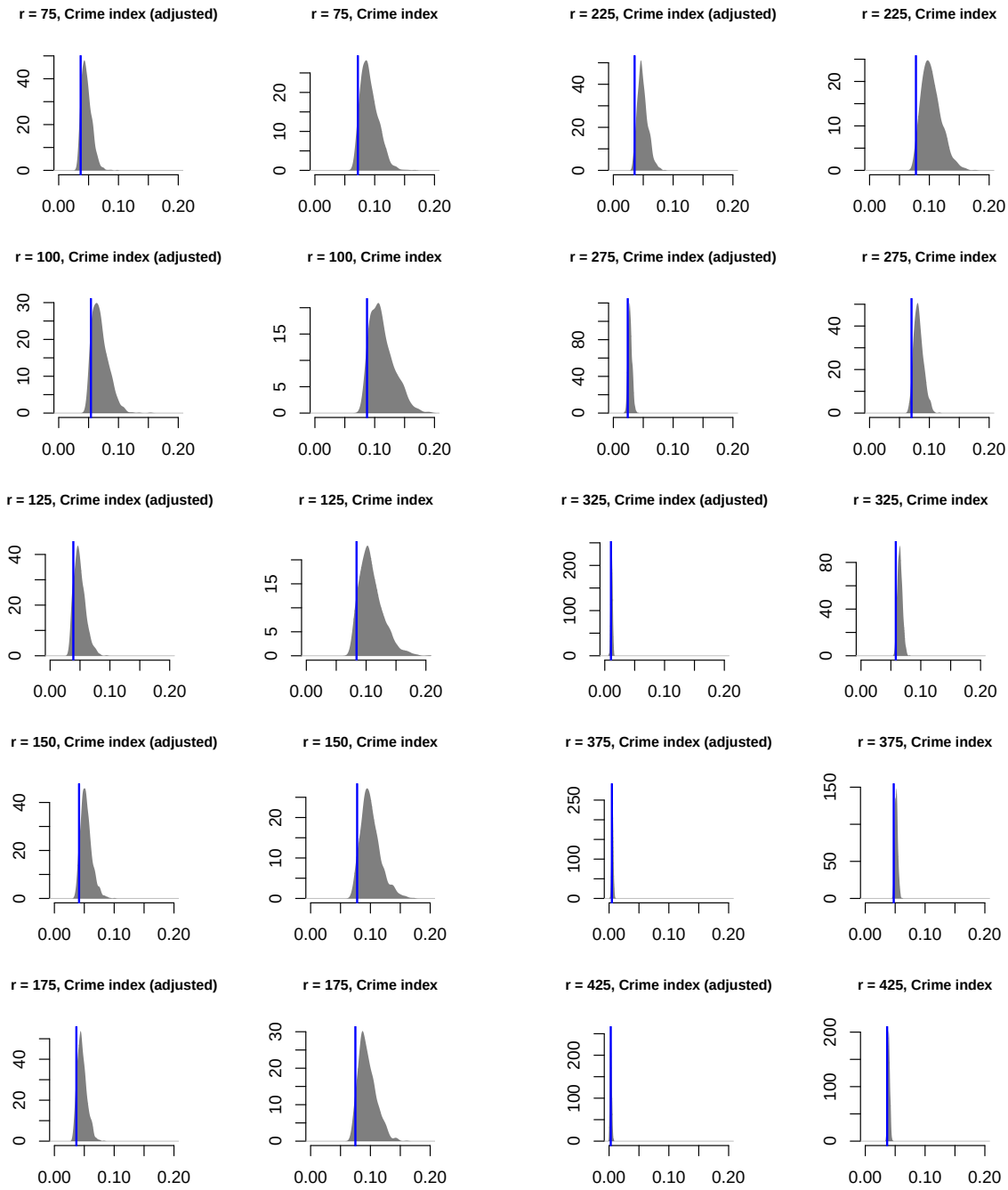

Figure 14. Randomization distributions and observed test statistics (blue) for 20 clique tests. The first and third columns use the adjusted crime index as the outcome, while the second and fourth columns use the raw crime index. The radius defining spillover exposure status varies from 75 meters (top left) to 425 meters (bottom right). Note that the adjusted crime index distributions have lower variance and are centered at smaller positive values compared to their raw index counterparts.