Network Effects in Field Experiments on Interactive Groups: Cases from Legislative Studies

Sayali Phadke * Bruce A. Desmarais †

May 22, 2017

Abstract

DRAFT, RESULTS MAY CHANGE

Most social processes involve complex interaction among units through some form of social, communication, or collaboration network. The stable unit treatment value assumption (SUTVA)—the assumption that a unit's outcome is unaffected by other units' treatment statuses—is required in conventional approaches to causal inference. When SUTVA is violated, as in networked social interaction, treatment effects spread to control units through the network structure. We evaluate the evidence for spillover effects in data from three field experiments on US state legislatures. Randomized field experiments represent the gold standard in causal inference when studying political elites. It is rarely possible to bring political elites into a controlled laboratory environment, and causal identification with observational data is fraught with problems. We review recently-developed methods for testing for causal effects—including interference effects—while relaxing SUTVA. We propose new specifications for treatment spillover models, and construct networks through geographical or ideological proximity and co-sponsorship. Considering different combinations of spillover models and networks, we evaluate the robustness of recently developed non-parametric tests for interference. The approaches we illustrate can be applied to any experimental setting in which interference is suspected.[1]

*PhD Student, Departments of Statistics, Pennsylvania State University, sayalip@psu.edu.

†Associate Professor, Department of Political Science, Pennsylvania State University, bdesmarais@psu.edu.

[1]Prepared for presentation at the 2016 Political Networks Conference, Washington University, St. Louis. Work supported in part by National Science Foundation grants SES-1558661, SES-1619644, CISE-1320219, and IGERT Grant DGE-1144860, Big Data Social Science. Any opinions, findings, and conclusions or recommendations are those of the authors and do not necessarily reflect those of the sponsor.

1 To-do before submission

- Ensure that literature review includes papers on [SP] [Roughly, yes]
  - Interference/diffusion/propagation models
  - Experiments on networks and their applications
  - Approaches to inference or estimation with propagation
  - Potential outcomes framework
  - Review of political networks
  - Review of field experiments
- Analyze without expected exposure adjustment (so just raw exposures) [SP: For all analysis] [Done for all]
- Run an equal number of replications on each analysis and save results + plots [SP] [Done; 5000 for Coppock and Bergan each]
- Broockman analysis: [Done]
  - Consider blackpercent or dempercent individually for the network [SP]
- Coppock analysis: [Done]
  - Do we want address the issue of tied neighbors?
  - Use number of shared committees in adjacency matrix and see the results [SP] Checked; preliminary raw analysis shows zero indirect effect

- [Priority] Need to decide if we should include both raw and expected results. Adding raw results would entail a discussion about why we think expected exposure Coppock's specification is redundant [Done. Decided to include only raw exposure, except in case of Coppock replication]

- A note on why the adjustment for expected exposure in Coppock's method is not critical [BD]

- A note explaining the intuition behind the meaning of movement in effect estimate due to spillover specification. Depending upon the context, estimate under SUTVA is either conservative or or optimistic. In some cases, effect spilling over to neighbors is good (GOTV mails). In others, it is may show that the treatment has worked worse than we thought (disease prevention). This may also depend upon the type of network being used (e.g. friendship network v/s enmity network). This leading into a note about SUTVA violations resulting in attenuation bias (example: Ghana experiment) [BD]

- A note that we are not distinguishing between contagion and spread of treatment itself, at this point [BD]

- A note on interpretation of CIs. Can we explain them in terms of 'if 100 such tests were conducted, 95% of the p-values would lie in this interval' [SP&BD]

- Include geographical network in Coppock and Bergan [SP][Can't find the code used to generate Coppock geographical net]

- Consider additional test statistic such as the Anderson-Darling test and other tests

mentioned in (**?**) (Mann-Whitney U test, Control Median test etc.) [SP]

- Consider weighted combinations of different networks to estimate a γ [Maybe; SP]
- Do we want to include network plots for all networks we have considered? [Could look into local egocentric networks; however this is low priority]
- Separate out the high and low support district in original data and conduct separate analyses. This would explain the direction of the spillover effect much better. Conduct this analysis with ideological network as well as committee network [Low priority: SP]
- Explore the idea of using communities to model spread across the network [Lower priority]

2 Introduction

In social science, researchers often focus on sets of actors who interact on a regular basis. Areas of social science research in which regular and familiar interaction constitutes the norm include the study of team performance in formal organizations (e.g., **?**), the study of students from the same school/classroom (e.g., **?**), and the study of political elites from the same political institution (e.g., **?**). The importance of collections of highly interactive individuals, which may have conventionally been termed "small groups." (**?**), spans disciplines; including social psychology, public health, political science and organizational studies. Given the rise of research focused on social media (**?**), which includes

the study of interpersonal interaction, but is focused on large groups, we will use the term, "interactive group research."

To draw causal inferences in the study of an interactive group, the most reliable approach is to run a randomized experiment in which the researcher controls one or more interventions. Due to the interactions between group members, the effect of an intervention on one group member may spread to others through the network of interactions. The conventional framework for causal inference relies on SUTVA (Stable Unit Treatment Value Assumption). SUTVA holds that one unit/subject is not affected by the treatment status of any other unit (**?**). However, SUTVA breaks down in a network setting (**?**) when there are post-treatment interactions among units. The violation of SUTVA is termed "interference". When testing for causal effects in experiments on interactive groups, there are two related motivations for researchers to test and account for interference. First, even if the researcher is not interested in the nature of interference, it may be necessary to account for interference to accurately identify the direct effects of interventions on units. Second, interference processes may play a major role in shaping individual or group outcomes, and may therefore be important to evaluate if the objective of the research is to understand or explain outcomes arising from the group's social process.

Social influence and other processes of interdependence are central to nearly every domain of social science, researchers are well aware of the major limitations associated with causal inferences regarding influence drawn from observational studies (**?**). As such, a growing body of research seeks to study interference through experimental interventions (e.g., **???????**). These studies follow a variety of approaches to designing the interventions and testing for interference effects. However, it is clear that the field has, as of yet,

converged upon a consistent methodological framework for testing for causal effects in the presence of interference.

We review a recently developed general framework, introduced by **?**, for testing causal hypotheses in the presence of interference. As an illustration, we then apply this methodology to data generated by field experiments on US state legislatures (**??**); experiments that were not intended to study interference. As part of our application, we discuss and illustrate several choices researchers need to make in testing interference hypotheses.

Our results are three-fold. First, we review a broad framework for evaluating interference hypotheses, and make the case that researchers should consider such hypotheses when evaluating the results of experiments on interactive groups. Second, we show that we consistently find evidence in support of interference hypotheses in data from field experiments on US state legislatures. Third, we encourage researchers to consider several modeling dimensions in formulating interference hypotheses.

3 Experimental Research on Interactive Groups

In laboratory experiments, especially those in which the treatment is administered and response measured while the subject is in the lab, it is feasible to assure that subjects do not interact, such that SUTVA holds. However, in administering laboratory experiments with the intent of generalizing results to a target population of interest, researchers face the dual tasks of (1) recruiting participants from a target population of interest and (2) recreating real-world conditions relevant to the effects of the treatment. Since it is often impossible to complete these tasks, laboratory experiments often suffer from a lack of

external validity, where external validity is understood to be the generalizability of the effects of the treatment in the laboratory to members of the target population outside of the lab setting. **?**, p.245 puts it concisely, "Laboratory experiments are considered externally invalid because findings from the population studied cannot be generalized to populations other than the one from which the sample was drawn." This quote is an over-simplification, but it underscores the advantage of field experiments relative to lab experiments. That is, field experiments are directed at the population of interest, in its natural setting (**?**).

One characteristic of field experiments is that the researchers cannot prevent subjects from interacting. Furthermore, we argue, several classes of field experiments are conducted within contexts that are likely to lead to interaction among subjects. Such interaction should lead to the possibility that the results of the experiment reflect the effects of the experimental conditions to which subjects were assigned, as well ass interference across subjects according to patterns of interaction.

The first class of field experimental designs we discuss involves experimenting on relatively small groups of subjects who work for the same organization, in either the same building or on the same campus. Examples from this class include **?**, who studied the work-effort effects of presenting comparative pay information to student workers at a single university; **?**, who studied the effects on motivation of having employees in a fundraising organization interact with the organization's beneficiaries; and **?**, who, looking at nurses and clerical staff in a single hospital, studied the effects of employee participation in decision making on several factors associated with employees' perceptions of their roles in and influence on the hospital's operations. These experiments present the ideal conditions for interference, since these experiments involve the administration of treat-

ment among individuals who interact regularly, and a considerable span of time between treatment and response measurement.

A class of field experiments that is closely related to those focused on employees working in the same organization regards experiments directed at populations that are connected in terms of an attribute or behavior that is likely to lead to interaction within the population. These can include field experiments that focus on populations of individuals living in the same city or neighborhood (e.g., **?**), or contributing to the the same web forum/blog (e.g., **?**). This category includes several field experiments from political science that focus on voter turnout, and draw upon populations within the same geographic area (e.g., **??**), or are friends on social media (e.g., **?**). Field experiments in this class likely present a lesser risk of interference than those focused on single organizations, as individuals in the same physical or virtual space as broad as a city or website may not necessarily know each other or cross paths. However, the sample sizes in this class of experimental designs tend to be much larger than those focused on single organizations. Due to the larger sample sizes, bias introduced through interference is more likely to result in confidence intervals that do not cover the true population parameters.

Natural social groupings represent attractive targets for field experiments. Subjects are comparable on many important dimensions, which reduces variance in the sample. However, experimentally manipulating a population of individuals who are close in organizational, geographic, or virtual space presents the risk of interference in the effects of the manipulation. Broad categories of social science field experiments, like those we review above, are focused on interactive social groups. In studies such as these, analysis of the effects of the experimental manipulations should relax SUTVA.

4 A Design-Based Test for Network Effects Models

? introduced a method for testing for interference effects, represented by models of network effects, using non-parametric testing under Fisher's inference framework. The model used in testing can include separate parameters for direct causal effects of the treatment and spillover effects that depends on how treatments are allocated across subjects situated in a network. The network effects model is tested against the sharp null hypothesis of no effects—the hypothesis that the treatment assignment has no effect on any unit.

To compare a hypothesized model to the sharp null of no effects, the first step is to define a test statistic to be used to compare outcome distributions of units across experimental conditions. **?** use the Kolmogorov-Smirnov (KS) test statistics to test for the difference across treatment and control groups. If the sharp null is correct, the test statistic should be small in absolute magnitude—indicating little difference across conditions. However, due to the interference among units, we cannot rely upon conventional notions of the sampling distribution of the test statistic to evaluate the statistical significance of differences across conditions.

Random permutations of the treatment vector are used to construct the sampling distribution of the test statistic under the sharp null. In each permutation, a new treatment assignment is drawn from the randomization distribution used in the experiment. Based on the re-randomized treatment vector, the hypothesized parameters and model of effects are used to remove the effect of the treatment on all of the subjects in the experiment. **?** refer to the outcome vector derived by removing the hypothesized effect of the experiment as the uniformity trial. The test statistic is then calculated to assess the differences across experimental conditions. A p-value testing the hypothesized parameter value—with the

alternative hypothesis being the sharp null—is calculated as the proportion of test statistics under permutation that exceed the observed test statistic value (i.e., the test statistic value when evaluated on the uniformity trial given the observed treatment vector). Given the setup of the test, higher p-values indicate greater support for the hypothesized model and parameter values. The intuition for this reversal in the conventional direction of the p-value is that the correct hypothesis will be more effective than any other hypothesized model at removing the differences across experimental categories.

5 Considerations in Testing for Interference

One of the virtues of controlled experiments, in which treatment allocation is randomized, is that the randomization design can be used as the basis for inference in statistical tests (i.e., design-based or randomization-based inference) (**?**). Testing using the Bowers et al. framework still relies on design-based inference, as the stochastic nature of the outcomes is assumed to arise from the distribution based on which the treatment was randomized. However, the hypothesis being tested is formulated as a model of causal direct and spillover effects. As these models of effects are more complicated than the conventional form of effects considered in experiments—a location shift—researchers must put more thought into the functional forms that describe the relationship between the treatment and outcome vectors. It is not possible to enumerate all of the choices available in specifying the model of effects, but we discuss a few salient dimensions below.

- **Network selection**: First, we need to determine the network through which units may be interacting. Consider the case of legislative networks, and the variety of net-

works discussed in this domain by **?** and **?**. Example networks include similarity in roll call voting (**?**), bill cosponsorship (**?**), overlapping committee membership (**?**), collaboration in press events (**?**), co-membership in caucuses (**?**), the proximity of members of Congress' DC offices (**?**), and connections between legislative staffers (**?**). Given a set of prospective networks, such as these, researchers must consider which single network, or combination of networks, will connect all actors and account for maximum spillover. Sometimes this choice may depend on availability of data. In an experiment involving Facebook users, their connections can be extracted. However, in studying the spreading of a treatment encouraging citizens to vote in an election via a mail or door-to-door GOTV campaign, it is difficult to gather reliable network information.

In our re-analysis of experiments on state legislators, we will illustrate these choices by considering the following networks:

  - Ideological network: Using ideological scores calculated based on roll call data
  - Committee networks: Using information about legislative committees on which pairs of legislators have served together
  - Co-sponsorship network: Using information about legislators who have co-sponsored a bill together
  - Geographical network: Using adjacency of districts from which legislators are elected

- **Neighborhood specification**: Once the researcher decides which network—or com-

bination of networks—to use in analysis, it is important to determine the neighborhood within which the effect of treatment can be spreading to control units. For example, **?** find that Facebook users' voter turnout, as expressed on their Facebook walls, influences not only their Facebook friends' turnout decisions, but also turnout of the friends of their friends. This means that the effects of of a Facebook user's turnout decision spread within a neighborhood of two hops through the friendship network. This specification decision becomes more complicated when the network is weighted (as in most of the legislative networks mentioned above). In that case spreading is likely a function of connection strength, but may also disappear at some threshold (e.g., assuming individuals cannot process the influence of all other nodes, especially those very far away in the network). In our consideration of state legislative networks, we specify the neighborhood in two ways when using the ideological similarity networks:

- Entire network: Treatment effect can propagate through the entire network—proportional to ideological similarity—to affect the outcome of control units
- K-nearest neighbors: Treatment effects outcomes of controls only based on their k-nearest neighbors

The definition of neighborhood depends on substantive knowledge about the interaction in a certain network. For example, the number of legislators in a state legislature is relatively small and everyone may potentially communicate with each other. However in looking at interpersonal communication networks, geographic distance may only be a meaningful measure of interaction likelihood for very short distances.

- **Diffusion model specification**: The above two specifications determine which units play a role in the interference reflected in the hypothesized model. Diffusion model specification involves defining how the treatment effect spread. Below we review several considerations in defining the functional form of the diffusion model:

  - The number vs. proportion of treated neighbors as the relevant quantity: Is a control unit influenced by the number of treated units with which it interacts (e.g., as in an epidemic network), or by the balance or proportion of its neighbors that are treated (e.g., as we would assume in a voting or opinion-spreading network). **?** specification assumes treatment spreads as a function of the number of treated neighbors. This specification choice likely comes down to whether the researcher assumes that the treatment and the lack of it are equally powerful forces, or whether change in the outcome can only result from exposure to treated units. and it is important to then account for the control group neighbors that would be offsetting the treatment spillover effect.
  - Form of spread (linear or non-linear): Finally, it is important to determine whether the functional form of propagation of treatment effect should be linear or non-linear. Does the second treated neighbor have the same effect on a node's outcome as the first treated neighbor, or does the effect diminish? Or, alternatively, is it a threshold effect that only manifests when the number of treated neighbors reaches a critical level?

- **Test statistics selection**: To evaluate differences across experimental conditions using the framework of **?**, the researcher must select a test statistic to evaluate the

differences in terms of the hypothesized model of effects. As discussed earlier, **?** uses the KS test statistics, which is problematic for categorical responses, especially those with three or more experimental categories. The test statistic used by **?**—the sum of squared errors from a linear probability model of the binary outcome—is an option for managing categorical outcomes. The Anderson-Darling test is an alternative to the KS test with multiple experimental categories (**?**). Other possible tests include the Mann-Whitney U and Control Median tests (**?**).

6 Research Design

To explore the consequences of interference in field experiments on interactive groups, and illustrate the dimensions researchers should consider in testing for interference, we re-analyze results from a number of field experiments focused on legislators. This work builds directly on build upon **?**. Since it is generally infeasible to recruit legislators for lab experiments, field experiments represent the best option for design-based causal identification of effects in research on legislative behavior. The literature offers many recent examples of field experiments in legislative studies (e.g., **??????**). In these experiments, the researcher introduces a manipulation (e.g., a communication from a constituent, or information about constituent preferences), and then observes legislators' behavior in terms of casting roll call votes or reacting to the communication on an individual basis. Since legislators regularly communicate and collaborate, there is little basis on which to claim that SUTVA holds. Field experiments on legislators represent a broad example of the conditions under which researchers should test for interference effects.

In each of the replications and extensions that follow, we test causal models that include network effects. In order to test these models, we must specify their functional forms and select the data to use in measuring the network. **For each replication, we consider two definitions of ties, two sources of data from which we construct the networks, and two specifications of the ways in which effects manifest on the network.** This exercise illustrates two important methodological considerations. First, network effects are evident in each replication. Second, findings regarding the nature of causal effects depend upon all three dimensions of specification we explore.

6.1 ?

Butler and Nickerson conducted an experiment on New Mexico legislators, to study the effect of learning public opinion from their constituencies. In 2008, a special session of the New Mexico State Legislature was called to vote on a bill regarding proposed spending plans for a budget surplus. A large-scale phone based survey gathered constituent opinions from across the state. Using matched-pair randomization, 35 out of 70 legislators were assigned to the treatment group and sent a letter containing district-specific support for the proposal in their own districts. The original paper conducts a direct comparison of outcomes across treatment and control group, and conclude that the treatment (receiving a letter) significantly affected the likelihood of legislators to vote in favor of tax rebate. Especially, since the proposal was popular, it reduced the likelihood of legislators from low support districts voting in favor of it.

? applied the **?** methodology to test for propagation of treatment in this experiment. These results indicated a significant interference effect. As detailed in section 3, Coppock

used a network based on ideological similarity, where outcome of a control unit is affected by the entire network, and modeled a linear effect of the direct and indirect treatment, on the outcome. We begin by replicating this analysis. In the extension, we consider two types of networks; based on ideological similarity and serving on committees together. We also vary the neighborhood specification to consider an effect from all other legislators—proportional to ideological distance or effect only from k-nearest neighbors. Results of this analysis are presented in section 6.1. The following list summarizes the steps taken to implement Coppock's analysis of interference in the **?** replication.

1. Calculate W-NOMINATE ideology score for each legislator using roll call vote data
2. Calculate ideological similarity as $Similarity_{i,j} = \frac{2-|ideo_i - ideo_j|}{2}$
3. Calculate raw exposure as $Raw\ exposure_i = \sum_{j=1}^{n} Similarity_{i,j} * z_j,\ j \neq i.$
4. Coppock introduces an adjustment for the expected exposures of legislators. Exposures are simulated under a large number of randomizations. Each randomization where legislator i is in treatment is indexed as k (k = 1, 2, ..., K) and where legislator i is in control is indexed as l (l = 1, 2, ..., L)

$$Expected\ exposure_{i,z_i=1} = \frac{\sum_{k=1}^{K}\sum_{j=1}^{n} Similarity_{i,j} * z_{j,k}}{K},\ j \neq i,\ z_{i,k} = 0$$

$$Expected\ exposure_{i,z_i=0} = \frac{\sum_{l=1}^{L}\sum_{j=1}^{n} Similarity_{i,j} * z_{j,l}}{L},\ j \neq i,\ z_{i,l} = 1$$

5. I imagine this is where we will first mention not using the adjustment factor. Hence adding the bullet ..... Therefore, the raw exposure for each unit is used to evaluate

the indirect effect.

6. Using the hypothesized parameter values, remove direct effects of treatment and indirect effects of residual exposure based on a linear regression form.

7. Regress hypothesized uniformity trials on direct and indirect treatment, and use the residual sum of squares (RSS) as test statistic.

8. p-value for each hypothesized treatment effect is the proportion of simulated RSS that is lower than the observed RSS (note that smaller RSS indicates more variance explained).

6.2 ?

The second dataset we work with comes from an experiment on the Michigan state legislators. This experiment was conducted on legislators from both houses, in the context of an anti-bullying legislation. Units were stratified based on various background variables. The treatment was receiving calls from constituents, expressing their views on the proposed bill. Treatment was given in three different doses, which differed in the number of calls places to the given legislator. Once again, the authors conducted an analysis under SUTVA and concluded that this treatment has a significant effect on the outcome in form of the final vote on the bill. They observed a 12 percentage point increase in the likelihood of voting in favor of anti-bullying bill, for those treated. Interference effects are not accounted for in the original analysis. Therefore we begin by building a model similar to the one in **?** and extend the analysis to consider an alternate neighborhood specification, that assumes effects to spread through k-nearest neighbors.

6.3 ?

The final dataset also originates from a field experiment on state legislators. This study aimed to understand whether politicians behave differently based on the expected electoral incentive. The question of whether there exists differential intrinsic motivation to help a constituent based on his or her race is addressed by studying all state legislators in the US serving in mid-November 2010. These 6,928 legislators received an email from an alias Tyrone Washington who was seeking help filing for unemployment benefits. Treatment in this case was whether Tyrone was based within or far from legislator's district. Legislators were block-randomized based on the state, party, and race to receive one of the two possible treatments. Analysis using Coarsened Exact Matching (CEM) observed a difference-in-difference estimate of 18.5 percentage points which was highly significant. Also supported by results from an OLS and logistic regression model, the paper concluded that extrinsic motivation guided response rates from non-black legislators, and the actions of black legislators were less affected by strength of political incentives. We introduce interference effects in the analysis of this experiment by setting up a model similar to the one in **?**, where three separate networks are considered. Adhering to the block-randomization scheme, neighborhood specification takes a block-diagonal form assuming that effects spread through the entire network within each state.

7 Analysis/Results

In this section we present results from replication and extension of the analysis in both studies. For each study, we present plots of p-values for the different model specifications/-

parameterizations and then summarize point estimates and confidence intervals at the end of the sections. Following **?**'s (**?**) recommendation, the parameter vector corresponding to the maximal p-value is considered the point estimate, and 95% confidence intervals represent bounds around each parameter that contain all values that correspond to a p-value greater than 0.05, fixing the other parameter values at their point estimates.

7.1 Results for ? data

The data for this replication was obtained from the publicly available data repository created by the authors. Records of standing committee membership in the 16 standing committees in place during the 2008 regular session was obtained from the New Mexico Legislative Council Service Librarian.

The p-value plot for replication of the main analysis in **?** is in Figure 1. The p-value is highest (0.997) when the direct effect is -0.25 and indirect effect is -0.15. Negative effects indicate that the treatment reduced the likelihood of voting in favor for those who received the treatment directly as well as those to whom it propagated, due to ideological similarity. Confidence intervals are drawn using dashed lines. Despite slight differences, likely owing to simulation error, these results align with Coppock's finding that the effects of the experiment included direct effects and interference.

The first extension we consider is a change in the neighborhood specification. Instead of looking at ideological similarity across the network, we consider only the nearest k neighbors at values k = 3,5,8,12. In Figure 2, we see that the direct and indirect effects which maximize the p-value, are lower (in absolute value) than those in the first specification. When we model interference between all legislators in the network, we observe

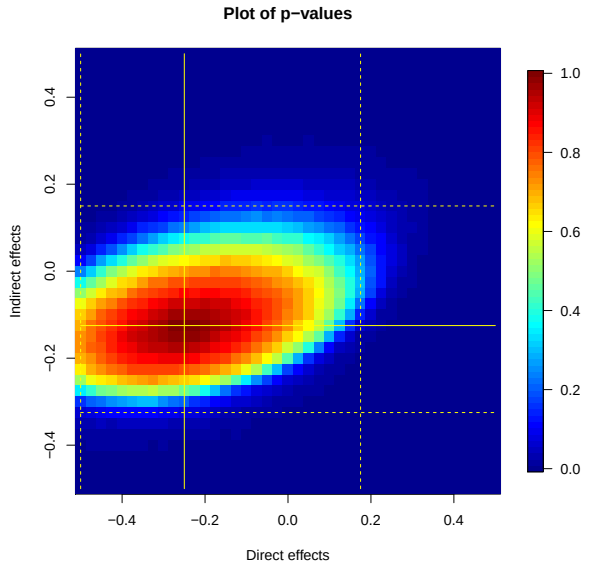

Figure 1: Main analysis for **?** data

higher spillover than when looking at only nearest neighbors. We see that the observed indirect effect is higher, when considering a a broader neighborhood (the entire network being the broadest neighborhood we can consider). Interestingly, we only see a result that is statistically significant, based on the 95% confidence interval, when looking at the indirect effect with a neighborhood defined as the twelve nearest neighbors.

In the next extension, we change the network itself, by considering committee network. Here, an undirected edge exists between legislators who served on standing committees together. We define the network at two thresholds—serving on at least one committee together and serving on at least two committees together. The results indicate that the committee network also carries the effect of treatment to control units. However, it is smaller in magnitude than that through ideological network. Indirect effects propagated through the committee network are not significant.

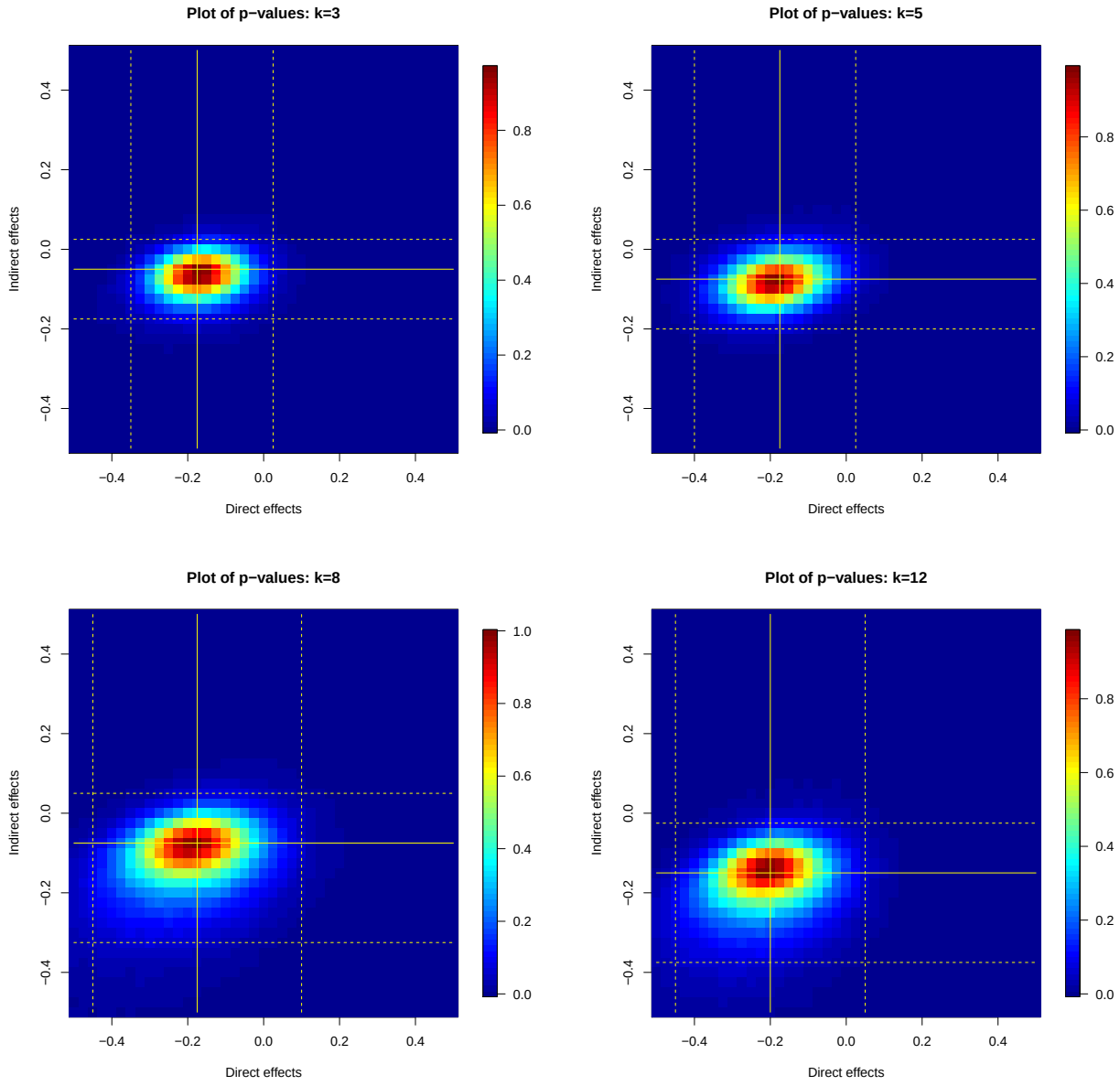

Figure 2: k-nearest ideological neighbors for **?** data

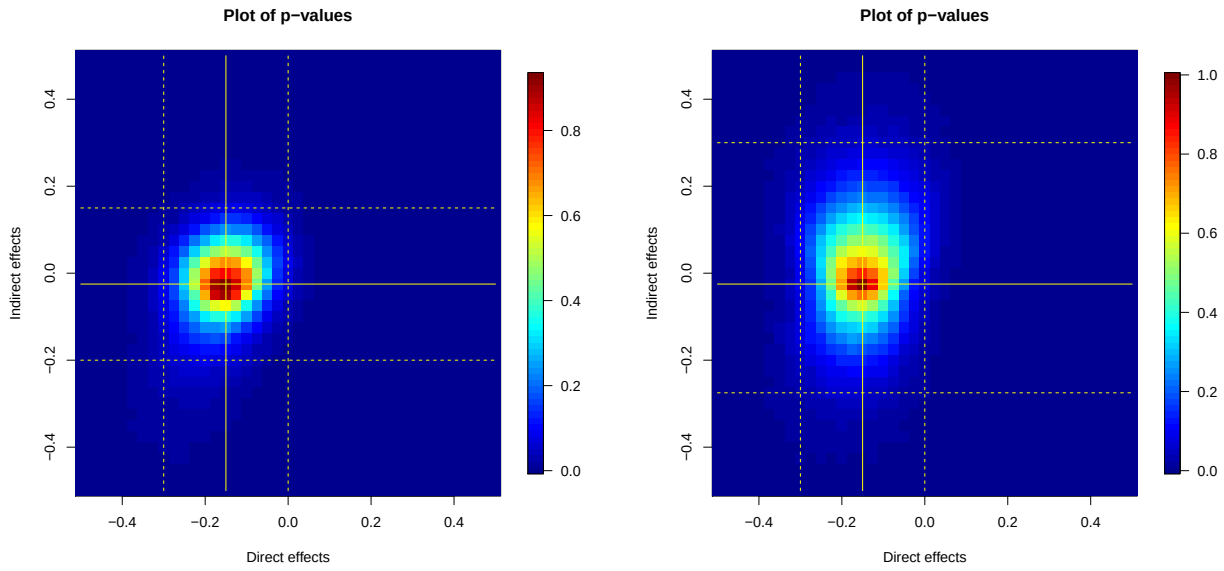

Figure 3: Committee network for **?** data

Table 1 summarizes the results for analyzing the **?** data under various specifications. We notice that the indirect effect is always negative. However, both the indirect and direct effect changes in magnitude depending on the specification. This emphasizes the importance of choosing a network that can best explain the phenomenon. Given that tax rebate is a fairly partisan topic, ideological similarity can explain information spillover better than mere membership on the same committees, which is likely to be more related to seniority and non-ideological consistency composition. The point estimate of the indirect effect being negative implies that as their exposure to the treatment increases, the likelihood of a legislator to vote in favor of the tax rebate decreases.

Table 1: Results from **?** data: Raw exposure for all analysis except Full Ideological Network

Model	Direct effect		Indirect effect	
	Estimate	95% CI	Estimate	95% CI
Ideology: full network	-0.175	(-0.35, -0.025)	-0.075	(-0.5, 0.5)
Ideology: 3nn	-0.175	(-0.35, 0.025)	-0.05	(-0.175, 0.025)
Ideology: 5nn	-0.175	(-0.4, 0.025)	-0.075	(-0.2, 0.025)
Ideology: 8nn	-0.175	(-0.45, 0.1)	-0.075	(-0.325, 0.05)
Ideology: 12nn	-0.2	(-0.45, 0.05)	-0.15	(-0.375, -0.025)
Committee: >0	-0.15	(-0.3, 0)	-0.025	(-0.275, 0.3)
Committee: >1	-0.15	(-0.03, 0)	-0.025	(-0.2, 0.15)

7.2 Results for ? data

This data has not been analyzed for indirect effects before. However, for all the reasons that we would expect to see interference in the **?** results, we would expect to see them in the **?** results. We conduct an analysis similar to that in **?**, where a network is constructed based on ideological scores of legislators, using roll call data. We find evidence of indirect effect in this analysis as well. However, these estimates have broader confidence intervals and none of them are significant. It is possible that we can attribute this to the nature of the bill. Voting behavior on an anti-bullying bill may not be governed by ideological coalitions. In the next iteration of this paper we plan to consider other networks such as the standing committee network. Figure 4 shows the plot of p-values for this analysis.

Figure 5 depicts results of analyzing the **?** data under the ideological network and considers k nearest neighbors (k = 3,5,8,12) based on ideological similarity to constitute the neighborhood. We see that the results regarding indirect effects are fairly similar and show evidence of interference effect.

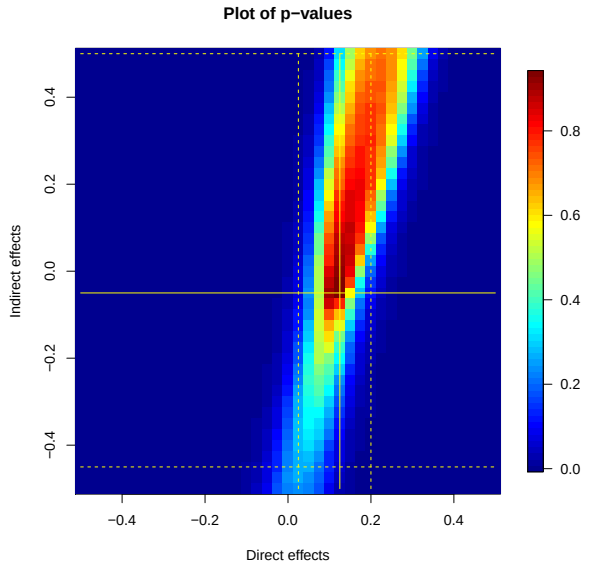

Figure 4: p-values: main analysis for **?** data

Table 2 summarizes results of the **?**. The point estimates are negative across neighborhood specifications under the ideological network, and positive for cosponsorship network. This indicates that as exposure through ideologically similar neighbors goes up, the likelihood of a legislator to vote in favor of the anti-bullying bill goes down, and this likelihood goes up if we consider connectivity via the co-sponsorship network. None of these effects are significant though, and as mentioned earlier, the issue addressed in this bill is not necessarily partisan. Therefore, we hesitate to conclude that talking to legislators with similar ideology makes a legislator more likely to favor bullying. We see no evidence of spillover when 12 nearest ideological neighbors are considered.

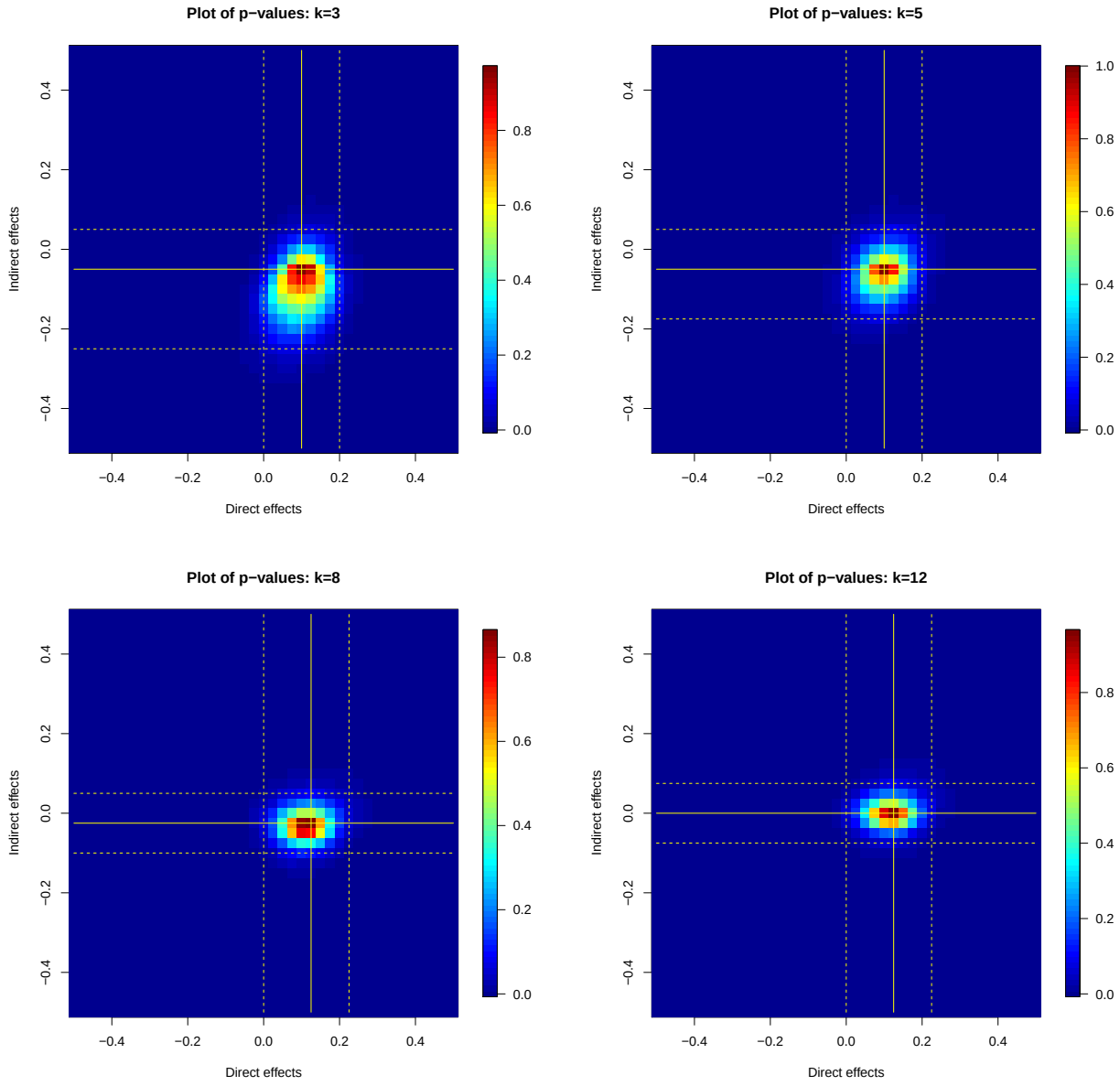

Figure 5: p-values: k-nearest neighbors for **?** data

Table 2: Results from **?** data: Raw exposure

Model	Direct effect		Indirect effect	
	Estimate	95% CI	Estimate	95% CI
Ideology: full network	0.125	(0.025, 0.2)	-0.05	(-0.45, 0.5)
Ideology: 3nn	0.1	(0, 0.2)	-0.05	(-0.25, 0.05)
Ideology: 5nn	0.1	(0, 0.2)	-0.05	(-0.175, 0.05)
Ideology: 8nn	0.125	(0, 0.225)	-0.025	(-0.1, 0.05)
Ideology: 12nn	0.125	(0, 0.225)	0	(-0.075, 0.075)
Cosponsorship	0.125	(0.025, 0.2)	0.075	(-0.5, 0.5)

7.3 Results for ? data

We now discuss the last application considered in this paper. As discussed in the earlier section, the author of this paper analyzed the data assuming SUTVA. However, it would be reasonable to expect that legislators within each state communicate with each other leading to spillover of treatment effect. We stipulate three possible networks through which interference may occur. These depend on two key covariates; Percentage of Democratic Vote (*demvotepercent*) in the district, and Percentage of Black Legislators (*blackpercent*) in the district. We create one network for each of the two covariates and a third combining the two. In networks based on individual variables, similarity score for legislators i and j based on variable X is defined as in Equation (1).

$$Similarity_{(i,j)} = \frac{2 - |x_i - x_j|}{2} \tag{1}$$

For the network based on two covariates, network is defined as the Euclidean distance between legislators i and j where *demvotepercent* (X) and *blackpercent* (Y) are equally weighted, as shown in Equation (2).

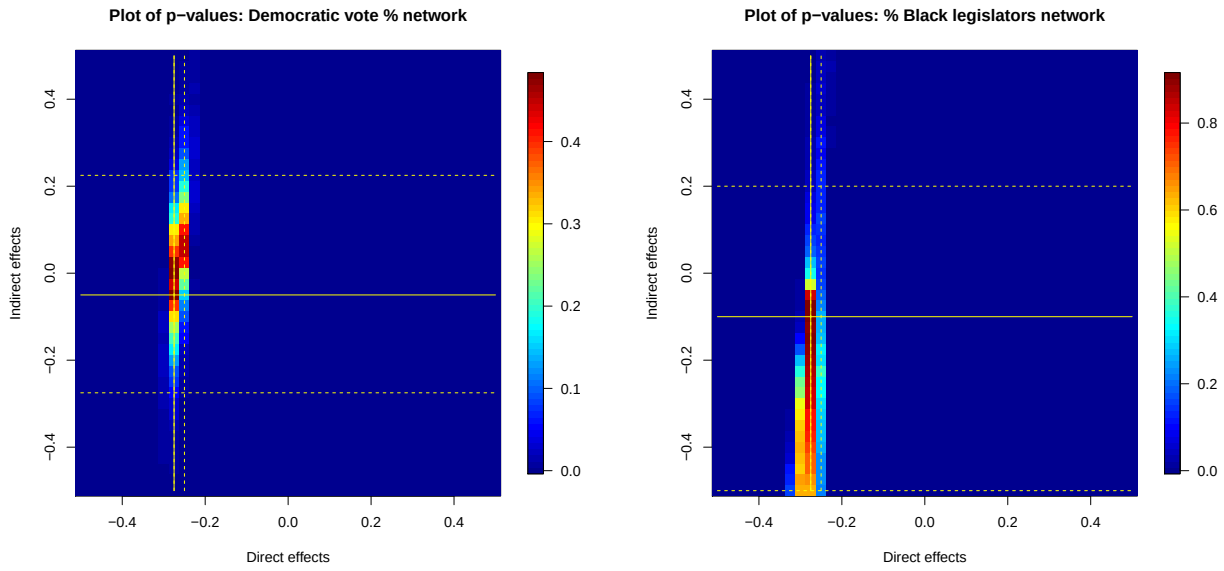

Figure 6: p-values for **?** data

$$Similarity_{(i,j)} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \tag{2}$$

In these block-diagonal networks, neighborhood of any legislator consists of all other legislators in his/her state. The closer any two units are on values of one of these variables, the stronger the tie, and higher the exposure to receiving indirect treatment. Figures 6 and 7 depict results of analyzing the **?** data under the combined network and the individual networks respectively. We see that all of these three specifications show evidence of spillover effect.

This is further detailed in table 5. The non-zero point estimates indicate spillover of treatment to control units. It is interesting to note though, that these values have different signs. There is positive spillover through individual networks and negative via the com-

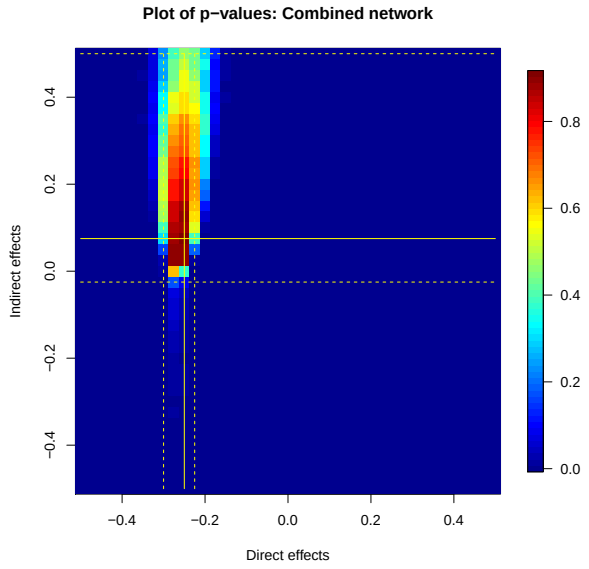

Figure 7: p-values for **?** data: Combined network

bined networks. Each of the confidence intervals crosses over zero, therefore the effects are not significant. However, there is still value to interpreting the effects. When exposure is via legislators with similar values of *demvotepercent* and *blackpercent*, the likelihood of the legislator to respond to Tyrone reduces. In contrast, as exposure through the combined network increases, the likelihood of responding to Tyrone increases. However, the setup of the similarity score is such, that for a given state, the number of legislators with either high or low *demvotepercent* and *blackpercent* can dominate the overall effect.

8 Discussion

The results from our replication of field experiments on legislatures underscore the importance and complexity of accounting for interference in field experiments on interactive

Table 3: Results from **?** data: Raw exposure

Model	Direct effect		Indirect effect	
	Estimate	95% CI	Estimate	95% CI
Democratic vote percentage	-0.275	(-0.275, -0.25)	-0.05	(-0.275, 0.225)
Percentage of black legislators	-0.275	(-0.275, -0.25)	-0.1	(-0.5, 0.2)
Mixture network	-0.25	(-0.3, -0.225)	0.075	(-0.025, 0.5)

groups. The replications and extensions of **?** and **?** demonstrate both the presence of interference effects, and the inferential consequences of choices in specifying both the network and the neighborhood through which treatment is hypothesized to propagate. These two applications are focused on roll call votes as the outcomes. These are important legislative actions, but may be constrained in terms of their susceptibility to experimental manipulations in that they are heavily centralized and subject to open and widespread scrutiny. The **?** study, in contrast, considers the effects of the manipulation on responses to individualized requests—requests that were initialized as part of the experiment. Results from this analysis emphasize the importance of thinking about not only the possible network, but also the resulting interpretation of the indirect effect parameter.

In the next iteration of this paper, we plan to extend the analysis in three directions. First, we will gather additional network data to use in testing for interference in—at the very least—the **?** replication. I don't remember why we emphasized on Bergan. Maybe because the indirect effects were zero when looking at expected exposure? One important caveat regarding our replications is that we do not have access to the multitude of relational data that scholars have used to define legislative networks. We will make an effort to gather additional legislative network data to use in our replications. Second, we have not yet considered the count vs balance dimension in the specification of the interference

models in our replication. Incorporating this dimension will allow us to compare models in which connections to control units counteract the effects to treated units, as compared to the current results, which attribute interference solely to the number of of treated units in a node's neighborhood. In the earlier draft, Broockman analysis was the third proposed change. However, that is now a part of the main paper. So we can select something from the first section. Maybe we look at a different outcome variable, or a different test statistic?

9 Conclusion

We make the case that many common domains in which political scientists employ field experiments present conditions under which we would expect interference, which represents a violation of SUTVA. These domains concern experiments on interactive social groups. We review the use of recently developed methods for testing causal models in the context of interference, and present several dimensions that researchers should consider in specifying causal models that involve interference. Through a broad replication and extension exercise focused on field experiments on legislatures, we illustrate consideration of these dimensions and show that evidence for interference exists, and depends upon the choices made in specifying the causal models.

Our arguments and results lead to three important implications in future research. First, when analyzing results from field experiments that target populations within which there is a risk of interaction, and therefore interference, researchers must consider the consequences of interference for their inferences. Second, our results underscore the opportunities for researchers to develop new theoretical inquiries focused on interference. Field ex-

periments on interactive groups have become commonplace in research designs in political science. Furthermore, we now have methods available for testing interference hypotheses. These two factors combine to open the door to principled empirical and experimental tests of interference hypotheses. Third, since the optimal experimental design depends upon the network and hypothesized model of effects (**?**), researchers should incorporate considerations regarding interference before fielding their experiments.