Political Analysis (2016) pp. 1–9
doi:10.1093/pan/mpw018

Research Note: A More Powerful Test Statistic for Reasoning about Interference between Units

Jake Bowers and Mark M. Fredrickson
Department of Political Science and Statistics, University of Illinois at Urbana-Champaign, 420 David Kinley Hall (DKH) MC-713, 1407 W Gregory Dr, Urbana, IL 61801, USA
e-mails: jwbowers@illinois.edu (corresponding author); fredric3@illinois.edu

Peter M. Aronow
Department of Political Science and Biostatistics, Yale University, 77 Prospect Street, New Haven, CT 06520, USA
e-mail: peter.aronow@yale.edu

Edited by Prof. Justin Grimmer

Bowers, Fredrickson, and Panagopoulos (2013, Reasoning about interference between units: A general framework, *Political Analysis* 21(1):97–124; henceforth BFP) showed that one could use Fisher's randomization-based hypothesis testing framework to assess counterfactual causal models of treatment propagation and spillover across social networks. This research note improves the statistical inference presented in BFP (2013) by substituting a test statistic based on a sum of squared residuals and incorporating information about the fixed network for the simple Kolmogorov–Smirnov test statistic (Hollander 1999, section 5.4) they used. This note incrementally improves the application of BFP's "reasoning about interference" approach. We do not offer general results about test statistics for multi-parameter causal models on social networks here, but instead hope to stimulate further, and deeper, work on test statistics and sharp hypothesis testing.

1 Overview: Randomization-Based Statistical Inference for Causal Effects from Social Network Experiments

In a randomized experiment with $n = 4$ subjects connected via a fixed network, the response of subject $i = 1$ might depend on the different ways that treatment is assigned to the *whole* network. When the treatment assignment vector, $\mathbf{z}$, provides treatment to persons 2 and 3, $\mathbf{z} = \{0, 1, 1, 0\}$, person $i = 1$ might respond one way, $y_{i=1,\mathbf{z}=\{0,1,1,0\}}$, and when treatment is assigned to persons 3 and 4, $\mathbf{z} = \{0, 0, 1, 1\}$ person $i = 1$ might act another way, $y_{i=1,\mathbf{z}=\{0,0,1,1\}}$. More generally, we might say that if the experiment had a causal effect on person i, then her outcome would differ under different realizations of the experimental treatment as a whole $y_{i,\mathbf{z}} \neq y_{i,\mathbf{z}'}$. The fundamental problem of causal inference reminds us that we can never see both states of the world: we only observe the outcome from person i under one treatment assignment vector, either $\mathbf{z}$ or some $\mathbf{z}'$ not both (Holland 1986; Brady 2008).[1]

Fisher's (1935, chap. 2) approach to design-based statistical inference as developed by Paul Rosenbaum (2010) begins with the premise of the fundamental problem of causal inference. Since we cannot observe all of the ways that a given person would respond to different treatments, the Fisher and Rosenbaum approach suggests that we focus on learning about how *models* of unobservable counterfactual outcomes relate to what we can observe. Although we do not know

Authors' note: Data and code to reproduce this document can be found at Bowers, Fredrickson and Aronow (2016).
[1]In simpler settings, where treatment given to one individual has no effect on any other individual, we tend to write $y_{i,Z_i=1} \neq y_{i,Z_i=0}$ to say that treatment had a causal effect on person i.

how person i would have acted under all possible experimental scenarios, we can learn how much information we have to dispel certain claims or hypotheses. This conceptual move—sidestepping the fundamental problem of causal inference via learning about claims made by scientists—drives hypothesis testing in general. Bowers, Fredrickson, and Panagopoulos (BFP) build on this insight by showing that models of counterfactual effects can involve statements about how treatment given to one node in a social network can influence other nodes. For example, they present a model that allows the effects of treatment to die off as the network distance between nodes increases.[2] They also show that the strength of evidence against the specific hypotheses implied by a given model varies with different features of the research design as well as the extent to which the true causal process diverged from the model. Since their simulated experiment involved two treatments, the only observations available to evaluate the model were comparisons of the assigned-to-treatment group and the assigned-to-control group. Since their model could imply not only shifts in the mean of the observed treatment versus control outcome distributions, but also changes in shape of those distributions, they used the Kolmogorov–Smirnov (KS) test statistic so that their tests would be sensitive to differences in the treatment and control distributions implied by different hypotheses and not only sensitive to differences in one aspect of those distributions (such as the differences in the mean).[3] So, in broad outline, the BFP approach involves: (1) the articulation of a model for how a treatment assignment vector can change outcomes for all subjects in the experiment (holding the network fixed) and (2) the use of a function to compare actually treated and control observations to summarize whether such a model is implausible (codified as a low p value) or to report that we have too little information available from the data and design about the model (codified as a high p value). This is classic hypothesis testing applied to an experiment on a social network with sharp null hypotheses.

Say Y_i is the observed outcome, and we hypothesize that units do not interfere and also that $y_{i,Z_i=1} = y_{i,Z_i=0} + \tau$. We can assess which (if any) hypothesized values of τ appear implausible from the perspective of the data by: (1) mapping the hypothesis about unobserved quantities to observed data using the identity $Y_i = Z_i y_{i,Z_i=1} + (1 - Z_i) y_{i,Z_i=0}$—noticing that if $y_{i,Z_i=1} = y_{i,Z_i=0} + \tau$ then $y_{i,Z_i=0} = Y_i - Z_i\tau$ (by substituting from the hypothesized relationship into the observed data identity); (2) using this result to adjust the observed outcome to represent what would be implied by the hypothesis for a given τ_0 such that $\tilde{y}_{i,Z_i=0} = Y_i - Z_i\tau_0$; and (3) under the hypothesis, $\tilde{y}_{i,Z_i=0}$ should have no systematic relationship with treatment assignment, so we can summarize this relationship with a test statistic, $\mathcal{T}(\tilde{y}_{i,Z_i=0}, Z_i)$. A distribution of values for this test statistic arises from repetitions of the treatment assignment process (new draws of $\mathbf{z}$ from all of the ways that such treatment assignment vectors could have been produced); and finally (4) a p value arises by comparing the observed test statistic, $\mathcal{T}(Y_i, Z_i)$ against the distribution of that test statistic that characterizes the hypothesis.

Notice that the test statistic choice matters in this process: the engine of statistical inference involves summarizing information against the hypothesized claim, yet different ways to summarize information might be more or less sensitive to substantively meaningful differences. The statistical power of a simple test of the sharp null hypothesis of no effects will vary as a function of the design of the study (e.g., total number of observations, proportion treated, blocking structure), characteristics of the outcome (e.g., continuous, binary, skewed, extreme points), and the way that a test statistic summarizes the outcome (does it compare means, standard deviations, medians, qqplots, or something else?).[4] In general, test statistics should be powerful against relevant alternatives. Rosenbaum (2002, section 2.4.4) provides more specific advice about the large sample performance of certain classes of test statistics and BFP repeat his general advice: "Select a test statistic [$\mathcal{T}$] that

[2]We present this model later in this article in equation (3). See the original paper itself for more details of the example model.

[3]If the empirical cumulative distribution function (ECDF) of the treated units is F_1 and the ECDF of the control units is F_0, then the KS test statistic is $\mathcal{T}(\mathbf{y}_0, \mathbf{z})_{\text{KS}} = \max_{i=1,\ldots,n} \left[F_1(y_{i,0}) - F_0(y_{i,0})\right]$, where $F(x) = (1/n)\sum_{i=1}^{n} I(x_i \leq x)$ records the proportion of the distribution of x at or below x_i (Hollander 1999, section 5.4).

[4]Some use the term "effective sample size"—which we first saw in Kish (1965) —to highlight the fact that statistical power depends on more than the number of rows in a given rectangular data set.

will be small when the treated and control distributions in the adjusted data . . . are similar, and large when the distributions diverge." Rosenbaum (2002, Proposition 4 and 5, section 2.9) presents results proving that test statistics with this property ("effect increasing" test statistics) produce an unbiased test of the hypothesis of no effects or positive effects when the positive effects involve one parameter. Such results mean that a test of the sharp null of no effects based on a known randomization using *any* effect increasing test statistic ought to be a valid test (in that the test should produce p values less than α no more than $100\alpha\%$ of the time when the null is true) even though different test statistics may imply different power against false hypotheses. Yet, when the models are complex and may involve increasing effects in the direction of one parameter and non-linear effects in the direction of another parameter, BFP showed that sometimes a KS test: (1) will have no power to address the model at all such that all hypothesized parameters would receive the same high p value; or (2) might describe all such hypothesized parameters as implausible. Thus, although in theory one may assess sharp multi-parameter hypotheses, in practice one may not learn much from such tests. BFP thus recommended simulation studies of the operating characteristics of tests as a piece of their workflow—because the theory justifying a simple one-dimensional effect increasing test statistics clearly did not cover multi-parameter situations like those easily arising from social network experiments.

2 Hypothesis Testing as Model Fit Assessment: The Sum-of-Squared-Residual Test Statistic

Fisher-/Rosenbaum-style randomization inference tends to use test statistics that compare two distributions. Simple models imply that the distribution of the outcome in the control remains fixed. For example, the implication of the constant, additive effects model, $\tilde{y}_{i,Z_i=0} = Y_i - Z_i\tau$, only changes the distribution of outcomes for units in the treated condition. Comparing the mean of $\tilde{y}_{i,Z_i=1}|Z_i = 1$ to the mean of $\tilde{y}_{i,Z_i=0}|Z_i = 0$ makes intuitive sense in this case and, if Y_i is Normal or at least unimodal without major outliers, then this test using means as test statistics might have optimal power. The complex model used as an example by BFP involved adjustments to both control and treated outcomes—some hypothesized parameters would cause shifts in variance, others in location. So, BFP proposed to use the KS test statistic to assess the relationship between $\tilde{y}_{i,Z_i=0,Z_{-i}=0}$ and Z_i (where $Z_{-i} = 0$ means "when all units other than i are also not treated").

Although thinking about test statistics as comparing distributions is natural and simple, one can also think about the process of hypothesis testing as a process of assessing model fit, and there are usually better ways to evaluate the fit of a model than comparing two marginal distributions: For example, the KS test uses the maximum difference in the empirical cumulative distributions of each treatment group calculated without regard for the relationship between the treated and control distributions, thereby ignoring information about the joint distribution that could increase the precision of the test. The simplest version of the sum-of-squared-residual (SSR) test statistic merely sums the difference between the mean of the outcome implied by the hypothesis and individual outcomes, thereby including the correlation between treated and control outcomes as a part of the distribution of the statistic:

$$\mathcal{T}(\mathbf{y}_0, \mathbf{z})_{\text{SSR}} \equiv \sum_i (\tilde{y}_{i,Z_i=0} - \bar{\tilde{y}}_{i,Z_i=0})^2 \tag{1}$$

To provide a little intuition about the SSR test statistic as compared to the KS test statistic, we created a small simulation study of 256 units with half assigned to a binary treatment completely at random. We compared the performance of the two test statistics on two outcomes with no interference—a Normal outcome and a skewed outcome, both with mean of 10 and standard deviation of 1, as shown in Fig. 1.

We compared two models of effects—a constant additive effects model in which $y_{i,Z_i=1} = y_{i,Z_i=0} + \tau$ and a constant multiplicative effects model in which $y_{i,Z_i=1} = y_{i,Z_i=0} \cdot \tau$. We set the truth to be the sharp null of no effects such that the true $\tau = 0$ for the additive model and the true $\tau = 1$ for the multiplicative model. To further explain the process of hypothesis testing and evaluation of the test

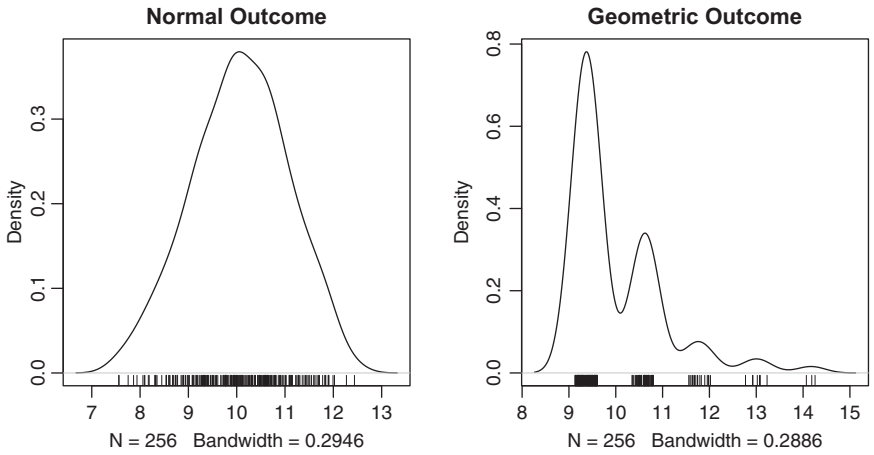

Fig. 1 Two outcomes (Y_i) simulated with no interference, $N = 256$, half randomized to binary treatment by complete randomization.

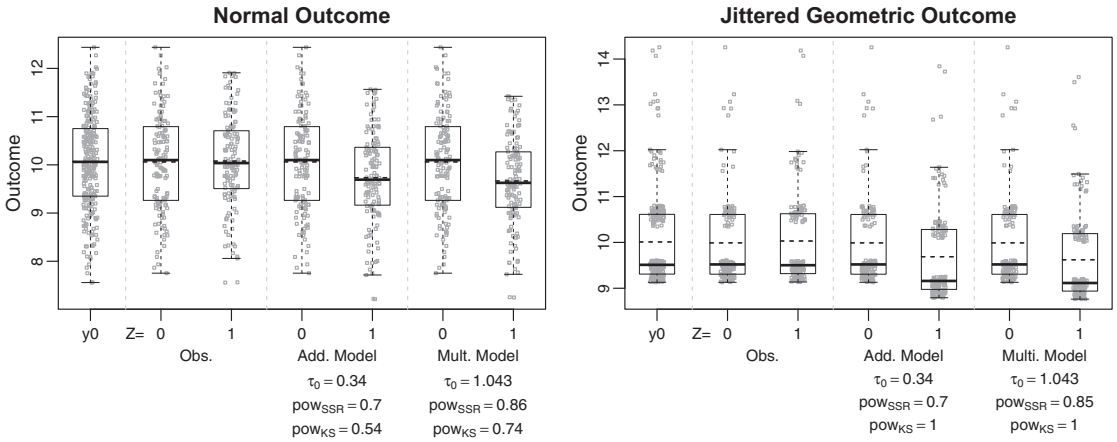

Fig. 2 The SSR test statistic has more power than the KS test statistic for the Normal outcome and less power for the skewed outcome. Each panel shows distributions of simulated data from left to right: outcome with no experiment (y_0), observed outcome after random assignment, outcomes implied by a constant additive model of effects ("Add. Model"), and outcomes implied by a constant multiplicative model ("Mult. Model"). The hypothesized model parameters τ_0 that produce the patterns, and the power of the tests using the KS and SSR test statistics, are printed below the models.

statistics, we display the results from a power analysis of one alternative hypothesis in Fig. 2. A given model and hypothesized parameter implies a distribution for observed outcomes in the treatment and control groups. Here, since we have no interference, the hypotheses only have different implications for the control group distribution. In this case, we chose the two parameters for the additive and multiplicative models which produced very similar implications (as shown by the similarity in the "Add. Model" and "Mult. Model" boxplots in both panels of the figure). We see that when we tested the hypothesis of $\tau_0 = 0.34$ from the additive model using the KS test statistic on Normal data, we produced p values lower than $\alpha = 0.05$ about 54% of the time across 1000 simulations. That is, the KS test has a power of about 0.54 to reject a false hypothesis with this Normal outcome and this constant additive effects model. The analogous power for the SSR test statistic was 0.70. For these particular parameters, we see the performance of the SSR test statistic as better than the KS test statistic with Normal outcomes, but not very effected by the model of effects (which makes some sense here because we are choosing parameters at which both models imply more or less the same patterns in outcomes). On non-Normal outcomes (shown in the

"Jittered Geometric Outcomes" panel), the SSR test statistic had less power. Again, the particular model of effect did not matter in this case because we chose the alternative hypothesis (τ_0 on the plots) to represent the case where both models implied similar outcome distributions.

When we assessed power across many alternative hypotheses in results not shown here but part of the reproduction archive for this paper (Bowers, Fredrickson, and Aronow 2016), our intuition was that the SSR would have more power than the KS test when the outcome was Normal and when the observational implication of the model of effects would shift means (i.e., the additive model). We used direct simulation of the randomization distribution to generate p values and repeated that process 1000 times to gauge the proportion of rejections of a range of false hypotheses (i.e., the power of the tests at many different values of τ_0).[5] The results bear out this intuition: the SSR has slightly more power than KS for Normal outcomes in both the additive and multiplicative effects conditions. The SSR has slightly less power than KS when the outcome is skewed for both models. In general, the SSR ought to be most powerful when the effect of the experiment involves a shift in the location of the distributions of the treated versus the controlgroups. And it should be less powerful when the effect of the experiment is mostly to leave the center of the distribution alone, but instead to stretch or compress the tails, or to concentrate the experimental effect on one quantile.

2.1 *The SSR Test Statistic with Network Information*

In the case where we know the fixed binary adjacency matrix of a social network, $\mathbf{S}$, where a given entry in the $n \times n$ matrix is 1 if two units are adjacent and 0 if two units are not adjacent, and where we imagine that network attributes (like degree) of a node play a role in the mechanism by which treatment propagates, the idea of assessing model fit rather than closeness of distributions leads naturally to the SSRs from a least squares regression of $\tilde{y}_{i,Z_i=0}$ on Z_i and $\mathbf{z}^T\mathbf{S}$ (i.e., the number of directly connected nodes assigned treatment) as well as the $\mathbf{1}^T\mathbf{S}$ (i.e., the degree of the node). If we collected Z_i, $\mathbf{z}^T\mathbf{S}$, and $\mathbf{1}^T\mathbf{S}$ into a matrix $\mathbf{X}$, and fit the $\tilde{y}_{i,Z_i=0}$ as a linear function of $\mathbf{X}$ with fitted coefficients $\hat{\boldsymbol{\beta}}$ then we could define the test statistic as

$$\mathcal{T}(y_0, \mathbf{z})_{\text{SSR}} \equiv \sum_i (\tilde{y}_{i,Z_i=0} - \mathbf{X}\hat{\boldsymbol{\beta}})^2. \tag{2}$$

2.2 *The SSR Test Statistic and the BFP Example Model*

As an example of the performance of these new statistics, we reanalyze the model and design from BFP. Their model of treatment propagation was

$$\mathcal{H}(\mathbf{y}_\mathbf{z}, \mathbf{w}, \beta, \tau) = \frac{\beta + (1 - w_i)(1 - \beta)\exp(-\tau^2\mathbf{w}^T\mathbf{S})}{\beta + (1 - z_i)(1 - \beta)\exp(-\tau^2\mathbf{z}^T\mathbf{S})}\mathbf{y}_\mathbf{z}. \tag{3}$$

Briefly, this model posits that treatment effects can depend on either direct assignment to treatment ($\mathbf{z}$) governed by β or spillover as an increasing (but flattening) function of the number of directly connected treated neighbors ($\mathbf{z}^T\mathbf{S}$) and is governed by τ. So, we have a model with two parameters. The network used by BFP involves 256 nodes connected in an undirected, random graph with node degree ranging from 0 to 10 (mean degree 4, 95% of nodes with degree between 1 and 8, five unconnected nodes with degree 0). Treatment is assigned to 50% of the nodes completely at random in the BFP example.

We assess three versions of the SSR test statistic versus three versions of the KS test statistic. The first, described above, we call the SSR + Design test statistic because it represents information

[5]Those interested in the code for this document can find it at https://github.com/jwbowers/TestStatRIInterference.

about how treatment is assigned to the nodes, $\mathbf{z}^T\mathbf{S}$. The second version of the SSR test statistic (SSR + Degree) only includes network degree, $\mathbf{1}^T\mathbf{S}$, and excludes information about the treatment status of other nodes. And the third version (SSR) includes only treatment assignment $\mathbf{z}$. The top row of Fig. 3 compares the power of the SSR + Design test statistic (upper left panel) to versions of this statistic that either only include fixed node degree (SSR + Degree), or no information about the network at all (SSR). For each test statistic, we tested the hypothesis $\tau = \tau_0, \beta = \tau_0$ by using a simulated permutation test: we sampled 1000 permutations. We executed that test 10,000 times for each pair of parameters. The proportion of p values from that test less than .05 is plotted in Fig. 3: darker values show fewer rejections, lighter values record more rejections. All of these test statistics are valid—they reject the true null of $\tau = .5, \beta = 2$ no more than 5% of the time at $\alpha = .05$—and the plots are darkest in the area where the two lines showing the true parameters intersect. All of the plots have some power to reject non-true alternatives—as we can see with the large white areas in all of the plots. Only when we add information about the number of treated neighbors to the SSR + Degree statistic does the plot show high power against all alternatives in the plane.

The bottom row of Fig. 3 demonstrates the power of the KS test. The bottom right-hand panel shows the test used in the BFP paper. Again, all of the tests are valid in the sense of rejecting the truth no more than 5% of the time when $\alpha = .05$ although all of these tests are conservative: the SSR-based tests rejected the truth roughly 4% of the 10,000 simulations but the KS tests rejected the truth roughly 2% of the time. The KS + Design and KS + Degree panels show the power of applying the KS test to residuals from linear models including network degree only (the + Degree version) or degree and also the number of treated neighbors (the +Design version). That is, whereas the SSR panels used the sum of squared residuals after accounting for network degree and/or number of treated neighbors, the KS + Design and KS + Degree panels apply the KS test to the raw residuals after adjusting for information about the design and network (or with no adjustment). These panels show (1) that inclusion of a quantity from the true model (number of treated neighbors) is not enough to increase power against all alternatives to the level shown by the SSR + Design test statistic and (2) that the KS tests and the SSR tests have different patterns of power—the KS tests appear be less powerful in general (more darker areas on the plots).

3 Application: Legislative Information Spillovers

Coppock (2014) presents a reanalysis of an experiment performed by Butler and Nickerson (2011). Leading up to a key vote on a budget bill, SB24, in the New Mexico State Senate, Butler and Nickerson randomly assigned legislators to see constituent survey results. The original analysis found that legislators assigned to receive constituent information were more likely to vote consistently with the preferences of their district. Given the easy nature of sharing information in field experiments (Winters, Testa, and Fredrickson 2012), Coppock reanalyzes the experimental results with an eye toward spillover, constructing a social network based on similarity on W-NOMINATE ideology scores. Coppock hypothesized that ideologically similar legislators would be more likely to share information. If the constituent survey information does induce a change in behavior, this effect may be observed not only in those legislators that are directly treated in the experiment, but also by those of similar ideology to those that are treated.

With distance matrix Γ, Coppock evaluated the model:

$$\mathbf{y}_0 = \mathbf{y}_\mathbf{z} - \beta_1\mathbf{z} - \beta_2 g(\Gamma\mathbf{z}),$$

where g is a function that normalizes the sums of distances to treated neighbors to have unit variance. This model has a direct effect (β_1) and an indirect effect (β_2) that is linear in the distances to treated neighbors. Coppock evaluated this spillover model using the SSR + Design statistic as presented in this article.

To demonstrate the performance of this statistic in an applied setting, such as this one in which the model is linear in the spillover effects, we repeat Coppock's analysis using both the

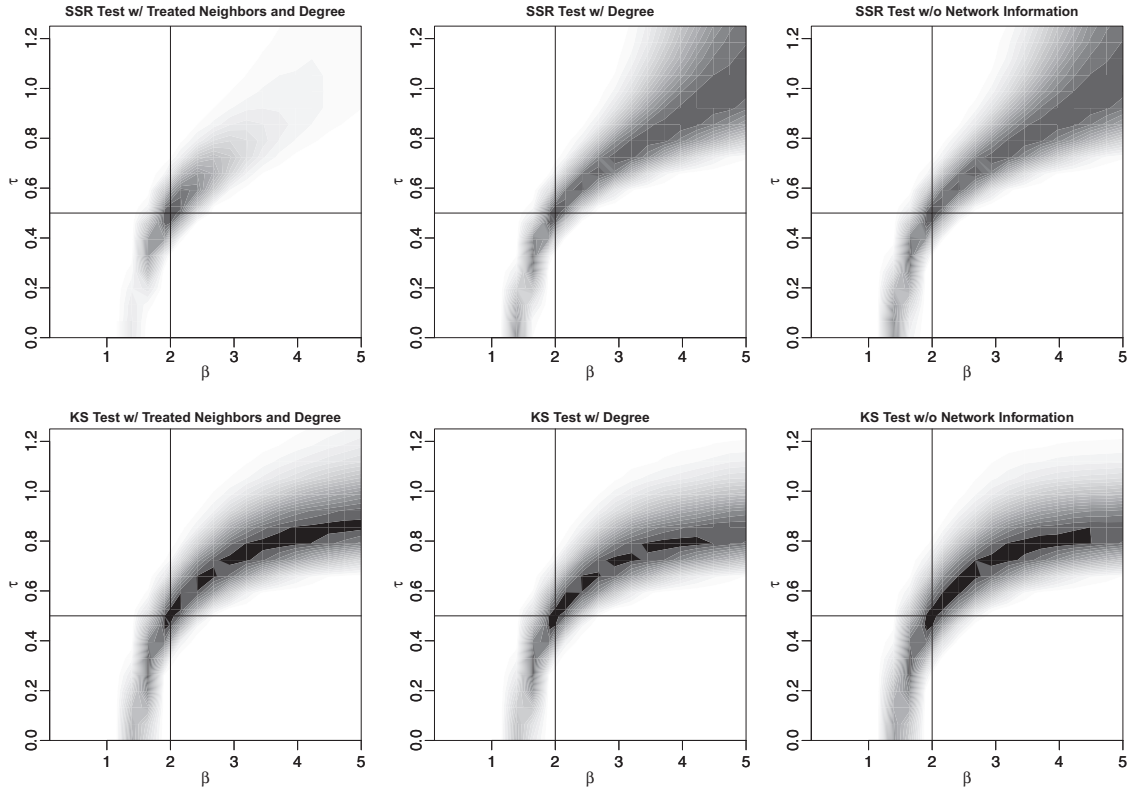

Fig. 3 Proportion of p values less than .05 for tests of joint hypotheses about τ and β for the model in equation (3). Darker values mean rare rejection. White means rejection always. Truth is shown at the intersection of the straight lines $\tau = .5, \beta = 2$. Each panel shows a different test statistic. The SSR tests refer to equation (2), the KS tests refer to the expression in footnote 3.

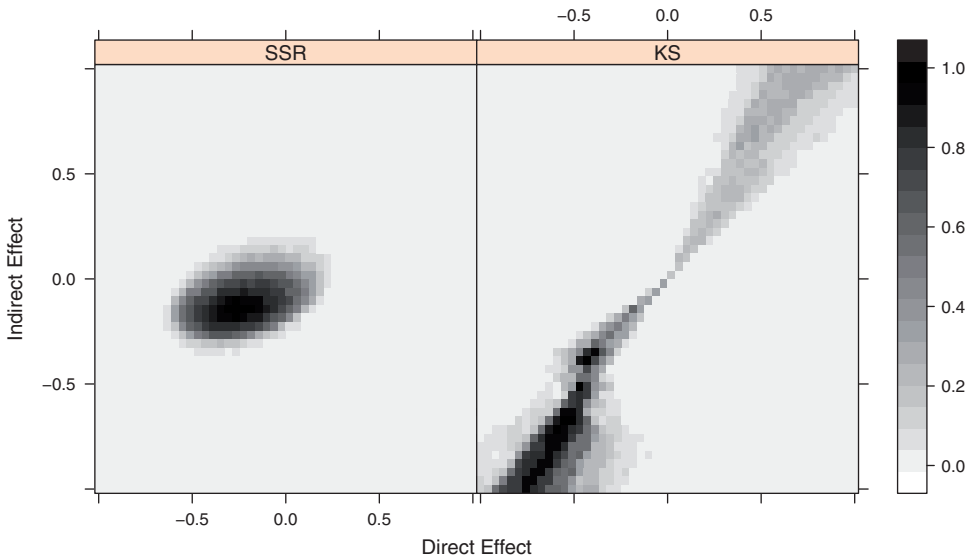

Fig. 4 Replication of Fig. 2 from Coppock (2014): plotting p values for direct and indirect model parameters. The original analysis used the SSR statistic, which nicely bounds the region of plausible hypotheses. The KS statistic, while eliminating more hypotheses in the neighborhood of (0, 0), fails to provide any such bounds.

SSR + Design statistic and the simple KS statistic. The results in Fig. 4 recreate Coppock's Fig. 2.[6] The KS test statistic, while being more powerful in the neighborhood of (0, 0), fails to bound the region of plausible hypotheses in the manner of the SSR + Design statistic. We find this quality of the test statistic to be particularly valuable. A powerful test statistic is a useful tool, but equally useful is the ability to succinctly describe the region in which we would fail to reject a null hypothesis. In this case, the SSR statistic provides such a region, making discussing and evaluating the results of Coppock's analysis much more straightforward: it would be very unlikely that the data were generated from the model with parameters outside the rectangle with corners (−0.6, 0.2) and (0.3, −0.4).

4 Discussion and Speculations

We cannot say here whether the SSR + Design test will provide the best power against relevant alternatives for all possible models of treatment effect propagation, network topologies, and designs: this article uses two models of effects, each of which was applied to a different network topology and experimental design.[7] However, we hope that the results from the examples presented here improve the application of the BFP approach and raise new questions for research. BFP are correct in the assertion that, regardless of the choice of test statistic selection, a set of implausible hypotheses is identified by the procedure. But we should not be led to believe that, for any given test statistic, some hypotheses are universally more plausible than others. Such inferences—comparing hypotheses—may depend on the test statistic used, and not necessarily reflect the plausibility of the model at hand. That is, the results of any hypothesis test (or confidence interval) tell us *both* about the test statistic *and* about the causal model under scrutiny.

In the example shown in Fig. 3, the SSR + Design test statistic had much better power than any other test statistic. But SSR from an ordinary least squares regression is not always appropriate: for example, when the probability of exposure to spillover is heterogeneous across individuals in a way not well captured by the $\mathbf{z}^T\mathbf{S}$ term or some other analogous term, we may wish to apply inverse probability weights so as to ensure representative samples of potential outcomes. This suggests a conjecture: that the *SSR* from an *inverse-probability-weighted* least squares regression is a more generally sensible test statistic for models that include interference.[8] Additionally, when nonlinear deviations from model predictions are of concern, a weighted variant of the Brownian distance covariance (Szekely and Rizzo 2009) or other *E* statistic may be more sensible than the sum of squared residuals.

References

Aronow, Peter M., and Cyrus Samii. 2012. Estimating average causal effects under general interference. Unpublished manuscript.

Bowers, Jake, Mark Fredrickson and Peter M. Aronow. 2016. Replication data for: Research Note: A more powerful test statistic for reasoning about interference between units. *Harvard Dataverse*. http://dx.doi.org/10.7910/DVN/V6ECYU.

Bowers, Jake, Mark M. Fredrickson, and Costas Panagopoulos. 2013. Reasoning about interference between units: a general framework. *Political Analysis* 21(1):97–124.

Brady, Henry E. 2008. Causation and explanation in social science. In: *Oxford handbook of political methodology*, eds. Janet M. Box-Steffensmeier, Henry E. Brady and Collier David, 217–70. Oxford, UK: Oxford University Press, http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199286546.001.0001/oxfordhb-9780199286546.

Butler, Daniel M., and David W. Nickerson. 2011. Can learning constituency opinion affect how legislators vote? Results from a field experiment. *Quarterly Journal of Political Science* 6(1):55–83.

Coppock, Alexander. 2014. Information spillovers: another look at experimental estimates of legislator responsiveness. *Journal of Experimental Political Science* 1(2):159–69.

Fisher, Ronald A. 1935. *The design of experiments*. Edinburgh: Oliver and Boyd.

[6]We are grateful to Alex Coppock for his help with both code and data. Figure 4 does not precisely match the figure in the original publication because we corrected a small programming error during replication.

[7]The BFP paper itself engages with some questions about the performance of this approach when the theoretical model is very different from the process generating the data, and we encourage readers to see that discussion in their section 5.2.

[8]Aronow and Samii (2012) use such weights for unbiased estimation of network-treatment-exposure probability weighted average treatment effects.

Holland, Paul W. 1986. Statistics and causal inference (with discussion). *Journal of the American Statistical Association* 81:945–70.

Hollander, Myles. 1999. *Nonparametric statistical methods,* 2nd ed. Hoboken, NJ: Wiley-Interscience.

Kish, Leslie 1965. *Survey sampling*. New York: John Wiley and Sons.

Rosenbaum, Paul R. 2002. *Observational studies*, 2nd ed. New York, NY: Springer.

———. 2010. *Design of observational studies*. New York, NY: Springer, http://www.springer.com/statistics/statistical+theory+and+methods/book/978-1-4419-1212-1.

Szekely, Gábor J., and Maria L. Rizzo. 2009. Brownian distance covariance. *Annals of Applied Statistics* 3(4):1236–65.

Winters, Matthew S., Paul Testa, and Mark Fredrickson. 2012. Using field experiments to understand information as an antidote to corruption. *Research in Experimental Economics* 15:213–246, http://www.emeraldinsight.com/doi/abs/10.1108/S0193-2306%282012%290000015010.