Causal Inference in Heterogeneous Networks

Matthew Simonson*

University of Pennsylvania

October 2021

Abstract

Social influence, peer effects, and spillover are important phenomena in social science, yet they remain difficult to measure, particularly in networks. Social ties induce dependencies between treated and untreated individuals thereby preventing the use of common asymptotic estimators. Recent advances in randomization inference offer a potential remedy, though their effectiveness has been limited by low statistical power. I extend two randomization inference approaches—parameteric and non-parametric—to networks with multiple types of social ties and individuals. In so doing, I restore statistical power while also giving social scientists the tools to determine which types of ties are most effective at transmitting influence. I test these methods though simulations as well as a replication analysis of an intervention aimed at reducing adolescent conflict, examining whether students enrolled in the anti-conflict program reshaped their friends' social norms and whether social cleavages inhibit the spread of social influence.

1 Introduction

1.1 A Motivating Example

How can we measure the spread of attitudes, beliefs, and behaviors in a social network? Causal inference faces unique challenges in a network context due to the ability of individuals to affect one another's outcomes. Consider a randomized control trial testing a program aimed at reducing gun violence among at-risk young men in an urban context. Half the eligible men in a neighborhood are selected at random to attend a violence-diversion program while the rest remain untreated by the intervention. In a non-network context, we could gauge the program's impact by comparing the arrest or victimization rates of the treated to the untreated subjects over the following year. But who in this study is

*I would like to thank Donghee Jo, Jennifer Larson, Cassie McMillan, Stephan McCabe and David Lazer for their support and feedback. This work was supported by the Harry Frank Guggenheim Foundation Dissertation Fellowship.

truly unaffected? Violence is inherently dyadic, and the reduction of violence in one group of people may also reduce violence in those they are used to fighting with. Thus, naively comparing the outcomes of the treatment and control groups might fail to turn up much of a gap, leading us to underestimate the program's impact. We might address this problem by having another group of young men in a different community serve as a control group. However, if this group is too close geographically, they might have friends of friends in the treatment group and thus indirectly feel its effects. If we choose a more distinct community in a different city, that city might be experiencing its own crime increase or decrease already, masking any difference in the two groups. And what if the spillover itself is the object of interest? If an untreated man refrains from violence, shall we attribute it to persuasion from a treated best friend, exposure to the behavior of treated acquaintances, or the disengagement of a treated enemy?

1.2 Problem and Objectives

While it may seem obvious that different types of relationships such as friends, acquaintances, and enemies will exert different levels of influence, this variation has been marginalized in the causal inference literature, in both a normative and a statistical sense. Scholars of causal inference frequently invoke assumptions that preclude heterogeneous relationships and individuals in order to simplify computation and construct unbiased estimators of quantities of interest (see Aronow and Samii 2017). In so doing, however, they sacrifice information which could allow one to determine a) whether any peer influence has occurred, and b) which individuals or relationships are most likely to influence others. These are important policy questions in some contexts, no less than determining average treatment effects. The aim of this study, therefore, is to demonstrate how heterogeneity in individuals and relationships can be leveraged to answer these questions.

1.3 History, Literature, and Nomenclature

Causal inference on networks has come a long way since Christakis and Fowler (2007) published their seminal study on the spread of obesity through a network. Their longitudinal analysis of patients in the Framingham Heart Study drew criticism for arguing that friends in the study experienced health outcomes such as weight loss or gain due to social influence. Critics argued that without a randomized or natural experiment, this research strategy failed to rule out the possibility of homophily—that people with similar hobbies, like exercise, may be more likely to become friends. Another threat to identification arises from the possibility the people linked by social ties may be more likely to be exposed to the same external stimuli. For instance, friends living near each other both experience the opening of a new McDonald's or gym (see Cohen-Cole and Fletcher 2008; Lyons 2011; Shalizi and Thomas 2011). Nevertheless, the obesity social contagion study and the attention it drew helped spark a surge of interest in determining the correct way to measure social influence

experimentally (e.g., Bond et al. 2012; Sinclair et al. 2012; VanderWeele 2011). At the same time, applied econometricians and statisticians who had spent years trying to clarify and ultimately relax their strict assumptions of causal identification began to turn their attention to settings where the subjects of experiments interfere with one another's outcomes (e.g., Hudgens and Halloran 2008; Manski 2013; Sobel 2006; Tchetgen Tchetgen and VanderWeele 2012). With the coming together of these research traditions, the past five years have witnessed a flourishing of interdisciplinary collaboration by biostatisticians, economists, sociologists, psychologists, computer scientists, statisticians, and political scientists on problems related to causal inference in networks (see Aronow et al. 2020, for an overview).

One of the byproducts of this confluence has been a bewildering alphabet soup of notation and nomenclature. In this article, I shall use *nodes* to refer to individuals, subjects, or experimental units; *alters* to refer to their friends, neighbors, or peers; and *edges* to refer to the links, ties, or relationships they share. Since *group* typically refers to people receiving the same treatment (e.g., treatment group, control group), I use *category* to refer to social groups such as gender, race, or tribe. Despite a long tradition in sociology of using "strength" to refer to the depth, power, or capacity of a relationship (Granovetter 1973), strength has another meaning in network science, so I shall instead refer to *weight* which is less ambiguous (except, perhaps, in the context of obesity contagion studies). The sets of nodes and edges together constitute a *network*. Spatial relationships can be represented by drawing weighted edges between homes, cities, and so forth. People found in households or classrooms where everyone knows everyone else can be represented by clusters of nodes which are fully connected internally and with few edges—or only low weight edges—between clusters.

When referring to that which is (perhaps) spreading across the network, I shall follow the tradition of political scientists who use *influence* or *spillover* (the latter, more often, in the context of an experiment). While I will sometimes use these terms interchangeably in a statistical context, the distinction between them is worth pausing to explore. Consider an impoverished middle school student who receives access to free breakfasts at school but does not take up this treatment for fear of being stigmatized. This student is clearly being influenced by his peers, but the treatment itself is not "spilling over," unless he shares that breakfast with someone else. A strand of the economics literature prefers the term *peer effects*, particularly in educational contexts. For example, in peer encouragement designs, the experimenter may encourage the formation of new network edges ex nihilo by randomly assigning students to the same dorm room (Sacerdote 2001) or study group (Carrell et al. 2013). In the present study, however, it is the assignment of treatment that is randomized, not the network itself (cf. Centola 2010; Salganik 2006). Likewise, although An (2018) and others have made modest headway studying inference in dynamic networks, here I will make the simplifying assumption that the network itself remains unchanged over the course of the study.

A related branch of research focuses on modeling the diffusion, propagation, or

contagion of rumors, information, behaviors, or disease over time (e.g., Banerjee et al. 2019; Larson and Lewis 2017). Compared to the spillover and interference literature, these scholars are much further advanced at incorporating node and edge heterogeneity into their predictive models (e.g., Sun et al. 2021). Although the framework of this study assumes that one has but a single snapshot of a network following treatment, the techniques presented here may prove useful to epidemiologists testing basic model assumptions and calculating model parameters. While the spillover literature often assumes that treatment effects attenuate as they get further from the source, there may be cases in which treatment effects continue to ripple through a network after the study's conclusion. Finally, the term *interference* is often applied to all these situations (as well as to studies where the nodes are animals, institutions, or vegetable plots), but the resemblance of the words *inference* and *interference* presents a potential source of confusion.[1]

1.4 Current Approaches

In spite of these fruitful cross-disciplinary collaborations, significant challenges remain. Most of the recent work on causal identification in networks has focused on developing unbiased estimators of quantities of policy interest such as average treatment effects (e.g., Aronow and Samii 2017). Although this approach holds promise, most of these methods to date require strong assumptions that are difficult to justify, particular in finite samples. Furthermore, one of the major justifications for conducting an intervention in a complex network is to witness complex effects that cannot be observed in other settings. For instance, a common assumption is that there are no higher order treatment effects—that is, nodes cannot influence anyone beyond their immediate alters. If first-order spillover is all we are interested in, and we assume all edges are identical (another common assumption), we could have saved ourselves a lot of statistical trouble by conducting an experiment on classrooms or households with no connections between them (e.g., Nickerson 2008).

Instead, this study focuses on extending recent advances in randomization inference, a technique first developed by Fisher (1935), to networks of heterogeneous nodes and edges. Randomization inference obviates the need for simplifying assumptions about the error distribution, thanks to Monte Carlo simulations of the treatment assignment process that actually took place. The p-values obtained in this way are exact—one can get arbitrarily close to the true p-value through repeated simulations. They are also said to be valid in the sense that a p-value of 0.05 will allow for false positives no more than 5% of the time. Fisher's original framework required a sharp null of no effects on

[1] Three reviews of the literature identify still more distinctions. Sacerdote (2011) states that class size or increased market demand constitute interference but not peer effects, since they do not flow directly from any particular peers but operate through externalities. Aronow et al. (2020) likewise considers contagion to be a subclass of interference in which the outcome spreads but the treatment itself does not. Manski (1993) lists a dozen other terms including herd behavior and imitation, all of which he groups under the heading of endogenous social effects.

any node, not merely no effect on average. Some practitioners criticize this approach as yielding results that are uninteresting from a policy perspective, since the test does not indicate which individuals experienced effects or even what direction they were in. However, many scholars contend that exact nulls can be quite informative (Xu and Basse 2021), while others are working to extend randomization inference to cover less stringent null hypotheses (Caughey et al. 2021; Wu and Ding 2020; Zhao and Ding 2021). In this study, I explore two strategies for randomization inference in networks: one which takes a purist approach to randomization without additional assumptions, and the other which adds parametric assumptions in order to estimate effect sizes and confidence intervals.

1.5 New Contributions

The randomization approach I discuss first—initially developed by Aronow (2012)—divides the sample into individuals eligible for treatment, but whose outcomes will not be measured, and nodes whose outcomes we measure but are not eligible to be treated. In other contexts, the nodes randomized for treatment/control and those whose outcomes we study are one and the same, so this new wrinkle reduces the effective sample size. Edge weight, however, offers a potential remedy. By using edge weight to guide the selection of which nodes to measure and which to randomize, we may be able to minimize this loss. This framework can further be strengthened through the development of test statistics that take edge weight into account in order to be more sensitive to spillover. I also suggest new null hypotheses and randomization techniques to test for spillover within and between categories. I call this framework the *nonparametric randomization approach* because it does not attempt to estimate model parameters, including effect sizes. The aim is merely to show whether or not a particular type of spillover occurs.

Conversely, we can use spillover to make inferences about the relative influence of different nodes and edges. First, we might ask if certain types of nodes exert more influence than others. For instance, social media users with a high number of followers often style themselves as "influencers." Do these *social referents*—nodes that others look to for appropriate behavior—actually have a larger impact than other users, and if so, do they have more influence on a per follower basis or only because they have a large audience? Second, we may have reason to think that certain edges fall into a strong or weak category, even if we don't have a number to quantify it. Are relationships that are "strong" in terms of emotional depth or frequency of contact actually more influential at recruiting someone for a cause or getting them to provide assistance)? Third, are cross-cleavage edges that connect members of two social categories—such as race, tribe, or gender—across a political or social cleavage less influential? Scholars have argued that the relative weakness of such edges at conveying information may increase the risk of conflict and undermine the provision of public goods (Habyarimana et al. 2007; Larson 2016). To address these questions, we need to estimate the relative magnitudes of spillover effects from different nodes and edges, not merely

show they are unlikely to equal zero. Doing so requires additional assumptions about spillover mechanisms and their functional form, though still within the context of randomization inference on a sharp null hypothesis. Thus, in Section 4, I extend what I call the *parametric randomization approach*—first developed by Bowers et al. (2013)—to quantify and compare the influence of different types of edges and nodes.

2 Setup and Notation

2.1 Weighted Networks

Before diving into a technical discussion about how edge weights can be used to improve inference, it is worth pausing to consider what such numbers might represent and how a researcher might collect such data. Weights can represent any sort of quantity that provides useful information about the relationship between two nodes. For instance, imagine a spatial model in which we filled in every cell in the matrix to represent the distance between two locations, as is seen in the legends of some inter-city road maps and travel guides. This sort of representation would be useful if we thought that every node had a possibility of influencing any other node directly and that distance was a good proxy for the likelihood of such influence. For instance, if we were to study a group of homeowners scattered across a town, we might use distance between their homes as a proxy for the likelihood that they knew each other or were likely to run into one another. Although both models represent networks derived from spatial relations, the key distinction here is that a traveler driving from one town to another must pass through all the intermediate towns along his route, while two homeowners need not know the neighbors who live between them. Thus, in addition to providing more information than an unweighted network, weighted networks can capture phenomena that unweighted networks cannot effectively represent. To model homeowners' relationships in this fashion, an unweighted network would either have to fill in every cell in the matrix with a "yes," at which point it would cease to tell us anything, or we would have to pick an arbitrary threshold a step beyond which any chance of relationship was ignored and within which it was uniformly a certainty.

Understanding unweighted networks as binary realizations of underlying weighted relationships shines a light on important issues of construct validity a researcher might otherwise miss. For instance, we might assume friendship is, by definition, a symmetric relationship: If I think of you as a friend, but you disagree, then we are not friends. However, suppose we ask a class of middle school girls to tell us who their friends are. Beth lists Anna as a friend, but Anna does not reciprocate. A researcher who has studied middle school students (or been one) will not find this asymmetry particularly surprising; perhaps Anna is mean, or she is a popular girl with a close group of friends and considers Beth an acquaintance. We can imagine that everyone, perhaps unconsciously, has a latent scale on which they rate the strength of their relationships, and, in Anna's

eyes, her relationship with Beth falls below the threshold for friendship. But what if the girls' thresholds are heterogeneous? For instance, Anna could be a rather private person who has a high threshold for how close she has to feel to someone before she considers them a friend. Thus, it is possible that Anna and Beth gave each other the same latent rating (7), but Beth's threshold is 5 and Anna's is 8. Now suppose there is another girl Claire with the same threshold as Beth (5), and that she and Beth both rate each other as 6, meaning they consider one another friends. All three girls would agree, if their preferences were made known, that Anna and Beth have a stronger relationship than Beth and Claire, yet Beth and Claire will show up as friends in our survey while Anna and Beth will not. What is a researcher to do? We could code all non-reciprocated friendships as friends, but then we would pick up pairs whose latent scores really are wildly different (for instance, if Claire gave Anna 10, but Anna gave Claire 1). We could simply allow friendships to be asymmetric, but we'd still be missing a directed tie from Anna to Beth that is far stronger than the edges between Beth and Claire that we do record.

Weighted edges offer us a way out of this conundrum, assuming we had a question that could capture them accurately. This raises an important question of construct validity, however—what aspect of friendship is actually relevant to our research question? If we are trying to measure the spread of a rumor, we might want to ask these girls whom they talk to the most (Banerjee et al. 2019). If our intervention is one of attitude change, we might ask them whom they look up to (Paluck 2011). Of course, we might not know ahead of time what would be the best question for a particular type of influence. Ognyanova (2020) finds that students are more likely to gain political knowledge from people with whom they report frequently socializing, not those they report frequently talking to about politics. The weights and directions of these edges will likely be relevant to the spillover of the intervention and moreover they minimize the paradoxical problems raised by filtering latent scales through subjective thresholds.

Of course, the scales could be heterogeneous as well (what Anna considers frequent communication Claire considers rare), but we can try to standardize them through precise definitions in our questions. For instance, a researcher might define "knew someone" in the text of a survey question as "someone whom you would have greeted on the street by name if you saw them." Conversely, rather than breaking edges down into multiple types, we can also aggregate multiple binary network questions to proxy for edge weight. Banerjee et al. (2013) and Larson and Lewis (2017) report networks between households created from aggregating over 12 and 7 edge questions respectively, but they treat their networks as unweighted. An alternative approach might be to count how many edges each household shared and assign weights accordingly, if the researchers believed edge weight might be relevant to the spread of information.

2.2 Network Notation

Let $\mathcal{V}$ be a set of n nodes and let G be an $n \times n$ matrix (called an *adjacency matrix* or *sociomatrix*) defined such that the entry G_{ij} indicates the *weight of an edge*

running from node i to node j (this relationship need not be reciprocated). In an unweighted context $G_{ij} \in \{0, 1\}$ where $G_{ij} = 0$ indicates the absence of any edge and $G_{ij} = 1$ merely implies that one exists. In general, however, G_{ij} can be any real number, including a negative number which would indicate a countervailing edge that somehow offsets a positively weighted edge of equal magnitude. We define the *weight of node* i as the sum of its edges weight $w_i = \sum_j G_{ij}$ and the *degree* of i as the number of edges it has: $d_i = \sum_j \mathbb{I}(G_{ij} \neq 0)$ where $\mathbb{I}$ is the indicator function mapping true statements to 1 and false statements to 0. In an unweighted network, node degree and weight are identical.

2.3 Potential Outcomes Notation

Let $Z_i \in \{1, 0\}$ be the treatment assigned to node i and let the treatment vector $Z \in \{1, 0\}^n$ represent the assignments for all the nodes in the network. We will restrict ourselves to binary treatments for pedagogical purposes. Let Y_i^{base} be the baseline outcome that node i would have exhibited had no one in the network been treated, and let the Y_i^{obs} be the outcome which was actually observed. Under the Neyman-Rubin potential outcome model (Imbens and Rubin 2015; Neyman 1923), each node is said to have a pre-ordained outcome for each level of treatment it could potentially receive. Thus, we can represent the vector of outcomes Y as a function of the treatment vector: $Y(Z)$. The outcomes $Y_i(\cdot)$ for a particular node i might depend not only on i's treatment assignment Z_i but also on the assignments of any other node in the network, yielding 2^n potential outcomes. In order to make inference tractable, we must start with at least some simplifying assumptions. For example, the asymptotic approach discussed in Section 1.4 (see Aronow and Samii 2017) makes stringent—sometimes implausible—assumptions in order to map $Y(Z_i)$ to only a handful of potential outcomes. The researcher then proceeds to invoke various estimators of different sorts of average effect (treatment compared to no treatment, treatment with spillover compared to treatment without spillover, etc.).

2.4 Null Hypothesis Significance Testing

Our approach is rather to test how plausible such an assumption is. If we find that the joint distribution of treatment assignments and outcomes is extremely unlikely under a given assumption, then we reject it and move on to test a new assumption in its place. We call such an assumption a null hypothesis. Adapting the language of Athey et al. (2018), hereafter AEI, we define a null hypothesis for a set of treatment vectors $\mathcal{Z}_{H_0}$ on a set of nodes $\mathcal{V}_{H_0}$ for a particular outcome variable $Y(Z)$ as follows:

Definition 1. *A null hypothesis* $H_0 : \{\mathcal{Z}_{H_0}, \mathcal{V}_{H_0}\}$ *on* $Y(Z)$ *is any set of assumptions restricting* $Y(Z)$.

Normally we may be used to thinking of a null hypothesis as being defined on the whole population $\mathcal{Z}$, but one can also study the null of "average treatment effects on the treated are zero" in which $\mathcal{Z}_{H_0} = \mathcal{Z}^{tr}$. What is particularly

unusual here is to consider the set of assignment vectors as well, which need not include all possible assignments.

The original *sharp* null proposed by Fisher was for "no treatment effect on any node under any assignment vector." This makes i's unseen potential outcomes easy to impute: they are all equal to Y_i^{base} including the one we observed: Y^{obs}. Thus, the outcome we observed is the one we would have seen under any other assignment vector Z. Here we generalize this idea to include other sharp nulls, defined on a particular of set of nodes $i \in \mathcal{V}_{H_0}$ and a particular set of treatment vectors $Z \in \mathcal{Z}_{H_0}$. Specifically:

Definition 2. *A sharp null* $H_0 : \{\mathcal{Z}_{H_0}, \mathcal{V}_{H_0}\}$ *on* $Y(Z)$ *is any set of assumptions restricting* $Y(Z)$ *such that for any* $Z \in \mathcal{Z}_{H_0}$ *we can impute* Z_i *for any other* $i \in \mathcal{V}_{H_0}$ *under any other* $Z' \in \mathcal{Z}_{H_0}$.

2.5 Sharp Nulls in Artificial Experiments

The key to the nonparametric approach is to carefully choose a subset of nodes $\mathcal{V}_{H_0}$ and a subset of nodes $\mathcal{Z}_{H_0}$ to set up a sharp null hypothesis of interest.

For instance, suppose we want to test a sharp null hypothesis of no spillover effects onto any node, anywhere in the network, under any circumstance. This would be very hard to prove, since we'd have to run through every possible treatment vector, but if we could just show that there was spillover on some small subset of nodes for some small subset of assignment vectors, then the whole hypothesis would be disproven. Statistics does not allow for such ironclad certainties, but it does allow for the next best thing: rendering a hypothesis highly implausible. Thus, analogously, if we can render this hypothesis only 5% plausible for some subset of nodes under some subset of assignment vectors, then this hypothesis would be only 5% plausible (or less) for the network as a whole. In sum, null hypotheses on subsets are valuable because rejecting them allows us to reject the original null for the full network.

Here is one way testing a sharp null on a subset of nodes and assignment vectors can work. Before running the experiment, designate half the nodes as eligible for treatment and reserve the other half for measuring outcomes. We'll refer to those eligible for treatment as the *randomization set* $\mathcal{R}$ and those we intend to measure outcomes in as the *measurement set* $\mathcal{M}$.[2] Since the nodes in $\mathcal{M}$ are not allowed to be treated, we will restrict our set of admissible treatment assignment vectors from all possible vectors in $\mathcal{Z}$ to only those which affect the randomization nodes. We'll call this subset set $\mathcal{Z}_{\mathcal{R}}$. If there really were no spillovers anywhere in the network, then the outcomes of the nodes in $\mathcal{M}$ should be unaffected when we re-shuffle the treatment assignment of the nodes in $\mathcal{R}$. Therefore, if the null hypothesis is true, then for any $i \in \mathcal{M}$ and any $Z \in \mathcal{Z}_{\mathcal{R}}$, we obtain $Y_i(Z) = Y_i^{obs}$. Thus, from the observed outcomes we can impute all other potential outcomes in our subset of nodes (for that particular subset of

[2] Aronow (2012) calls these sets variant and fixed respectively; AEI refer to them as auxiliary and focal. I prefer to name these sets after their intended purpose to make them easier to keep track of.

treatments). Hence, the null hypothesis is sharp for these subsets. AEI refer to this setup as an "artificial experiment" and point out that you can still set one up even after the real experiment has taken place. Simply choose a subset of treatment vectors $\mathcal{Z}_\mathcal{R}$ that preserves nodes in $\mathcal{M}$ with their observed treatment statuses, be they treatment or control. For details on how to use edge weights to optimize set selection, see Appendix A.

Now comes the statistical proof-by-contradiction (or more accurately, almost-proof-by-almost-contraction.) We select a test statistic we expect will be sensitive to violations of H_0. For that we need a statistic that measures the relationships between the outcomes of the measurement nodes $Y_\mathcal{M}^{obs}$ and the treatment statuses of the randomization nodes in $Z_\mathcal{R}^{obs}$. For instance, we could use the correlation between $Y_\mathcal{M}^{obs}$ and the number of treated neighbors each node in $\mathcal{M}$ has. If there is enough spillover, we should see a strong correlation. In contrast, if we randomly scramble the treatment status of the randomization nodes while keeping $Y_\mathcal{M}^{obs}$ the same, the result should be a weak (if any) correlation arising merely due to chance. Scrambling the treatment statuses of nodes in $\mathcal{R}$ is equivalent to drawing another hypothetical treatment vector $Z_\mathcal{R}^{hyp}$ from $\mathcal{Z}_\mathcal{R}$, and under our sharp null hypothesis, we know that $Y_\mathcal{M}^{hyp} = Y_\mathcal{M}^{obs}$. Thus, maintaining our (potentially faulty) premise that there are no spillover effects, we can check what the correction between $Y_\mathcal{M}^{hyp}$ and the number of treated neighbors would be under any realization of the experiment. Run 1000 such simulations, selecting $Z_\mathcal{R}^{hyp}$ randomly from $\mathcal{Z}_\mathcal{R}$.[3] If fewer than 5% of those produce correlations as strong as the one we observed, then we can reject the null as implausible not only for this special subset but for the entire network. Analogously, if we were told that pandas had gone extinct, we would not have to search the world to render this claim highly doubtful—we could limit our search to China. If we spotted what looked like a panda and could say with 95% certainty that's what we saw, then rumors of the panda's demise would be thrown into serious doubt not only in China but as a general statement about their existence in the world.

By the same token, finding strong evidence of spillover from $\mathcal{R}$ onto $\mathcal{M}$ does not mean there will be spillover effects between any two subsets. It just means that our original hypothesis of "no spillover anywhere" was false. As a result, $\mathcal{M}$ does not need to be a representative sample of $\mathcal{V}$, but can be chosen strategically. That is, we can look for pandas where they are most likely to be found.

[3] If the original randomization scheme in the experiment was blocked or has some other feature that made some realizations of $Z_\mathcal{R}^{hyp}$ more likely, then we must set up a script to draw hypothetical treatment vectors according to this distribution. Under normal circumstances, however, it would be sufficient to permute the treatment values. Under ideal circumstances, the researcher would simply rerun whatever lines of code were used to assign treatment in the first place. This becomes more complex, however, when the admissible treatment vectors in the artificial experiment constitute a strict subset of the original $\mathcal{Z}$ because the original experiment did not leave aside measurement nodes or because one wishes to test an additional null hypothesis after successfully rejecting the first one.

3 New Null Hypotheses

3.1 Multimodal Networks

Suppose the network is partitioned into subsets of individuals which we shall call categories. These categories could be based on a political or social cleavage or they could simply represent nodes that are likely to interact differently based on their role or function. While studies of school networks typically include only the students, we could imagine a study incorporating the teachers as well. Teachers may be more likely to influence students than be influenced by them. Similarly, Black and White students may have less influence across racial categories than within them. Of course, influence may be dependent on the type of spillover. Girls and boys may be unlikely to influence each other when it comes to fashion trends, but just as likely to exert influence across genders as within them when it comes to a new app or technology.

Categories can be indexed by their members such that C_i refers to the group that i belongs to and $C_i = C_j$ means that i and j are members of the same category. Given a treatment vector z, the notation z_{C_i} refers to the subvector of z containing the treatment assignments of C_i's members. We are now ready to define our first null hypothesis:

Hypothesis 1 (No Cross-Category Spillovers)**.** *Let $i \in \mathcal{V}$ and let C_i be i's category. Then for any treatment assignment vectors z and z', if $z_{C_i} = {z'}_{C_i}$ then $Y_i(z) = Y_i(z')$.*

In other words, so long as the assignment of the members of i's group remain undisturbed, i's outcome will remain the same, regardless of what happens to everyone else in the network.

To test the no cross-category spillover hypothesis, we can break it down into sub-hypotheses for each category. To test whether a particular category C_i is impervious to outside influence, simply set $\mathcal{M} = C_i$ and $\mathcal{R} = \mathcal{V} \setminus C_i$ (that is, any nodes not in C_i). The results of testing any of one of these sub-hypotheses may be interesting in its own right (for instance, rejecting the hypothesis that students have influence on their teachers), or we may be more interested in aggregating the results. See Vovk and Wang (2019) for suggestions about how to safely aggregate p-values.

Another hypothesis of interest might be whether there are spillovers within a group.

Hypothesis 2 (No Intra-Category Spillovers)**.** *Let $i \in \mathcal{V}$ and let C_i be i's category. Then for any treatment assignment vectors z and z', if $z_i = z'_i$ and $z_{\mathcal{V} \setminus C_i} = {z'}_{\mathcal{V} \setminus C_i}$ then $Y_i(z) = Y_i(z')$.*

This hypothesis is more difficult to test since i is now within the category we are permitted to manipulate. Once again, we begin by breaking down the larger hypothesis into sub-hypotheses and testing each one on its own. This time, we can essentially discard the nodes outside of C_i since they cannot be randomized but also are not relevant to measurement. In essence, we now must conduct a

standard null hypothesis of "no spillovers" within C_i as if C_i were the entire network.

3.2 Multiplex Networks

In addition to multiple types of nodes, networks can also have multiple types of edges. For instance, in Paluck, Shepherd, and Aronow's (2016) anti-conflict norms experiment on 56 middle schools, students listed up to ten other students they spend time with and up to two "best friends." If we assume that students are likely to start by listing students they feel closer to (or who are more salient to them and easy to remember), then we might use the order of these names as weights. However, we may be better off keeping these distinctions categorical if we don't know the function that maps these weights to influence. For instance, as we go down a list of names, does influence decrease linearly or with $f(x) = 1/x$? Is a best friend twice or three times as influential as any other person one spends time with? Thus, even in cases where an edge level variable is ordered, we may be best off treating it categorical rather than continuous. Furthermore, nodes can share multiple overlapping ties. As previously mentioned, Larson and Lewis (2017) measured seven types of edges with separate questions. Rather than aggregating these edges into a single network, or turning the number of edges between two nodes into weights, we can test which edge type is best suited to transmit spillover.

Even higher-order spillover from the alters of one's alters can be seen as a multiplex network. For instance, to capture the effects of nodes two steps away, one can construct a new adjacency matrix G' such that $G'_{ij} = \sum_h G_{ih}G_{hk}$. We could set all edges in G' to equal 1, or we could allow G' to encode the number of paths between the two nodes, since presumably the amount of influence will be related to the number of paths there are for influence to travel between an origin and destination. Interestingly, if G is weighted and $0 < G_{ih}, G_{hj} < 1$, then G'_{ij} will have an even smaller value than either G_{ih} or G_{hj}. One way to interpret this fact is that each edge in G corresponds to the probability of influence flowing from one node to the next, with the attrition of influence increasing over longer paths.

Once we have two networks, we can test the null hypothesis outlined in AEI that after accounting for the edges in one network, we will see no influence traveling across the other. Formally, this is stated:

Hypothesis 3 (No Spillovers via an Alternate Network)**.** *For any nodes i, j such that $G_{ij} \neq 0$, and any two treatment assignment vectors z and z', if $z_j = z'_j$ then $Y_i(z) = Y_i(z')$.*

That is, as long as none of i's alters in network G change their treatment assignment, then i's outcome will be unaffected. To test this null hypothesis, we must bear in mind that if i is a measurement node, then none of i's neighbors in the original network can be randomized. Some of them may fall into $\mathcal{M}$, while others will be part of the third set of so-called buffer nodes excluded from measurement and randomization. If nodes existing in tight-knit clusters with

identical connections to the outside, the third category may not be necessary, but most likely we will need a pretty large network to obtain a sufficient effective sample size. See AEI for suggestions about set selection.

4 Estimating Effect Magnitudes with Parametric Randomization

As discussed in Section 1.5, Bowers et al. (2013) propose a parametric randomization approach to estimating the magnitude of spillover effects. In exchange for making assumptions about the functional form of the influence mechanism, we can obtain confidence intervals while still using randomization inference and a sharp null. In this section, I demonstrate for the first time how AEI's nonparametric approach to using measurement and randomization sets can be combined with Bowers, Fredrickson, and Panagopoulos's (2013) parametric approach to estimate relative weight of different edge types and categories.

4.1 Estimating One Parameter

One begins by thinking about the mechanism through which spillover operates and making an educated guess as to its functional form. Ideally, this functional form should be informed by prior studies or by the rejection of other forms using the nonparametric approach. For instance, we might assume that spillover involves the number of treated alters and that the functional form is linear. Thus, if τ is the spillover effect, then

$$Y_i(z) = Y_i^{base} + \tau \sum_j G_{ij} z_j$$

We can then transform this into a formula for recovering the baseline values of Y:

$$Y_i(z) - \tau \sum_j G_{ij} z_j = Y_i^{base}$$

Since we can transform any two realizations of Y_i to the Y_i^{base}, we can also transform them into each other:

$$Y_i^{hyp}(z^{hyp}) - \tau \sum_j G_{ij} z_j^{hyp} = Y_i^{obs}(z^{obs}) - \tau \sum_j G_{ij} z_j^{obs}$$

And finally, we can recover the values of any hypothetical Y_i^{hyp} given our observed values and the hypothetical assignment vector we have randomly drawn:

$$Y_i^{hyp}\left(z^{hyp}\right) = Y_i^{obs}\left(z^{obs}\right) + \tau \sum_j G_{ij} \left(z_j^{hyp} - z_j^{obs}\right)$$

For any value of τ, we can now impute any values of Y. The null hypothesis of "spillover effects equal to τ" is, therefore, sharp for every node in $\mathcal{M}$ and every treatment assignment vector that leaves the treatment values $z_{\mathcal{M}}$ unchanged. The price of admission to this new set of null hypotheses was assuming that the spillover mechanism follows this particular functional form. Thus, we are required to draw on our substantive knowledge of the phenomenon we are studying to come up with an appropriate influence mechanism, but we leave it to hypothesis testing to parameterize it.

Our first step toward estimating treatment effect is to set $\tau = 0$, thus implying that the usual null hypotheses of no treatment effects: all potential outcomes of Y under any z can be imputed because they are equal to Y_i^{base} and hence to the value Y_i^{obs} that we actually observed. Our next null hypothesis sets $\tau = 0.1$, then 0.2, etc. Once we have tested an arbitrary sequence of values, we reject those with $p < 0.05$ or some other threshold and accept those with larger p-values as defining our confidence interval. Bowers et al. (2013) suggest accepting the τ with the largest p-value as our point estimate.

4.2 Estimating Two Parameters

Bowers et al. (2013) do not restrict us to using separate measurement and randomization sets as Aronow (2012) and AEI do. As a result, Bowers et al. (2013) are forced to include a second parameter in their model, μ, that picks up direct treatment effects since some nodes will have received treatment, as will everyone else after enough randomizations. By limiting our sharp null to a separate measurement set without any treated nodes, we can avoid needing to test out values for both τ and μ and focus solely on τ, our object of primary interest. This restriction also frees us up to have a second parameter represent something else. For instance, the young men in the opening example of this article were embedded in a network of friends and a network of enemies, but we don't know which type of relationship is stronger. Rather than guessing at the weights and naïvely trying to combine them, we can model these as separate layers of a multiplex network. Assuming friend and foe effects τ and μ are linear, our model looks like:

$$Y_i(z) = Y_i^{base} + \tau \sum_j G_{ij}^{friend} z_j + \mu \sum_j G_{ij}^{foe} z_j$$

And our final equation becomes

$$Y_i^{hyp}\left(z^{hyp}\right) = Y_i^{obs}\left(z^{obs}\right) + \tau \sum_j G_{ij}^{friend}\left(z_j^{hyp} - z_j^{obs}\right) + \mu \sum_j G_{ij}^{foe}\left(z_j^{hyp} - z_j^{obs}\right)$$

By plugging in different values for both τ, the friend effect, and μ, the foe effect, we can estimate a confidence region across both parameters jointly. We can now answer questions not only about networks that are inherently multilayered, but also about edges we would normally conceive of as belonging to the same network. For instance, we could set up an experiment in which treated nodes

are induced to call on alters for a favor. If we treated edges between people of different religions or ethnic groups as different types, then we can test whether the two have the same impact or if edges within a category are more effective at obtaining small favors. To test whether a certain type of people such as social referents have a strong impact, we can simply treat all edges coming from a social referent as a tie type.

5 Test Statistics

In this section, I review the test statistics proposed by various authors and show how each one can be generalized to incorporate edge weights (weighted networks) and multiple types of edges (multiplex networks).

5.1 Edge-Level Contrast

First proposed by Bond et al. (2012) and corrected by AEI, the edge-level contrast statistic is constructed as follows: assign to each cross-set edge the outcome of its measurement node and the treatment status of its randomization node. For example, if an edge connects a treated randomization node to a measurement node with outcome $Y(i) = 0.5$, then we call that a treated edge with an outcome 0.5. The edge-level contrast statistic is then the difference in means between the treated edges and control edges.

$$T_{elc} = \frac{\sum_{i\in\mathcal{M},j\in\mathcal{R}} G_{ij}Y_i^{obs}z_j}{\sum_{i\in\mathcal{M},j\in\mathcal{R}}, G_{ij}z_j} - \frac{\sum_{i\in\mathcal{M},j\in\mathcal{R}} G_{ij}Y_i^{obs}(1-z_j)}{\sum_{i\in\mathcal{M},j\in\mathcal{R}}, G_{ij}(1-z_j)} \tag{1}$$

In a weighted context, the formula remains the same; the entries in the identity matrix G are no longer just 1s and 0s, and this mean automatically becomes a weighted mean. If we are dealing with multiple edge types, we might start by making an educated guess as to what the ratio of their effectiveness is when it comes to spillover and specific weights accordingly. Misspecifying these weights won't affect the validity of the test, but it could affect power. For instance, if you've made your test statistic highly sensitive to a tie type that transmits very little, while making it insensitive to a type that transmits a lot, then it may be relatively easier to achieve even bigger values in a randomization. Alternatively, we can test different types of edges separately and either average or take the max of the two test statistics each time.

5.2 Score

When the measurement set includes treated units, AEI recommend regression-based approaches that adjust for direct treatment effects. Begin by assuming the null hypothesis is correct and furthermore that effects can be modeled with a linear equation:

$$Y_i^{obs} = \alpha + \tau\ z_i + \beta\ X_i + \epsilon_i \tag{2}$$

I have added to their original equation βX to account for covariates, which may be particularly important in a natural experiment. Note that there are no spillover effects in this model. After estimating the coefficients α, τ, β using OLS, examine the covariance between the residuals $\hat{\epsilon}$ and the suspected source of spillover, modeled according to a spillover function $S(i)$.

$$T_{score} = \text{Cov}\left(\widehat{\epsilon_i}\ , S(i)\right) \tag{3}$$

$S(i)$ does not need to be correctly specified, but it should reflect an educated guess about how spillover is likely to work in order to obtain maximal power. For instance, based on our substantive knowledge, we might expect spillover $S(i)$ to be the number of i's neighbors that are treated, $\sum_j G_{ij} z_i$, or perhaps the fraction of i's neighbors that are treated, $\frac{\sum_j G_{ij} z_i}{\sum_j G_{ij}}$. In the latter case, after solving for $\widehat{\epsilon_i}$, the score statistic becomes:

$$T_{score} = \text{Cov}\left(\quad Y_i^{obs} - \hat{\alpha} - \hat{\tau}\ z_i - \hat{\beta}\ X_i\ , \quad \frac{\sum G_{ij} z_i}{\sum G_{ij}} \quad \middle| \quad i \in \mathcal{M} \right) \tag{4}$$

The premise of this statistic is that if the fraction of treated neighbors actually affects the measurement nodes' outcomes, then it should be correlated with those outcomes, particularly after subtracting off whatever variation can be explained by direct effects and covariates. If there are no spillover effects, then the residuals should be randomly distributed and uncorrelated with any statistic that depends solely on $z_{\mathcal{R}}$. If the spillover function is misspecified, or if the null model leaves out an important term, the test will still be valid but less powerful since the correlation will be weaker due to noise. The score statistic generalizes to a weighted context as easily as the edge-level contrast statistic did. The residuals are unaffected, while $S(i)$ picks up the weights automatically, thanks to G. If the edges are of different types, one could test each of them separately and aggregate the results, though correcting for multiple comparisons is worth bearing in mind if there are more than a couple of categories.

5.3 Sum of Squared Residuals

Bowers et al. (2016) suggest a similar method to AEI, also based on residuals from a regression. Instead of running a regression for the null model, run a regression for the model you think is correct. Then use the sum of squared residuals (SSR) to test model fit. In the authors' parametric framework, i can be any node since our exposure mapping allows us to infer potential outcomes for all units. In AEI's non-parametric framework, $i \in \mathcal{M}$. SSR has yet to be used in a non-parametric context.

If treated alters are actually exerting spillover effects, then there should be less noise—and hence a lower SSR—using the actual treatment assignment than when pretending that randomly chosen neighbors are treated. Like the score statistic, SSR should be most powerful when spillovers are correctly specified; it remains to be shown how well it performs under various types of misspecification.

Normally, test statistics are designed to reject the null when the observed statistic is larger—not smaller—than the simulated ones. If the researcher does not want to adjust their code to test for $T_{obs} < T_{sim}$ instead of $T_{obs} > T_{sim}$, Loh and Ren (2020) suggest either using $\frac{1}{SSR}$ or using the coefficient of determination R^2. Since $R^2 = 1 - \frac{SS}{SSR}$ and the numerator $SS = \sum_i \left(Y_i^{obs} - \bar{Y}_i\right)^2$ will not change under simulation randomizations, R^2 monotonically decreases as SSR increases. The p-values obtained using SSR, $\frac{1}{SSR}$, and R^2 will, therefore, be identical.

These statistics can incorporate weights automatically through the exposure mapping function for Y^{obs}. If weights are relevant to spillover, even weights that are misspecified—up to a point—should improve power over no weights, particularly if the weights are ranked correctly (e.g., best friend effects > acquaintance effects). Out of all the test statistics discussed, this one offers perhaps the most potential for testing different edge types. One can simply include a coefficient for each in the model.

6 Simulations and Replications

For pedagogical purposes, I provide in this section a brief demonstration of how one might go about implementing some of these techniques. All the code is available upon request and will eventually be made public as part of a new package in R statistical software. The data come from Paluck, Shepherd, and Aronow's (2016) anti-conflict social norms experiment and are publicly available on the ICPSR database at the time of this writing. The original experiment placed 28 schools in treatment and with those schools designated a small portion of students as eligible for treatment based, in part, on the number of students who reported spending time with them. The aim was to treat varying numbers of social referents in each school to gauge the impact of social referents on each school's overall "climate of conflict" compared to control schools. Within treated schools, researchers hoped to observe additional effects from spending time with a treated student, particular if that student was a social referent. Since the overwhelming majority of students were not eligible for treatment, I relied on Aronow's (2012) advice to simply designate the ineligible students as $\mathcal{M}$ and all eligible students as $\mathcal{R}$. Although Aronow was a coauthor of this study and had published on both the parametric and nonparametric randomization approaches, Paluck et al. (2016) do not appear to have used either technique (though future papers on this study continue to be published; see Gomila et al. (2020)). For randomization, I followed the study's original blocking scheme. For my initial simulations, I chose a single school from those that received treatment. Using the actual data on treatment eligibility and network edges for friends G^{friend} and best friends G^{best} I generated fake outcomes according to the following formula:

$$Y_i^{obs} = Y_i^{base} + \aleph_i Z_i + \beth_i \frac{\sum_j G_{ij}^{friend} z_j}{\sum_j G_{ij}^{friend}} + \gimel_i \frac{\sum_j G_{ij}^{best} z_j}{\sum_j G_{ij}^{best}} \tag{5}$$

where $Y_i^{base} \sim \mathcal{N}(1.0, \quad 0.01), \aleph_i \sim \mathcal{N}(0.6, \quad 0.01)$, $\beth_i \sim \mathcal{N}(0.8, \quad 0.01), \gimel_i \sim$

$\mathcal{N}(1.2,\ 0.01)$. A variance of 0.01 implies a standard deviation of 0.1 so 95% of effect were within ± 0.2 of the mean. Thus, the noise in the baseline, treatment, and spillover effects was substantial, though not enough to drown out the signal. The outcomes here are linear-in-means for both mean of treated friend edges and mean of treated best friend edges. Thus, best friends have a stronger effect than friends both on a per capita basis (since friends were capped at 10 and best friends at 2) and overall.

I then attempted to recover the underlying parameters through my parametric and nonparametric approaches. Running 1000 randomizations took only a couple of seconds, including calculating simulated spillovers $S(Z)$. This, too, was for greater realism since a researcher's educated guess about the functional form of the influence mechanism may not be quite right. After randomizing, I then calculated 441 p-values searching for $\beth$ and $\gimel$ over a grid of $\beth, \gimel \in -0.2,\ 0,\ 0.2, \ldots, 1.8$. For my test statistic, I used the SSR as recommended by Bowers et al. (2016) and Loh and Ren (2020). I deliberately misspecified the spillover function S so that the presumed spillover operated through the number of treated edges rather than the proportion and merged friend and best friend together:

$$Y_i^{obs} = Y_i^{base} + \aleph_i Z_i + \beth_i \sum_j G_{ij}^{all\ friends} z_j \tag{6}$$

However, I used the correct model for transforming $\left(Y^{obs}, Z^{hyp}\right) \rightarrow Y^{hyp}$ when iterating over the grid. Each p-value took a second or two to calculate, which is similar to what AEI reported obtaining MATLAB without parallelization (mine was over 4 cores).

The results shown in Figure 1 are a bit surprising. The true value falls neatly in the middle of the confidence region. However, the region widens rather than narrows as it moves away in either direction. I suspect this has something to do with treating the two edge types as interchangeable. I then repeated this simulation comparing the various test statistics discussed above plus a difference-in-means estimator which compared nodes with spillover greater than zero (under the deliberately misspecified exposure model) to those with none. The weighted edge level contrast statistic likewise created weights for each edge using the correct ratio but on the false assumption that edges of the same type would have equal weight rather than be divided by the degree of their source node. For the categorical SSR, I maintained this misspecified functional form but allowed the friend and best friend nodes to have different effects rather than combining them:

$$Y_i^{obs} = Y_i^{base} + \aleph_i Z_i + \beth_i \sum_j G_{ij}^{all\ friends} z_j + \gimel_i \sum_j G_{ij}^{best} z_j \tag{7}$$

All the test statistics included the true values in their confidence regions (see Figure 2). However, most displayed the same saddle point property as SSR and also did not identify the true coordinates with the highest p-value. The categorical SSR, however, stood out, precisely identifying the coordinates

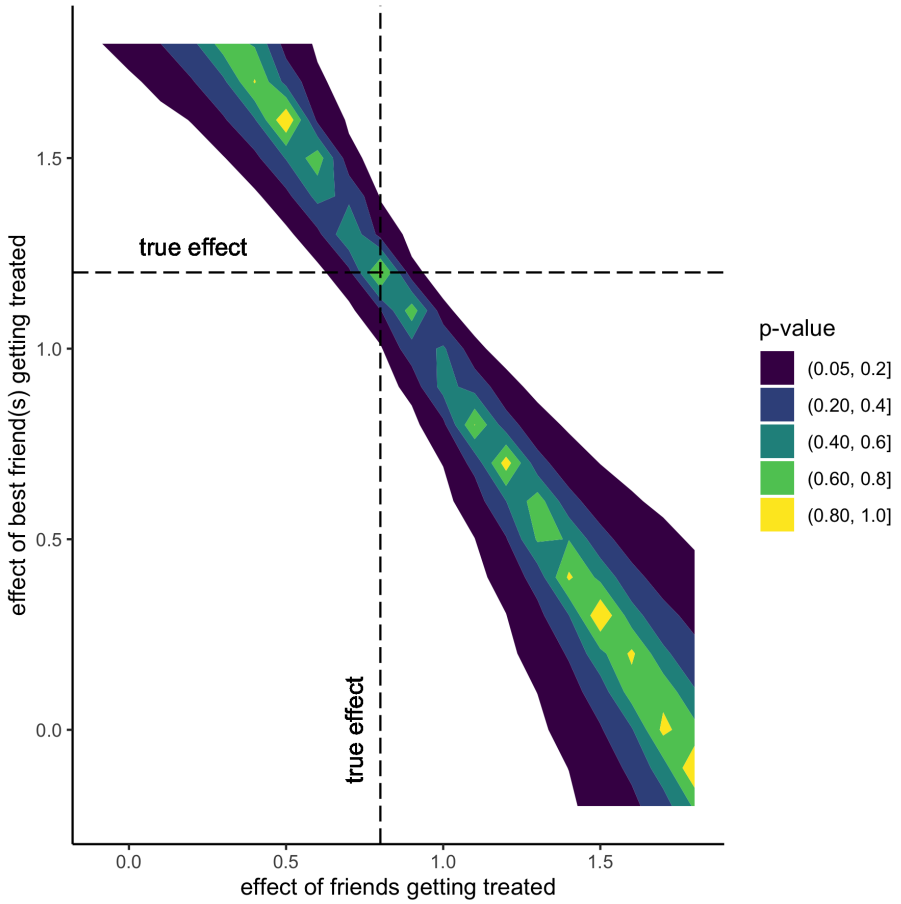

Figure 1: Heatmap for School 4 simulation using SSR as the test statistic and searching over a 21-by-21 grid with increments of 0.1.

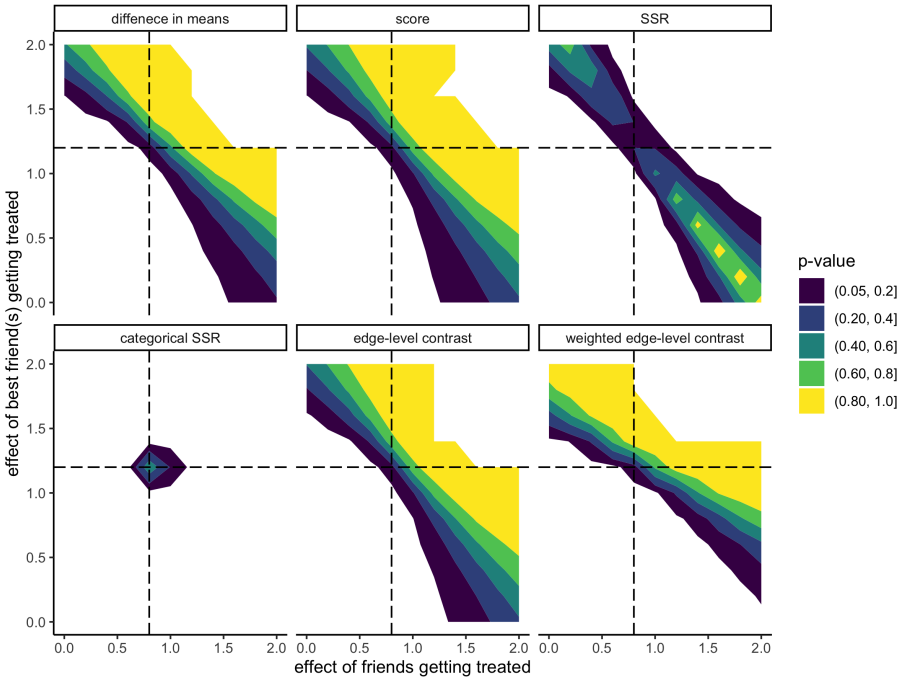

Figure 2: Comparing six test statistics within one school.

forming a tight confidence region around it. Note that all models reject complete null spillover effects of $(0, 0)$.

Finally, I attempted to reproduce results from Paluck et al. (2016) using these techniques, none of which appeared in the original paper. Although direct effects were easy to detect, I did not find evidence of spillovers for shifts in conflict norms. One of the strongest peer-to-peer spillover effects identified in the paper was the distribution of a wristband from students in the anti-conflict program to their peers, and I did find strong evidence of wristband sharing using the nonparametric approach. I saw no evidence of cross-gender or cross-ethnic/racial edges being less influential than within-category ties.

7 Conclusion

Throughout this study, I have argued that taking node and edge heterogeneity into account is a valuable tool to enhance our ability to detect spillover effects and social influence. I laid out options for first, testing if weaker edges have any effect, and second estimating the relative magnitudes of the two edge types. This approach can be applied not only to strong and weak edges but to multiplex and multimodal networks with different categories of nodes and edges. Thus, we can now causally address under what circumstances edges spanning a conflict cleavage are weaker or less influential than those within a social group. We can also learn which group of individuals would benefit most from an intervention that relies on spillover and which members of their network to target as seeds.

Future work on causal inference should strive to continue developing more powerful test statistics and to bring new advances in randomization inference of non-sharp nulls into a network context. Incorporating pretreatment variables into a score statistic or SSR test will also allow these techniques to be applied more widely to quasi-experimental contexts. Practitioners, meanwhile, should start to employ these techniques to address important questions in political science, sociology, and economics about the impact of social cleavages on social capital, political mobilization, and economic development. With time, both advantages in network causal inference and greater awareness of the benefits of randomization inference will hopefully make network causal inference one of the primary tools in every quantitative social scientist's toolbox.

References

An, W. (2018, August). Causal Inference with Networked Treatment Diffusion. *Sociological Methodology 48*(1), 152–181.

Aronow, P. M. (2012, February). A General Method for Detecting Interference Between Units in Randomized Experiments. *Sociological Methods & Research 41*(1), 3–16.

Aronow, P. M., D. Eckles, C. Samii, and S. Zonszein (2020). Spillover Effects

in Experimental Data. In J. Druckman and D. P. Green (Eds.), *Advances in Experimental Political Science*. Cambridge: Cambridge University Press.

Aronow, P. M. and C. Samii (2017, December). Estimating Average Causal Effects Under eneral interference, with application to a social network experiment. *The Annals of Applied Statistics 11*(4), 1912–1947.

Athey, S., D. Eckles, and G. W. Imbens (2018, January). Exact p-Values for Network Interference. *Journal of the American Statistical Association 113*(521), 230–240.

Banerjee, A., A. G. Chandrasekhar, E. Duflo, and M. O. Jackson (2013, July). The Diffusion of Microfinance. *Science 341*(6144), 1236498.

Banerjee, A., A. G. Chandrasekhar, E. Duflo, and M. O. Jackson (2019, November). Using Gossips to Spread Information: Theory and Evidence from Two Randomized Controlled Trials. *The Review of Economic Studies 86*(6), 2453–2490.

Barabási, A.-L. (2016, July). *Network Science*. Cambridge University Press.

Basse, G., A. Feller, and P. Toulis (2019, June). Randomization tests of causal effects under interference. *Biometrika 106*(2), 487–494.

Bond, R. M., C. J. Fariss, J. J. Jones, A. D. I. Kramer, C. Marlow, J. E. Settle, and J. H. Fowler (2012, September). A 61-million-person experiment in social influence and political mobilization. *Nature 489*(7415), 295–298.

Bowers, J., M. M. Fredrickson, and P. M. Aronow (2016). Research Note: A More Powerful Test Statistic for Reasoning about Interference between Units. *Political Analysis 24*(3), 395–403.

Bowers, J., M. M. Fredrickson, and C. Panagopoulos (2013). Reasoning about Interference Between Units: A General Framework. *Political Analysis 21*(1), 97–124.

Carrell, S. E., B. Sacerdote, and J. West (2013). From Natural Variation to Optimal Policy? The Importance of Endogenous Peer Group Formation. *Econometrica 81*(3), 855–882.

Caughey, D., A. Dafoe, X. Li, and L. Miratrix (2021, January). Randomization Inference beyond the Sharp Null: Bounded Null Hypotheses and Quantiles of Individual Treatment Effects. *arXiv:2101.09195 [math, stat]*.

Centola, D. (2010, September). The Spread of Behavior in an Online Social Network Experiment. *Science 329*(5996), 1194–1197.

Christakis, N. A. and J. H. Fowler (2007). The spread of obesity in a large social network over 32 years. *New England Journal of Medicine 357*(4), 370–379.

Cohen-Cole, E. and J. M. Fletcher (2008, December). Detecting implausible social network effects in acne, height, and headaches: longitudinal analysis. *BMJ 337*(dec04 2), a2533–a2533.

Fisher, R. A. (1935). *The design of experiments*. Edinburgh, London: Oliver and Boyd.

Gomila, R., H. Shepherd, and E. L. Paluck (2020, May). Network insiders and observers: who can identify influential people? *Behavioural Public Policy*, 1–28.

Granovetter, M. S. (1973). The Strength of Weak Ties. *American Journal of Sociology 78*(6), 1360–1380.

Habyarimana, J., M. Humphreys, D. Posner, and J. M. Weinstein (2007). Placing and passing: Evidence from Uganda on ethnic identification and ethnic deception. In *Annual Meeting of the American Political Science Association*.

Hudgens, M. G. and M. E. Halloran (2008, June). Toward Causal Inference With Interference. *Journal of the American Statistical Association 103*(482), 832–842.

Imbens, G. W. and D. B. Rubin (2015, April). *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press.

Larson, J. M. (2016, May). Interethnic conflict and the potential dangers of cross-group ties. *Journal of Peace Research 53*(3), 459–471.

Larson, J. M. and J. I. Lewis (2017, April). Ethnic Networks. *American Journal of Political Science 61*(2), 350–364.

Loh, W. W. and D. Ren (2020, October). Estimating social influence in a social network using potential outcomes. *Psychological Methods*.

Lyons, R. (2011, January). The Spread of Evidence-Poor Medicine via Flawed Social-Network Analysis. *Statistics, Politics, and Policy 2*(1).

Manski, C. F. (1993, July). Identification of Endogenous Social Effects: The Reflection Problem. *The Review of Economic Studies 60*(3), 531–542.

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

Neyman, J. S. (1923). On the Application of Probability Theory to Agricultural Experiments: Essay on Principles, Section 9. *Roczniki Nauk Roiniczych X*, 1–51.

Nickerson, D. W. (2008, February). Is Voting Contagious? Evidence from Two Field Experiments. *American Political Science Review 102*(01), 49–57.

Ognyanova, K. (2020, May). Contagious Politics: Tie Strength and the Spread of Political Knowledge. *Communication Research*.

Paluck, E. L. (2011, March). Peer pressure against prejudice: A high school field experiment examining social network change. *Journal of Experimental Social Psychology 47*(2), 350–358.

Paluck, E. L., H. Shepherd, and P. M. Aronow (2016). Changing climates of conflict: A social network experiment in 56 schools. *Proceedings of the National Academy of Sciences 113*(3), 566–571.

Sacerdote, B. (2001). Peer effects with random assignment: Results for Dartmouth roommates. *The Quarterly journal of economics 116*(2), 681–704.

Sacerdote, B. (2011). Peer Effects in Education: How Might They Work, How Big Are They and How Much Do We Know Thus Far? In *Handbook of the Economics of Education*, Volume 3, pp. 249–277. Elsevier.

Salganik, M. J. (2006, February). Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market. *Science 311*(5762), 854–856.

Shalizi, C. R. and A. C. Thomas (2011, May). Homophily and Contagion Are Generically Confounded in Observational Social Network Studies. *Sociological methods & research 40*(2), 211–239.

Sinclair, B., M. McConnell, and D. P. Green (2012). Detecting Spillover Effects: Design and Analysis of Multilevel Experiments. *American Journal of Political Science 56*(4), 1055–1069.

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.

Sun, K., W. Wang, L. Gao, Y. Wang, K. Luo, L. Ren, Z. Zhan, X. Chen, S. Zhao, Y. Huang, Q. Sun, Z. Liu, M. Litvinova, A. Vespignani, M. Ajelli, C. Viboud, and H. Yu (2021, January). Transmission heterogeneities, kinetics, and controllability of SARS-CoV-2. *Science 371*(6526).

Tchetgen Tchetgen, E. J. and T. J. VanderWeele (2012, February). On causal inference in the presence of interference. *Statistical Methods in Medical Research 21*(1), 55–75.

VanderWeele, T. J. (2011, May). Sensitivity Analysis for Contagion Effects in Social Networks. *Sociological Methods & Research 40*(2), 240–255.

Vovk, V. and R. Wang (2019, October). Combining p-values via averaging. *arXiv:1212.4966 [math, stat]*. arXiv: 1212.4966.

Wu, J. and P. Ding (2020, April). Randomization Tests for Weak Null Hypotheses in Randomized Experiments. *Journal of the American Statistical Association 0*(0), 1–16.

Xu, H. and G. Basse (2021, February). Randomization Inference for Composite Experiments with Spillovers and Peer Effects. *arXiv:2103.00567 [math, stat]*.

Zhao, A. and P. Ding (2021, June). Covariate-adjusted Fisher randomization tests for the average treatment effect. *Journal of Econometrics*.

Appendices

A Set Selection: General Considerations

In order to maximize our chances of observing spillover, we need the number of measurement nodes $|\mathcal{M}|$, the number of treated nodes we're permitted to randomize $|\mathcal{R}_{tr}|$, and the number of control nodes we're permitted to randomize $|\mathcal{R}_{ctl}|$. If we think of $\mathcal{R}$ as a satellite transmitting a signal and $\mathcal{M}$ as a satellite dish designed to receive it, then we want both a powerful transmitter and a powerful receiver. All else being equal, this power is a function of sample size. In the absence of other criteria, such as a subset of nodes that were ineligible to participate in the original experiment, Aronow (2012) suggests choosing a value of $|\mathcal{M}|$ to maximize $|\mathcal{M}||\mathcal{R}_{ctl}||\mathcal{R}_{tr}|$. However, the desire to achieve optimal sizes of each of set (measurement and randomization) and each group (treatment and control) should be balanced by other criteria.

Basse et al. (2019) suggest that if the experiment has already taken place, and we don't expect nodes that are treated to be further influenced by their alters' outcomes, we should choose $\mathcal{M}$ from among the control nodes to keep all the treated nodes in $\mathcal{R}$. Their general advice is to avoid including nodes in $\mathcal{M}$ that we think are unlikely to be affected by the effects we are looking for. A natural extension of the authors' logic is that we should think through potential *influence mechanisms* and determine which we think is most likely. If we think, based on our substantive knowledge of the field, that older women, millionaires, introverts, White Republican males, etc. are unlikely to be affected by our treatment but do have a chance of spreading its effects, we should assign them to $\mathcal{R}$. By the same token, if there is a group of individuals who may be influenced by their peers but aren't likely to respond to, or pass along, a treatment they receive directly, we should place them in $\mathcal{M}$. For example, in a study of a program in 56 middle schools designed to reduce conflict, Paluck et al. (2016) predict that social referents—nodes with a high indegree—are more likely to influence their fellow middle school students' conflict behaviors. Thus, if they are correct, placing all the social referents in randomization would lead to stronger spillover effects. Even if the entire randomization set consisted of social referents, inference would still be improved. From the point of view of $i \in \mathcal{M}$ with an alter $j \in \mathcal{R}$, the difference between j being assigned to treatment or control would be more dramatic if j were a social referent than if j were an ordinary friend. However, the authors also sought to test this hypothesis and for

that needed a mixture of social referents and ordinary nodes in $\mathcal{R}$ in order to compare their effects (in fact, they performed a parallel experiment across 56 schools, varying the proportion of social referents in $\mathcal{R}$.)

One desideratum might be to maximize the number of *cross-set* edges connecting nodes in $\mathcal{M}$ to nodes in $\mathcal{R}$. Any edge that connects two nodes in the same set is, in some sense, a lost opportunity. Finding the set with the maximum number of cross-set edges by brute force, however, may be computationally intractable. AEI suggests a greedy "edge comparison" algorithm to find an approximate optimum which I present below, followed by commentary and extensions to weighted networks.

B Edge Comparisons

Procedure 1. *Begin with all nodes assigned to $\mathcal{M}$ and then begin switching nodes to $\mathcal{R}$ one by one. In each step, examine each measurement node to see how many cross-set edges would be created or lost by switching it to randomization. Choose the node that would generate the biggest net increase in cross-set edges. Stop when no more improvements can be attained.*

In an unweighted network, this algorithm will invariably start by switching the highest degree node to randomization. In fact, it will likely continue to select nodes in reverse order of degree for some time, particularly if those nodes are spread out in the network with few connections to one another. Eventually, each high-degree node's "measurement degree" (number of neighbors in measurement) will be degraded enough by loss of neighbors to randomization that low degrees nodes in more peripheral parts of the network will receive priority. This will be particularly true of networks that display assortativity—the tendency of nodes to share edges with nodes of a similar weight or degree. This is a common feature of social networks where popular people tend to be friends with one another (Barabási 2016). If high degree nodes are all in one cluster and low degree nodes are in separate clusters or branch out from the central cluster in long chains, then eventually some of these low degree nodes will have more neighbors left in measurement than the high degree ones. Nevertheless, unless the degree distribution has a low variance (e.g., if nodes are not permitted to have a degree greater than 5), then nearly all the highest degree nodes are likely to end up in measurement.

Depending on our expected influence mechanism, this may be a feature or a bug. If the treatment spillover involves a finite resource, as with the school breakfast discussed in the introduction, then placing a high degree node in $\mathcal{R}$ means that each of their alters will have a smaller chance of receiving this resource—if it is invisible—or, if it can be divided, will receive a smaller share. However, in the Paluck et al. (2016) anti-conflict experiment, high degree nodes were seen as advantageous to spillover, since these individuals were more likely to be role models and the behaviors spread via imitation. We must also consider the direction spillover flows in, relative to the direction of the edges. In the first example, it was the node's outdegree (the number of students they would

consider friends) that matters, since sharing the breakfast was their decision. In the second case, it was the node's indegree (the number of students who consider them a friend) that matters, since the influence actually flows backwards relative to the edge's direction. And finally, if we have data on multiple types of edges, this too should be taken into consideration. Helping someone by sharing food may require a stronger edge, or an edge that implies in-person interaction, whereas weak edges and interaction over social media may suffice for spreading behaviors through exposure. Thus, if we think through the likely influence mechanisms and their relationship to node heterogeneity, we can better equip our artificial experiment with the power to detect effects.

Even if one node has the potential to influence a lot of alters, however, that does not necessarily mean we can make $\mathcal{R}$ smaller while still preserving power. A treated node with ten alters may influence the same number of nodes as 10 treated nodes with one alter each (more, in fact, since there is likely to be overlap among the latter). However, when we run our simulated randomizations, the 10 alters of the high degree node will have their spillovers turned off or on all at once, while the ten alters of ten nodes will exhibit much more internal variation across randomizations, making true effects easier to detect. Thus, while high degree nodes may spread more of an impact, the correlated nature of that impact can reduce the effective sample size and power.

It does not take much of a stretch to extend edge comparisons to a weighted context. Rather than try to maximize the number of cross-set edges, one could instead strive to maximize the total cross-set weight. However, just as adding a node with a cross-set degree of 10 yields smaller gains than adding 10 cross-set edges from distinct nodes, adding a weight 2 cross-set edge will yield smaller gains than adding two weight 1 cross-set edges. Weights, thus, do fully free us from our reliance on edge counts since the latter represent less dependence across units. A better strategy might be to maximize both quantities. For instance, in each step, one could rank all remaining nodes in $\mathcal{M}$ their impact on the total inter-set weight if they were switched to $\mathcal{R}$. Among those in the 90th percentile, choose the node with the highest cross-set degree. Alternatively, one could sort the nodes into cross-set degree-based deciles and use weight as the tiebreaker. If, based on one's substantive knowledge of the likely influence mechanism, it seems preferable to have high-degree, high-weight nodes in $\mathcal{M}$, one can simply start with all nodes in $\mathcal{R}$ rather than $\mathcal{M}$ and run the procedure in reverse.

C ϵ-nets

AEI's other method of set selection aims to maximize the potential for spillover onto each measurement node. This approach too sees within-set edges as edges wasted, but it prioritizes avoiding edges within $\mathcal{M}$. In fact, this procedure is likely to result in a substantial number of wasted edges in $\mathcal{R}$.

Procedure 2. *Begin with all nodes unassigned. Select a node for measurement at random. Immediately assign all of its neighbors to randomization. Repeat this procedure until all the nodes have been selected for one of the two sets.*

Right away, this method presents two obvious inefficiencies. First, due to triadic closure in social networks (one's friends tend to be friends with each other), there is likely to be a substantial number of edges between nodes assigned simultaneously to randomization. Second, these rings of randomization nodes surrounding a measurement node can end up back-to-back. For instance, imagine that nodes are squares on a chessboard connected to the squares they share a side with (no diagonals). We choose a square near the middle for $\mathcal{M}$ and immediately the 4 squares surrounding it are assigned to $\mathcal{R}$. If the next node we choose is two squares to the left, then there the two will, rather conveniently, share a randomization node, and only 3 new nodes need to be placed in $\mathcal{R}$. However, if we choose a node three squares away, then we will end up with two randomization nodes back-to-back between them, sharing an edge. Certainly, we could do better by choosing new nodes in a more systematic fashion. For instance, we could start by selecting a square, preferably in a corner so that it only has two alters to begin with. After surrounding it with randomization squares, choose a square adjacent to one of those squares as the next square assigned to measurement. Eventually you will end up with all the white squares in one set and the black squares in the other, ensuring a perfect 50/50 split with zero edges within categories.

Moving from a chessboard to a complex network, the same principle applies. While we cannot hope to tile a random graph as efficiently as a lattice—unless all circuits are of even length—we can certainly cut down on wasted edges by choosing measurement nodes that are adjacent nodes already placed in randomization. In fact, in each step, we should choose unassigned nodes that will be adjacent to the most randomization nodes. In essence, we are back to the cross-set edges maximization criterion, except that we are assigned nodes in such a way that it is impossible for two nodes in $\mathcal{M}$ to share an edge. This makes the computation considerably easier, improving run time. Each new addition can only improve the total cross-set edge count, since as an unassigned node putting it in one set or the other does not subtract cross-set edges. One limitation to bear in mind, however, is that sharing measurement nodes reduces the variation in the randomizations, since the two nodes that share a neighbor will be correlated in their outcomes.

To incorporate weights into this procedure, we have a couple of options. We can set a weight threshold, below which edges get ignored. This will allow a few edges to connect nodes within the measurement set, but they will be weak ones, so the loss is minor. Rather than setting an absolute threshold, we might set a threshold for each node, ignoring its weakest edges or the bottom quartile of its edges. The tradeoff is that by allowing some less important edges to go to waste within the measurement set, we obtain a more equal number of measurement and randomization nodes overall. If the network is not too large, one could try running this algorithm several times with different thresholds until the optimal number of nodes in each set is achieved. Our second option for leveraging information about weights would be to select new nodes for treatment based on their contribution to the total cross-set weight. Once again, it might behoove us to come up with a scheme that incorporated total weight and total

count in order to maximize the potential for spillover while preserving as much potential variation as possible.

The ϵ-nets approach does not inherently favor high degree/high weight nodes for one set or the other, though they are likely to be assigned more quickly than other nodes due to the friendship paradox. Since we assign nodes that are adjacent to previous assigned nodes, the "next" nodes in a chain will be the opposite set from the previous one, unless chains intersect. So, while $\mathcal{R}$ is likely to end up bigger than $\mathcal{M}$, there will not necessarily be a higher proportion of high degree/weight nodes in $\mathcal{R}$ than low degree/weight nodes. However, if we wanted to place high degree nodes in $\mathcal{M}$, we could simply select unassigned nodes with the highest weight or degree each time we need a new measurement node. The reverse is not possible, however, without significantly modifying the procedure. While it makes sense to surround a measurement node with randomization nodes, each of which gets randomized independently, we gain a lot less when we surround randomization nodes with measurement nodes, since all those measurement nodes will be exposed or not exposed simultaneously. This is the same issue we saw earlier with the 10-degree node versus ten 1-degree nodes dilemma.