Research Note: A more powerful test statistic for reasoning about interference between units

Jake Bowers [*] Mark Fredrickson [†] Peter M. Aronow [‡]

August 26, 2015

Abstract

Bowers, Fredrickson and Panagopoulos (2012) 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 infernce presented in Bowers, Fredrickson and Panagopoulos (2012) (henceforth BFP) 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, §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.

Key Words: Causal effect; Interference; Randomized experiment; Randomization inference; Fisher's Sharp Null Hypothesis; SUTVA

[*] Associate Professor, Dept of Political Science and Statistics, University of Illinois @ Urbana-Champaign (`jwbowers@illinois.edu`).

[†] Ph.D. Student in Statistics and Political Science, University of Illinois @ Urbana-Champaign.

[‡] Assistant Professor, Dept of Political Science and Biostatistics, Yale University

1 Background on randomization based statistical inference for causal effects from social network experiments

In a randomized experiment with $n = 4$ subjects connected via a fixed network, subject $i = 1$ might respond differently to different ways that treatment is assigned to the whole network. When the treatment assignment vector provides treatment to persons 2 and 3, $\mathbf{z} = \{0, 1, 1, 0\}$ she might act one way, $y_{i=1,\mathbf{z}=\{0,1,1,0\}}$ and when treatment is assigned to persons 3 and 4, $\mathbf{z} = \{0, 0, 1, 1\}$ she might act another way, $y_{i=1,\mathbf{z}=\{0,0,1,1\}}$ That is, if the experiment had a causal effect on person i, then her outcome would differ under different realizations of the experimental treatment $y_{i,\mathbf{z}} \neq y_{i,\mathbf{z}'}$. 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).

R.A. Fisher's approach to design-based statistical inference (Fisher, 1935, Chap 2) as developed by Paul Rosenbaum (Rosenbaum, 2010) begins with the premise of the fundamental problem of causal inference. Given the unobservability of potential outcomes, 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 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. 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.[1] 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

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

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 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 merely sensitive to differences in one aspect of those distributions (such as the differences in the mean).[2] 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) use a function to comparing actually treated and control observations summarize whether such a model is implausible (codified as a low p-value) or whether 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.

So, 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 given τ_0 such that $\widetilde{y}_{i,Z_i=0} = Y_i - Z_i \tau_0$; and (3) Under the hypothesis, $\widetilde{y}_{i,Z_i=0}$ should have no systematic relationship with treatment assignment, so we can summarize this relationship with a test statistic, $\mathcal{T}(\widetilde{y}_{i,Z_i=0}, Z_i)$ and the distribution of values for this test statistic arising from repetitions of treatment assignment (new draws of $\mathbf{z}$ from all of the ways that such treatment assignment vectors could have been

[2] 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 $\mathcal{T}(\mathbf{y_0}, \mathbf{z})_{\text{KS}} = \max\limits_{i=1,\dots,n} [F_1(y_{i,\mathbf{0}}) - F_0(y_{i,\mathbf{0}})]$, 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, §5.4).

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 (proportion treated, blocking structure, etc), characteristics of the outcome (continuous, binary, skewed, extreme points, etc), and the way that a test statistic summarizes the outcome (does it compare means, standard deviations, medians, qqplots, etc). In general, test statistics should be powerful against relevant alternatives. Rosenbaum (2002, §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 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, §2.9) presents results proving that test statistics with this property ("effect increasing" test statistics), produces an unbiased test of the hypothesis of no effects or positive effects when the positive effects involve one parameter. 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 will have (1) no power to address the model at all such that all hypothesized parameters would receive the same implausibility assessment; (2) or might reject all such hypothesized parameters. Thus, although in theory one may assess sharp hypotheses about multiparameter settings, 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 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 SSR test statistic

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, $\widetilde{y}_{i,Z_i=0} = Y_i - Z_i\tau$ only changes the distribution of outcomes for units in the treated condition. Comparing the mean of $\widetilde{y}_{i,Z_i=1}|Z_i = 1$ to the mean of $\widetilde{y}_{i,Z_i=0}|Z_i = 0$ makes intuitive sense in this case, and, if Y_i is Normal or at least unimodal without major outliers, then this test 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 $\widetilde{y}_{i,Z_i=0,Z_{-i}=0}$ and Z_i (where $Z_{-i} = 0$ means "when all units other than i are also not treated".)

Yet, 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. In the case where we know the fixed adjacency matrix of the network, $\mathbf{S}$, 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 the sum-of-squared-residuals (SSR) from a least squares regression of $\widetilde{y}_{i,Z_i=0}$ 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 $\widetilde{y}_{i,Z_i=0}$ as a linear function of $\mathbf{X}$ with coefficients $\boldsymbol{\beta}$ when we could define the test statistic as:

$$\mathcal{T}(\mathbf{y_0}, \mathbf{z})_{\text{SSR}} \equiv \sum_i (\widetilde{y}_{i,Z_i=0} - \mathbf{X}\hat{\boldsymbol{\beta}})^2 \tag{1}$$

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

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

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 nodes with degree 0 [i.e. unconnected]). 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 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 figure 1 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 Figure 1: 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 $\alpha = .05$ — 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. However, only when we add information about the number of treated neighbors to the SSR+Degree statistic, do we see high power against all alternatives in the plane.

The bottom row of Figure 1 demonstrates the power of the KS test. The bottom right

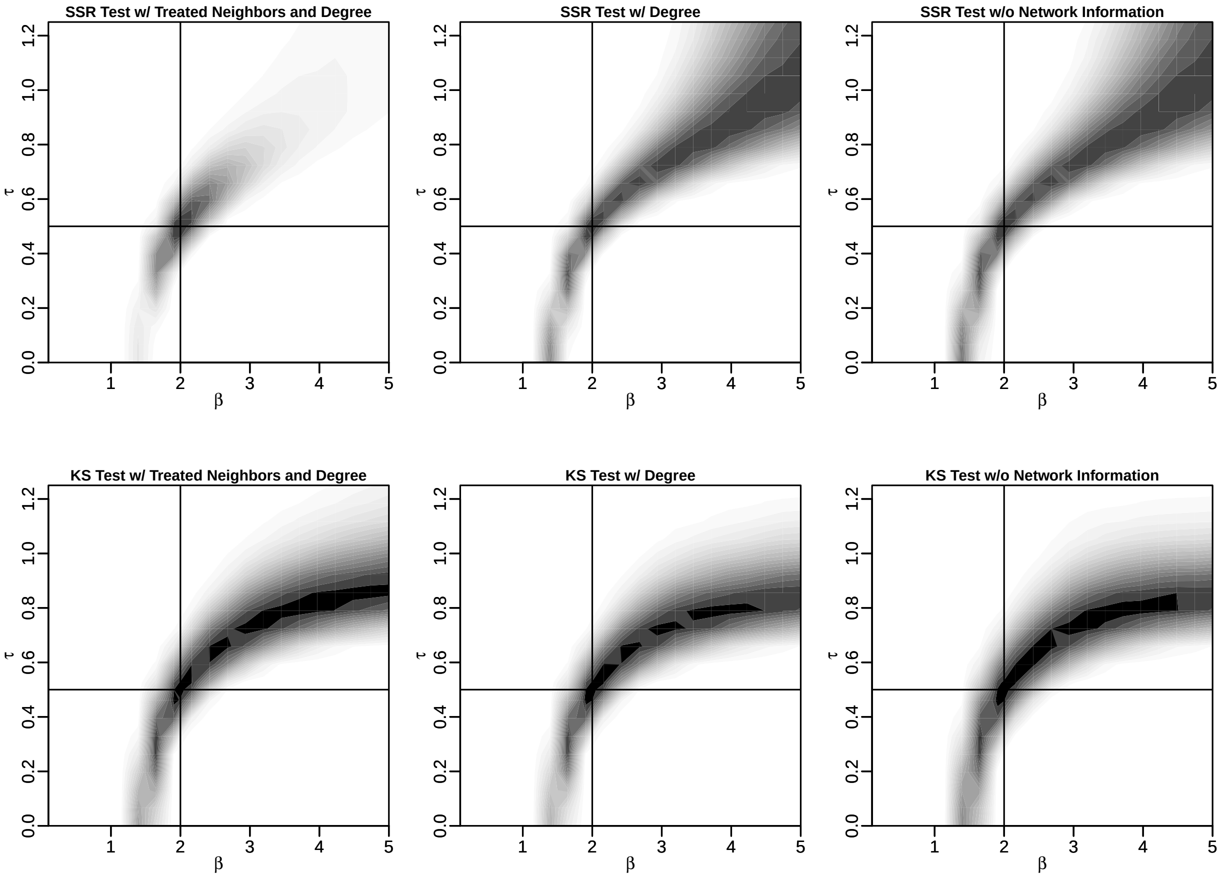

Figure 1: Proportion of p-values less than .05 for tests of joint hypotheses about τ and β for the model in equation 2. Darker values mean rare rejection. White means rejection always. Truth is shown at the intersection of the straight lines $\tau = .5, \beta = 2$. Each panel shows a different test statistic. The SSR Tests refer to equation 1, the KS tests refer to the expression in footnote 2.

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 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. However, we hope that this research note both improves the application of the BFP approach and raises 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, that 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 creation) tell us *both* about the test statistic *and* about the causal model under scrutiny.

In the example above, 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 more generally a sensible test statistic for models that include interference.[3] 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 SSR.

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

References

Aronow, Peter M. and Cyrus Samii. 2012. "Estimating Average Causal Effects Under General Interference.".

Bowers, Jake, Mark M. Fredrickson and Costas Panagopoulos. 2012. "Reasoning about Interference Between Units: A General Framework." *Political Analysis* 21(1):97 – 124.

Brady, Henry E. 2008. "Causation and explanation in social science." *Oxford handbook of political methodology* pp. 217–270.

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

Holland, P. W. 1986. "Statistics and Causal Inference (with discussion)." *Journal of the American Statistical Association* 81:945–970.

Hollander, Myles. 1999. *Nonparametric Statistical Methods, 2nd Edition.* Second ed. Wiley-Interscience.

Rosenbaum, Paul R. 2002. *Observational Studies.* Second ed. Springer-Verlag.

Rosenbaum, Paul R. 2010. *Design of Observational Studies.* Springer.
URL: *http://www.springer.com/statistics/statistical+theory+and+methods/book/978-1-4419-1212-1*

Szekely, G.J. and M.L. Rizzo. 2009. "Brownian distance covariance." *The annals of applied statistics* 3(4):1236–1265.