Estimating Heterogeneous Spillover Effects

Daniel Kent*

July 13, 2020

Abstract

In social environments, interference between units is likely the norm, not the exception. This poses a problem for estimating causal effects, where the potential outcomes framework assumes that one unit's treatment assignment has no effect on another unit's outcome. In response to this concern, one increasingly popular approach for handling interference between units is the estimation of spillover effects, where sharing a networked tie to a treated unit confers indirect treatment exposure. However, like average treatment effects, there are good reasons to expect that spillovers vary in magnitude and direction across contexts. In order to capture this variation, I approach spillovers through the lens of heterogeneous treatment effects, which can be modeled with causal random forests. After formally discussing a research design that combines these techniques, I apply the combined procedure to a study about the efficacy of school programs meant to decrease bullying.

*kent.249@osu.edu; Ph.D. Candidate, Department of Political Science, The Ohio State University

1 Introduction

In many social settings, researchers are interested in estimating the causal effect of a treatment. These causal estimates are generally produced by comparing the average outcomes for units that receive treatment and units that do not. While many concerns arise in observational settings when comparing treatment and control groups, two prominent threats to the utility of a straightforward comparison of means are: heterogeneous treatment effects and treatments spilling over from one unit to another. In cases of the former, given sufficiently heterogeneous effects, estimates of average effects will not map onto the actual group-level effects. In cases of the latter, comparing treated to untreated units will return a biased estimate, lowering an effect estimate because observations in the control group will be mistakenly labelled as having no treatment exposure of any kind.

While techniques have been developed for both situations separately, what if spillover effects are present *and* they are heterogeneous? Across applications such as get-out-the-vote (GOTV) programs, foreign aid provision, and school programs, spillover effects are a widely-recognized reality, where observations tied to treated observations in some way receive indirect treatment exposure. In addition, these applications often exhibit heterogeneous treatment effects, with the treatment's magnitude and direction varying widely across contexts. It therefore follows that, across domains, spillover effects should not only be considered, but also checked for effect heterogeneity. Yet, while heterogeneous treatment and spillover effects are methodologically important and potentially simultaneously present in many contexts, little guidance exists on how the two can be combined in a causal inference framework.

This paper discusses and presents an approach that combines recent advances in both literatures. The proposed method works in three steps: defining potential outcomes in terms of both direct and indirect treatment exposure, weighting observations by their probability of receiving indirect treatment exposure, and then incorporating these potential outcomes and sample weights into a causal random forest. In terms of disaggregating potential outcomes, Aronow et al. (2020) demonstrate that once units are differentiated by their treatment sta-

tus and whether they share a social tie to a treated unit, then indirectly exposed units can be compared to units with no indirect exposure, providing a causal estimate of the sample's spillover effects. This comparison is equivalent to calculations of an average treatment effect if no spillovers are present, meaning comparing groups according to their spillover conditions is compatible with methods for identifying heterogeneous treatment effects. (Imai and Ratkovic, 2013; Wager and Athey, 2018, e.g.,) The two primary differences between the proposed and existing heterogeneous treatment effect estimation procedures are that 1) indirect treatment exposure is substituted for directly receiving treatment and 2) the two groups under comparison are control group units in an experimental setting that are indirectly exposed and unexposed.

After reviewing the relevant methods separately, I further discuss how the two are compatible theoretically and in practice. After formally discussing the relevant measures, I demonstrate how available R packages can be used in tandem to estimate conditional average treatment effects, coupled with a monte carlo simulation to demonstrate the software combination provides accurate results. The simulation also returns strong support for including sample weights for the probability of indirect treatment exposure when estimating a causal forest for spillover effects. Lastly, I apply the combined methods to a study on spillovers from anti-bullying programs in school settings, where treatment effects are heterogeneous across schools, demonstrating how a pilot study can inform expectations of where subsequent efforts will be effective and where they are likely to produce undesired outcomes. The application also demonstrates how outcomes can be disaggregated to differentiate between trends due to direct and indirect treatment.

2 Related Work

This paper seeks to synthesize, from a causal inference perspective, two popular methodological research topics: heterogenous treatment effects and spillover effects. In many contexts,

spillover effects and effect heterogeneity are plausibly both present, making accurate estimation important for academic research, product design, and policy implementation. Both methodological literatures are growing and tend to approach estimation procedures from the potential outcomes framework, where causal effects are understood to represent the difference between a unit's outcome if that unit does and does not receive treatment. More specifically, the two recent methodological developments – with accompanying software – that drive this proposed approach are Aronow et al. (2020) and Wager and Athey (2018). From a technical perspective, this article's contribution is that it provides a theoretical and applied guide to combining the methods effectively. Indeed, while developed separately, the two methods can be viewed as complementary. I formally discuss the two approaches in turn before discussing how the two methods can be synthesized.

2.1 Spillover Effects

A fundamental concern when estimating causal effects is the presence of "interference" between units, a class of instances where one unit's outcome is not only a result of its treatment status, but also the treatment status of other units. (Hudgens and Halloran, 2008; Rubin, 1990; Taylor and Eckles, 2018; VanderWeele and Tchetgen, 2011) One way that interference occurs, and this paper's focus, is through spillover effects, where indirect treatment exposure is conferred through a tie linking two units.[1] In these situations, treatment literally spills over from one unit to another through their shared tie. These spillovers can be geographic[2], social[3], or through linked processes in a complex system.[4] I mention these different settings because, while I discuss spillover effects specifically in terms of social networks throughout the rest of this paper, the proposed framework is readily adjustable to other contexts.

This paper's proposed method draws primarily upon Aronow et al. (2020), who frame

[1] Recent examples of studies estimating spillovers include Baicker (2005); Cheung and Ping (2004); Haushofer and Shapiro (2018); Jones et al. (2017); Ng (2000); Nickerson (2008); Paluck et al. (2016)

[2] E.g., A change in one location impact neighboring locales.

[3] A change to one person can impact their friends and family.

[4] Consider the ongoing public health pandemic, which was shortly followed by a massive drop in oil prices.

spillovers through four experimental conditions. Each unit's condition can be derived as long as treatment assignment and the adjacency matrix of ties between units are known. Given knowledge of ties and treatment, units can be differentiated based on whether they received treatment and/or indirect exposure through a potential spillover. Aronow et al. label these conditions the "exposure mapping", with four possible conditions:

- d_{11}: Direct and indirect treatment exposure
- d_{10}: Isolated direct exposure
- d_{01}: Indirect exposure
- d_{00}: No exposure

Under these conditions, indirect treatment exposure occurs when a unit shares a networked tie with a unit that received direct treatment exposure. A unit's potential outcome is then labelled $y_i(d_k)$, as opposed to the common notation $y_i(1)$ and $y_i(0)$ for whether a unit is in the treatment or control group. For example, in the later example of spillover effects from anti-bullying programs in schools, indirect exposure occurs when a student is a friend with a student that was assigned to an anti-bullying program, where the anti-bullying program is the treatment of interest. However, in order to estimate these indirect effects, both the school's friendship network and the students assigned to the anti-bullying program must be known.[5]

The categories then allow the estimation of $\tau(d_k, d_l)$, or the causal effect of being in exposure condition d_k rather than exposure condition d_l. In terms of spillover effects, two comparisons are especially relevant: 1) $\tau d_{01}, d_{00}$ – comparing units with indirect and no exposure in the control group and 2) $\tau d_{11}, d_{10}$ – comparing units in the treatment group that also receive indirect exposure to those only with direct treatment. These comparisons allow us to move from only comparing average outcomes in the control and treatment groups to also comparing units within each direct treatment condition based upon indirect exposure. The most straightforward version of these spillover effects, and my subsequent focus, is on the first comparison:

[5] Or the social network and treatment assignments must be estimable from the available data.

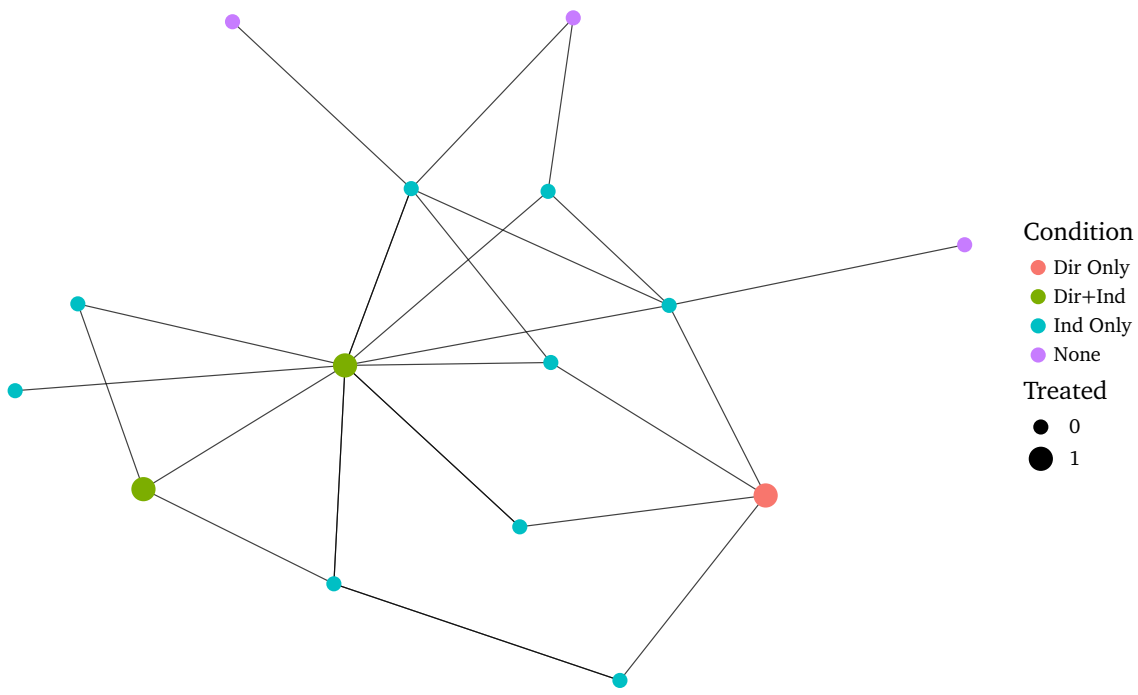

Figure 1: Example network, where larger nodes are the treated units and ties are pathways through which indirect exposure can occur. Node color represents the exposure condition, spanning: no exposure (purple), indirect exposure only (light blue), direct treatment only (red), and both direct treatment and indirect exposure (green).

units in the control group with indirect exposure and units in the control group with no exposure. However, for any comparison of groups k and l, then Aronow et al. demonstrate that the average causal effect can be estimated with:

$$\tau(d_k, d_l) = \frac{1}{N}\sum_{i=1}^{N} y_i(d_k) - \frac{1}{N}\sum_{i=1}^{N} y_i(d_l) = \mu(d_k) - \mu(d_l). \tag{1}$$

where $\mu(d_k) = \frac{1}{N}\sum_{i=1}^{N} y_i(d_k)$ is the average potential outcome for units with exposure k. Equation 1 is therefore a difference of means, just with a more fine-grained set of potential outcomes than the usual formulation of an average treatment effect.

In most settings, however, even if treatment is assigned at random, the preexisting social ties are not random. This means the probability of $y_i(d_k)$ is not equal across all units. As a solution, Aronow et al. (2017) propose estimating π_i, or observation i's *generalized probability*

of exposure. For our framework, $\boldsymbol{\pi}_i = (\pi_i(d_{11}), \pi_i(d_{10}), \pi_i(d_{01}), \pi_i(d_{00}))$. Once these probabilities are estimated, they can then be used as weights for the observed outcomes, approximating "as if" random assignment to the observed potential outcome for each unit.[6] In order to estimate $\boldsymbol{\pi}_i$ for each unit, Aronow et al. (2020) provide software that calculates the range of possible treatment assignments, denoted Ω, and then calculates the proportion of times that each unit i finds itself in exposure condition d_k for the observed network.[7] Given a sufficiently large set of treatment assignments, where $|\Omega|$ is high, treatment vectors are sampled finitely to approximate the range of values in Ω.[8]

The resulting probability estimates of each unit i being in exposure condition d_k can then be used for the Horvitz-Thompson inverse probability estimator:

$$\widehat{y^T_{HT}}(d_k) = \sum_{i=1}^{N} \boldsymbol{I}(D_i = d_k) \frac{Y_i}{\pi_i(d_k)}. \tag{2}$$

Once each unit's observed outcome is inversely weighted by its estimated probability of being in the observed exposure condition, we can combine Equations 1 and 2 to produce:

$$\widehat{\tau_{HT}}(d_k, d_l) = \widehat{\mu_{HT}}(d_k) - \widehat{\mu_{HT}}(d_l) = \frac{1}{N}\left[\widehat{y^T_{HT}}(d_k) - \widehat{y^T_{HT}}(d_l)\right]. \tag{3}$$

where $\widehat{\tau_{HT}}(d_k, d_l)$ is the Horvitz-Thompson weighted estimate for the average causal effect of being exposed to k rather than l. In this sense, Aronow et al. (2020) argue that spillover effects can be estimated through the familiar difference-of-means framework once exposure conditions are calculated, but that the observed potential outcomes should also be inversely weighted by probability estimates.

Because Equation 3 is a difference of means with added weights, it is readily applicable to the methods I discuss next about heterogeneous treatment effects. Across techniques, heteroge-

[6] This approach is similar to that employed in Ugander et al. (2013).

[7] See https://github.com/szonszein/interference for the exact code and an example on p11 of Aronow et al. (2020).

[8] While these probabilities are currently estimated only using the observed ties, an interesting next step in this literature could be estimating exposure weights by directly modelling a network's tie-formation using tools such as exponential random graph models (ERGMs) or other inferential network analysis tools. (e.g., Cranmer et al., 2017; Hunter et al., 2008; Minhas et al., 2019; Victor et al., 2017; Wilson et al., 2017)

neous treatment effect estimators calculate the difference of means across potential outcomes, conditional on covariate values. However, I also find in monte carlo simulations that applying the Horvitz-Thompson weights from Aronow et al. (2020) almost always increases the accuracy of heterogeneous treatment effect estimators when they are applied to spillover effects. I discuss the proposed method for heterogeneous treatment effects next.

2.2 Heterogenous Treatment Effects

Machine learning methods are increasingly popular tools for uncovering the presence of heterogeneous treatment effects, with a recent focus on forest-based algorithms. (Athey and Imbens, 2016; Green and Kern, 2012; Grimmer et al., 2017; Hill, 2011; Hill and Su, 2013; Hill et al., 2020; Wager and Athey, 2018) Generally speaking, there are two approaches to modeling effect heterogeneity. Given a theoretically informed set of moderators (James and Brett, 1984), one can denote and compare outcomes across subsamples or fit an interaction term in a regression. (Braumoeller, 2004; Brambor et al., 2006; Esarey and Sumner, 2018, e.g.,) Alternatively, a situation of interest may be characterized by unknown heterogeneity across an unknown number of variables. Here, machine learning algorithms can be used to sort through the possible moderators with either the aforementioned forest-based methods or a form of variable selection (Imai and Ratkovic, 2013), using the available data to assess moderator plausibility. Considered this way, machine learning approaches to uncovering effect heterogeneity can be incredibly valuable for taking a pilot study or A/B test and uncovering where the proposed policy or product change has the largest or smallest predicted effects when implemented broadly in one's population of interest.

This paper builds off Athey and Imbens (2016) and Wager and Athey (2018), who develop a procedure for estimating heterogeneous treatment effects with "causal forests" fit through a two-stage 'honest' approach. In the honest approach, when training a model, half of the data is used to construct each tree's splits and the other half of the data is used to assess the causal effects assigned to observations that fall in each tree's terminal leaves. While this comes at the

cost of not using all available information when building the structure of the component trees, Wager and Athey demonstrate two substantial benefits in return. First, the honest approach enables observation-level standard errors. Second, it lowers the probability of overfitting, which in this framework occurs when outlier observations have disproportionate influence on the final causal effect estimates. Importantly, on the latter, random forests are generally recognized to have a tendency to mistake high-variance noise for the data's true systematic trend because doing so decreases training set predictive error. In response, calculating each leaf's predictions on data not used for training lowers this risk, because outlier observations can only influence one portion of the training process, not both. For example, while an outlier observation may disproportionally influence the covariate values where splits occur, it will not subsequently inform the prediction made for out-of-sample observations that fall under the resulting leaves. Indeed, this 'honest' aspect of the model is not uniquely helpful for causal forests and has been demonstrated to improve test-set predictive accuracy for a range of forest-based algorithms.

Beyond the honest sampling approach, causal forests are different from random forests because they are not used to maximize within-leaf predictive accuracy. Rather, splits are chosen on covariate values that maximize the difference between the average treatment effect estimates in the two resulting nodes. As a short aside, in a random forest, a model's accuracy can be evaluated using test set data where true outcomes are known and can be compared to predictions, allowing a straightforward verification of a model's predictive accuracy. However, causal effects are never observable because of the fundamental problem of causal inference, so the causal forest's predictive accuracy cannot be confirmed by comparing unit-level predicted causal effects to the true causal effects, which are unobserved.[9] Returning to the causal forest's mechanics, by differentiating average treatment effects across nodes through covariate splits, the causal forest estimates conditional average treatment effects (CATEs):

$$\tau(x) = \mathbb{E}\left[Y^{(1)} - Y^{(0)} \mid X = x\right] \tag{4}$$

[9]This discussion is meant to raise a sense of empirical caution with these models. While they are verified to work in simulation, their final estimates should be carefully assessed.

giving us a difference of conditional means at a proposed covariate value.

Formally, the method begins with n units, where a tuple (X_i, Y_i, W_i) is observed with: feature vector $X_i \in \mathbb{R}^p$, response $Y_i \in \mathbb{R}$, and treatment assignment $W_i \in \{0, 1\}$. The causal forest then estimates heterogeneous treatment effects by fitting a series of decision trees under the following steps:[10]

1. Draw, with replacement, a random subsample from the dataset
2. Repeatedly split the root into child nodes repeatedly, with the following steps:
    - Select a random subset of variables
    - At each variable x, possible values for splitting, v are considered. Each potential split (x, v) is evaluated by how much it increases heterogeneity in the resulting CATE estimates across leafs.
    - Observations within the splitting variable x that are less than or equal to v are placed in the left resulting node and values greater than or equal to v in the right resulting node.
    - If a node lacks valid splits or potential splits do not improve fit, a node is not split any more and considered a leaf.

However, since we are working in the honest framework, the independent half of the training dataset not used for model building then determines the estimated causal effects for observations in the terminal nodes for each tree, where the estimated CATE for any observation in a node is the average difference between its control and treated units. These leaves allow for estimating complex moderating effects for each observation, producing a final set of unit-level causal effect estimates, determined by where that observations falls in the causal forest. If one is more interested in how effects vary across groups, then these observation-level estimates can be aggregated by groups and then compared.

[10] This explanation draws on the accessible tutorial at https://grf-labs.github.io/grf/REFERENCE.html.

More formally, when building the forest, Wager and Athey denote the conditional average treatment effect of a potential split at $X = x$ as:

$$\hat{\tau}(x) = \frac{1}{|\{i : W_i = 1, X_i \in L\}|} \sum_{\{i:W_i=1, X_i \in L\}}^{Y_i} - \frac{1}{|\{i : W_i = 0, X_i \in L\}|} \sum_{\{i:W_i=0, X_i \in L\}}^{Y_i} \tag{5}$$

where Y_i is the outcome, W_i is treatment assignment, and $i \in L(x)$ are indices for the data X which decide the split. $|\{i : W_i = 1, X_i \in L\}|$ and $|\{i : W_i = 0, X_i \in L\}|$ denote the number of observations in both conditions (treated vs untreated for the proposed covariate split). Once the ensemble of trees are fit, the causal forest takes the aggregated ensemble, where each tree's estimated causal effect for a covariate condition, x, can be expressed as $\hat{\tau}_b(x)$ and averaged for the whole causal forest, giving us:

$$\hat{\tau}(x) = \frac{1}{B} \sum_{b=1}^{B} \hat{\tau}_b(x). \tag{6}$$

Equation 6 is therefore the estimated causal effect of receiving treatment (or spillover in our context), conditional on covariates $X = x$.

In summary, the causal forest predicts the difference between each observation's potential outcomes, based on its covariate values. Generally the terms 'predict' and 'causal' are not used in tandem. However, in this case, the causal forest's goal is to predict the causal effect of treatment on an observation, conditional upon that observation's covariate values. This is accomplished by building a random forest where each trees' splits are decided upon by splitting at the covariate value that maximizes the difference between the average treatment effect in each resulting node. Each tree within the final ensemble provides a predicted treatment effect for observations that fall within the final leafs, where the causal forest makes final predictions by averaging each tree's predictions. These final predictions are produced at the unit-level, where each observation's predicted causal effect of treatment assignment is calculated by the final leaf it is placed in within the component regression trees. In the next section I outline how this approach can be merged with the aforementioned discussion of spillover effects to estimate heterogeneity in spillover effects.

3 Method

In this section I discuss how the two previously discussed approaches to spillover effects and heterogeneous treatment effects are directly compatible and propose a workflow for combining the two in applied research. First, I demonstrate how the two are at their core a comparison of average outcomes across two groups, differentiated by a causal variable (which can be direct treatment or indirect treatment exposure through a networked tie). Second, I discuss the implementation of the Horvitz-Thompson inverse probability weights in a causal forest, accounting for spillover assignment being non-random, even if treatments are assigned at random. Lastly, I discuss how to evaluate output from a causal forest in terms of spillover effects and their heterogeneity.

Spillover effects and heterogeneous treatment effects are both at their core about comparing means across groups that do or do not receive exposure to a causal variable (treatment or spillover). Indeed, although the notation in the aforementioned papers is more complicated, due to the nuances of the groups being compared and estimating sample weights, the core formulation of an average treatment effect is the foundation of both approaches:

$$\text{ATE} = \mathbb{E}[y|t = \text{treatment}] - \mathbb{E}[y|t = \text{control}]. \tag{7}$$

In the potential outcomes framework, an average treatment effect takes the average outcome in the treatment group and compares it to the average outcome in the control group. With both potential outcomes unobservable simultaneously for all observations, comparing these two averages is recognized as providing an unbiased and useful estimate of the true causal effect of a treatment variable, if treatment is assigned at random and effects are relatively homogenous across units. Although the situation-specific methodological challenges drive which groups are compared and how those groups are distinguished, the core calculations always take the difference of means across groups with and without a causal variable.

Keeping this comparison of average outcomes in mind, we can see Equation 7 in the more complicated Equations 1 and 5. Restating in turn, Equation 1 is the primary formulation for

spillover effects:

$$\tau(d_k, d_l) = \frac{1}{N}\sum_{i=1}^{N} y_i(d_k) - \frac{1}{N}\sum_{i=1}^{N} y_i(d_l)$$

and Equation 5 denotes calculations for heterogenous treatment effects:

$$\hat{\tau}(x) = \frac{1}{|\{i : W_i = 1, X_i \in L\}|} \sum_{\{i:W_i=1, X_i \in L\}}^{Y_i} - \frac{1}{|\{i : W_i = 0, X_i \in L\}|} \sum_{\{i:W_i=0, X_i \in L\}}^{Y_i} .$$

For both of these equations, the effect of interest is comparing the average outcome across groups that do and do not receive some causal assignment. In the spillover equation, d_k, is the group with indirect treatment exposure and d_l receives no indirect exposure. For the heterogeneous treatment effect equation, $W_i = 1$ denotes receiving treatment and $W_i = 0$ denotes being in the control group. Both then are averaging outcomes, y_i, across the two groups.

Considering how the two equations can be combined, let both d_k and d_l be units in the control group, varying by whether they share a network tie with treated units, conferring indirect treatment exposure. Then the average causal effect of spillover exposure is the difference between the average outcome in both groups k and l. Incorporating Equation 3, we can then restate this difference in means to be the difference in weighted averages, taking each unit's probability of indirect treatment exposure into account. Moreover, in equation 5, if $X_i \in L$ is restated as all covariate values and W_i denotes spillover exposure, rather then treatment exposure, ***then the two equations will return identical outcomes***. This is the core argument for this paper: *because the two methods start with the same methodological assumptions and framework, the two can be used in tandem.*

Next, I outline a series of steps with R code, demonstrating how to combine the methods in practice. In order to do so, the following data is necessary: a network of ties between units, treatment assignment, and observation-level covariates to condition on. Fortunately, recent software developments for modeling heterogeneous treatment effects allow for a straightfor-

ward implementation of sample weights, allowing the use of the Horvitz-Thompson inverse probability weights.

3.1 R Procedure

Before turning to simulations and an application, this section outlines the necessary R code for combining the interference and grf packages. Both packages are recent improvements in the estimation of spillover effects and heterogeneous treatment effects, respectively. After installing the libraries, this code assumes the following objects in R:

- `net`: network object
- `treat`: vector with each node's treatment assignment
- `X`: covariate values for each node
- `y`: outcome variable for each node
- `W`: binary variable denoting whether or not a node receives indirect treatment exposure

Once these objects are defined, the following code can be easily applied to one's data with minimal adjustments.

First, we extract the adjacency matrix from our network of interest:

```
# Extract adjacency matrix from network object
adjmat <- treat_schools_net(net)
```

Then we calculate the probability that each node ends up in each of the four exposure conditions outline in Aronow et al..

```
# Store treatment assignment for each node
d <- make_exposure_map_AS(adjmat, treat, hop = 1)
# hop = 1 for single tie conferring indirect exposure

# Create vectors of different treatment assignments
potential_tr_vector <- make_tr_vec_permutation(
    N = nrow(d), # number of observations
    p = p_treat, # exogenously set probability of treatment
    R = 30, # number of treatment vectors to generate
    seed = 123 # random seed for replication
```

```
)

# Calculate exposure probabilities
obs_prob_exposure <- make_exposure_prob(
    potential_tr_vector, # treatment assignments
    adjmat, # network adjacency matrix
    make_exposure_map_AS, # exposure mappings
    list(hop = 1) # spillover with 1 tie
)

# Exposure conditions
exp_probs <- t(make_prob_exposure_cond(obs_prob_exposure))
```

Now that we have the probability of exposure for each condition for each node, we can apply those probabilities as weights in the causal forest. For this causal forest, we will estimate $\tau_i(d_{01}, d_{00})$, or the expected difference between indirect exposure and no exposure for each unit, conditional on its covariate values. This code fits and evaluates the causal forest on all available data, but can be easily amended for the training-test framework if one wishes to approximate the process of using a pilot study to make predictions about expected effects on new data. Here, let `y` be the outcome of interest, `X` is a matrix of covariates, and `W` is the treatment (here a spillover). Notably, the causal forest includes the option to also match by inverse-propensity scores, estimating within-leaf treatment effects by matching units together based on their estimated probability of receiving treatment. In light of King and Nielsen (2019), I eschew this option and only use the inverse-probability estimates as sample weights, rather than matching on them.[11]

```
cf <- causal_forest(
    X = as.matrix(X), # Covariates
    Y = y$outcome, # Outcome
    W = W$spillover, # Spillover assignment
    sample.weights = 1/exp_probs$ind, # Probability weight
    seed = 123 # random seed for replication
)
```

We can then test for effect heterogeneity in the entire causal forest with the following

[11] For more information on implementing a causal forest, https://www.markhw.com/blog/causalforestintro is a helpful resource.

command:

```
test_calibration(cf)
```

If the term `differential.forest.prediction` in the resulting summary table is positive and significant, then that implies support for effect heterogeneity and we can reject the null of homogenous effects. If the null of homogenous effects has been rejected, then we can use a variable importance plot to see which variables tended to produce the most effect heterogeneity when used to generate splits in the causal forest.[12]

```
cf %>%
     variable_importance() %>%
     as.data.frame() %>%
     mutate(variable = colnames(cf$X.orig)) %>%
     arrange(desc(V1))
```

If the estimated spillover effects are heterogeneous and certain variables are highlighted by the variable importance plot, then we can turn to using observation-level predicted effects to visualize the heterogeneity of interest. If no new test set is entered for prediction, then the default prediction format is to use 'out-of-bag' (OOB) predictions, where each observation's predicted outcome is made using trees where it was not used for training. Because random forests build each tree using a random sample of data, the expectation is that each data point is not used for building every tree, meaning there are certain trees where the data point approximates a new test set observation that the model is unfamiliar with. Notably, a benefit of the honest approach is that the model produces observation-level standard errors, which `estimate.variance = TRUE` allows.[13]

```
# If no test set, using OOB predictions
predict(cf, estimate.variance = TRUE)
```

[12]This code is drawn from https://www.markhw.com/blog/causalforestintro, where additional useful code is available for using a causal forest in a social sciences context.

[13]If all observations are plotted together at once, then the standard errors can produce a messy visual which is difficult to interpret. But if one is interested in a single or handful of cases, then the standard errors about predicted observation-level causal effects, conditional on covariates, are incredibly helpful for probabilistically comparing effect estimates.

```
# If there is a test set
predict(cf, as.matrix(test_x), estimate.variance = TRUE)
```

The resulting predictions of observation-level spillover effects can then be visualized and interpreted as necessary for one's purposes. In the subsequent example of spillover effects in anti-bullying programs, I aggregate predicted effects across schools for a set of four models, each predicting a different outcome.

4 Simulation

In observational studies, heterogeneous treatment effects are estimated based on the observed outcomes for each unit and their covariate values, through which moderating effects can occur. In this section I simulate data where spillover effects occur through indirect treatment exposure and all potential outcomes are generated for each observation. Each observation's potential outcomes ($d_{11}, d_{10}, d_{01}, d_{00}$) are simulated to be a function of: whether or not it receives direct treatment, it is tied to a treated unit, and its covariate values. Spillovers are then programmed in to explicitly occur heterogeneously, conditional on covariate values. The two variables that I vary throughout the simulation are 1) the number of observations and 2) the heterogeneity of the true spillover effects. The simulations therefore tests the ability of the combination of causal forests and estimated spillover exposure weights to correctly predict a unit's spillover effect, keeping the number of observations and effect heterogeneity in mind.

Since the causal forest's final product is a set of observation-level predictions of causal effects, it follows to simulate observations where potential outcomes are known and to evaluate how closely the causal forest predicts the relevant differences between the potential outcomes of interest. Alongside evaluating whether or not the proposed procedure returns accurate predictions of the true causal effect, I compare the final predictions when the inverse probability weights are included and not included, testing whether weighting observations by their probability of receiving indirect treatment exposure improves the causal forest's accuracy.

This simulation's goal is to demonstrate that the causal forest can be used to model hetero-

geneous spillover effects and that the causal forest effectively implements the inverse probability weights. While forest-based methods are confirmed to hold promise for heterogeneous treatment effects and the inverse probability weights are recognized to assist in spillover studies, I hope to convince the reader that the combination of the approaches works as expected. Across simulations, the causal forest returns accurate unit-level predictions of spillover effects. Moreover, while implementing inverse probability weights almost never produces less accurate predictions, it generally improves or does not change predictive accuracy. In game-theoretic terms, including the probability weights is a weakly-dominant strategy; including the weights almost never leaves the researcher worse off, but generally helps.

4.1 Simulation Setup

Each simulation iteration is an experiment with sample size n, a network of ties between units, and treatment assignment z. Our effect of interest is the spillover effect, where treatment is randomly assigned and then units with a network tie to treated units receive indirect exposure. Our two exposure conditions of interest, then, are indirect exposure only (d_{01}) and no exposure (d_{00}). This provides the following function for an average spillover effect:

$$\tau = \mathbb{E}[y(d_{01}) - y(d_{00})] \tag{8}$$

which is explicitly programmed to vary across covariate values, creating the conditional average spillover effect:

$$\tau(x) = \mathbb{E}[y(d_{01}) - y(d_{00}) \mid X = x]. \tag{9}$$

The benefit of simulation is that we are able to know every observation's entire set of potential outcomes. While the estimator faces the fundamental problem of causal inference, only encountering one potential outcome for each observation, we are able to verify the causal forest's predicted observation-level causal effects through the known potential outcomes. Each observation is assigned 5 covariate values, X, drawn from a standard normal distribution, and a set of 5 spillover effects for those covariates are drawn from a normal distribution with mean

0 and varying standard deviation (one of our varying parameters). To put this in familiar terms, let the spillover coefficients be labelled β. The true direct treatment effect is held arbitrarily at 2. Each unit then receives some random noise that cannot be modelled, ε_i. Each unit's potential outcomes are simulated as follows:

- Direct and indirect exposure (d_{11}): $2 + X_i * \beta + \varepsilon_i$
- Direct exposure only (d_{10}): $2 + \varepsilon_i$
- Indirect exposure only (d_{01}): $X_i * \beta + \varepsilon_i$
- No exposure (d_{00}): ε_i

For each unit, every potential outcome is calculated so that the true effects can be compared against the causal forest's predictions. The simulated experiment's treatment assignment and network ties are also held constant within each simulation iteration. Once the probabilities for each condition are estimated, $\pi_i = (\pi_i(d_{11}), \pi_i(d_{10}), \pi_i(d_{01}), \pi_i(d_{00}))$, a causal forest is fit to compare observations with indirect exposure only to observations with no exposure, conditional on their covariate values. The resulting model provides predicted spillover effects for each observation, which are estimated by comparing the average outcome for observations with similar covariate values (fall in the same random forest leaf) that receive indirect exposure and those that do not. We can then see how well the causal forest's difference-based predictions map onto the true observation-level effects.

Lastly, across simulations I vary two parameters: the number of observations and the heterogeneity of effects. For each simulation I also fit a causal forest without the estimated inverse probability of treatment weights and then a causal forest with the weights, allowing a comparison of whether the Horovitz-Thompson weights improve the model.

4.2 Simulation Results

Figures 2 and 3 include simulation results. Across simulations the causal forest without weights and with weights both tend to correctly estimate the true observation-level spillover effect. This

is reflected in Figure 2, where the x-axis is a model's root-mean squared error for a single iteration, which is calculated by comparing the observation-level predicted spillover effects to the true effects. In Figure 2, both the weighted and unweighted models disproportionately tend to accurately predict the true causal effect, with the histograms heavily peaked at zero. However, foreshadowing Figure 3, the peak at zero is higher for the weighted model. Therefore, we see that the causal forest is generally effective at estimating heterogeneous spillover effects, but tends to benefit from the inverse probability weights.

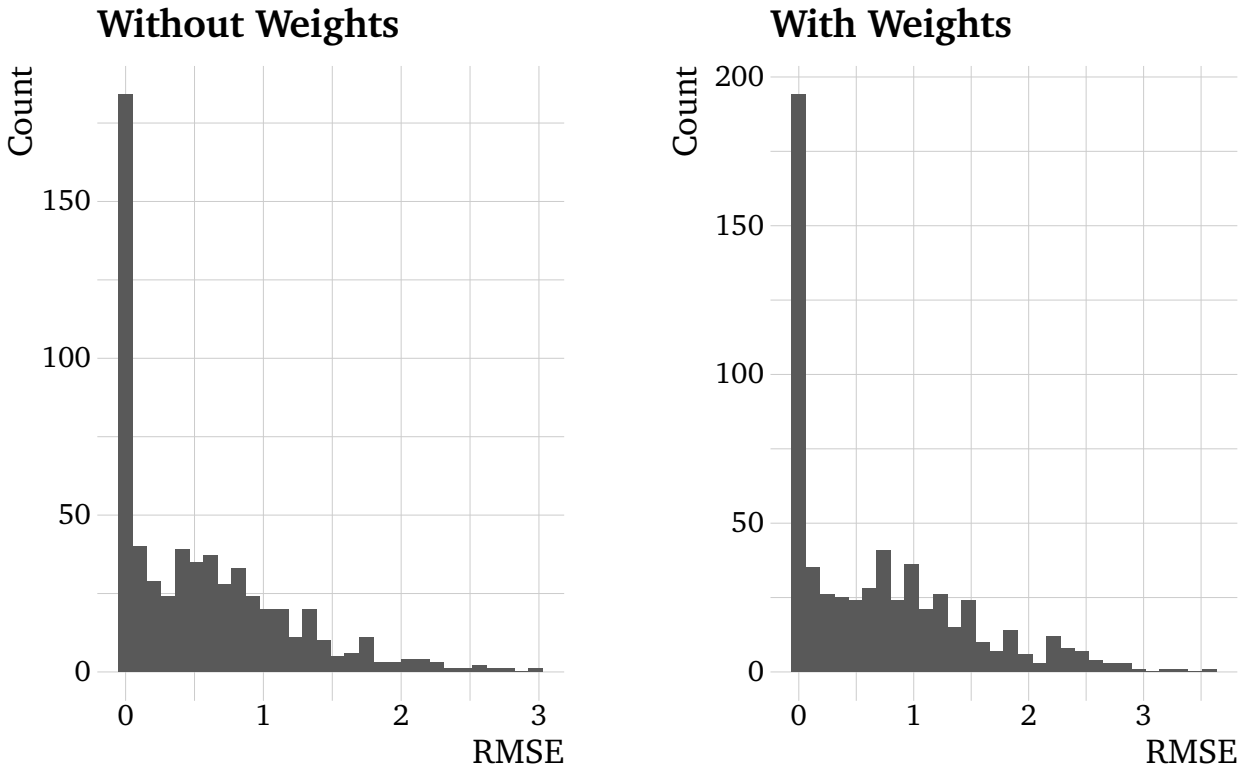

Figure 2: Histogram of the root mean squared error

Decomposing the added benefit of the probability weights, Figure 3 compares the RMSE for the weighted and unweighted models in each simulation. Positive y-axis values represent a decrease in RMSE when using weights, or an increase in predictive accuracy. X-axis values represent the standard deviation in spillover effects – where a larger standard deviation represents more heterogeneous effects. The color of each boxplot is the number of observations in the network. We see that as the number of observations increases, so too does the value of the

inverse probability weights. This is likely due to the fact that the more observations there are, then the more distinct the probability of being in each exposure condition becomes, increasing the impact of the weights. In addition, the more heterogeneous the underlying effects are, then the more these weights tend to increase model accuracy for larger networks.

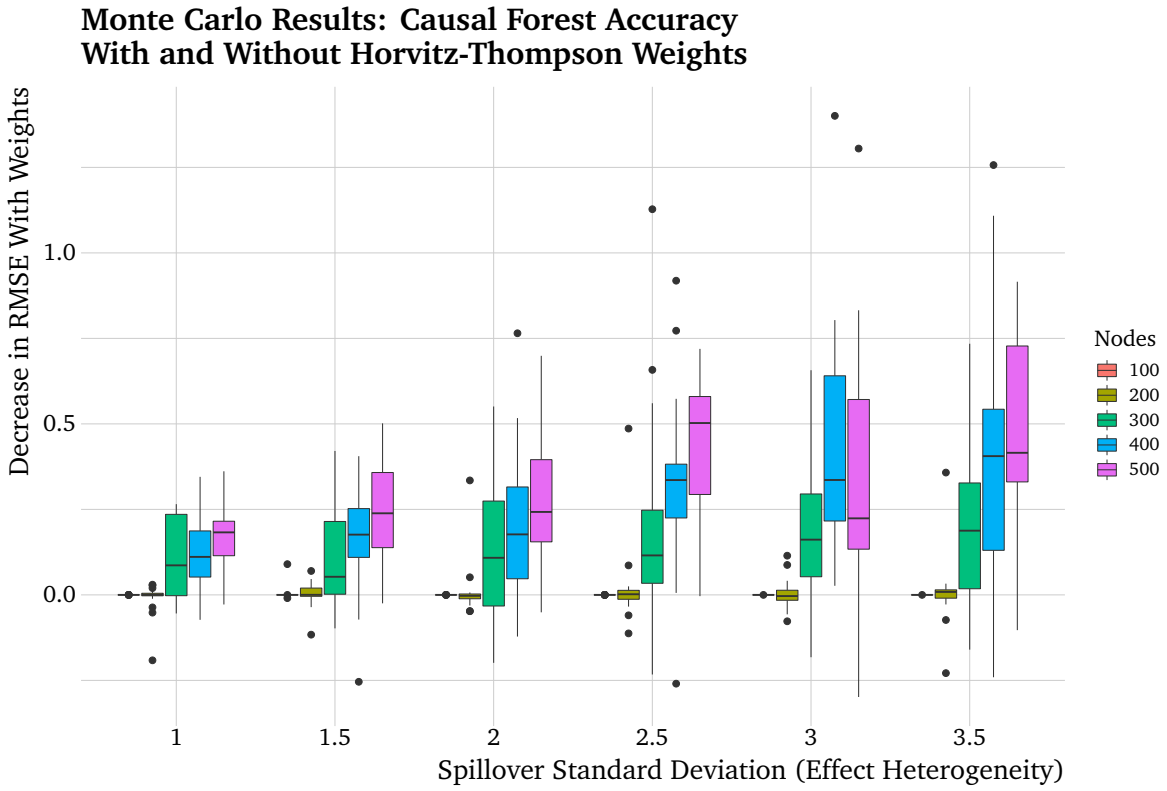

Figure 3: RMSE for a weighted causal forest minus the RMSE for an unweighted causal forest. The heterogeneity of the true spillover effects increases along the x-axis and the number of observations per simulation is color-coded.

Reiterating the aforementioned point about including weights being akin to a weakly-dominant strategy in game theory, only in a very small handful of simulations does including weights decrease the causal forest's accuracy. Moreover, if the network includes fewer than 200 nodes, then the weights tend to make no difference. However, as effects become more heterogeneous and the network's size grows, then the probability weights tend to increasingly improve model accuracy, relative to an unweighted model.

5 Application: Anti-Bullying Programs in School

Turning to a demonstration of the method in an applied setting, Paluck et al. (2016) provide an example of quantifying spillover effects in a randomized controlled trial. In their study, Paluck et al. measure social networks within schools in New Jersey and then assign prominent nodes – popular students – at random to an anti-bullying program. The study demonstrates the effectiveness of anti-bullying programs in schools, not just comparing subsequent aggregate behavior at the school level, but also comparing outcomes of students that are friends with students in the anti-bullying program to students who are not friends with any attendees – spillover effects. On average, Paluck et al. find that the anti-bullying program, though implemented with a small number of students, tends to decrease bullying at the school-level and behaviors differ meaningfully between those with indirect exposure through friends and those with no exposure.

The study is normatively important and methodologically thorough. It also includes all necessary pieces for examining the degree to which the anti-bullying program's spillover effects also vary across groups and are heterogeneous. In this section I extend the paper's results, estimating each student's probability weights for each exposure condition and then fitting a causal forest for each of the four dependent variables included in the study. The four dependent variables are: 1) a perceived norm against bullying within the school, 2) frequency of discussions about bullying among friends, 3) wearing a wristband to signify support for the anti-bullying program, and 4) subsequent conflictual behavior between peers at the school (based on administrative records). Evidence of an effective anti-bullying program is a positive association with 1-3 and a negative association with 4. Covariates included in the causal forest are: the school's ID, the student's grade level, ethnicity, gender, proxies for income, and the student's recent disciplinary record.

I fit a causal forest with data on students in treated schools, comparing students in the control group (not assigned to the anti-bullying program) who are friends with students in the program to control group students not friends with students in the program. Another form

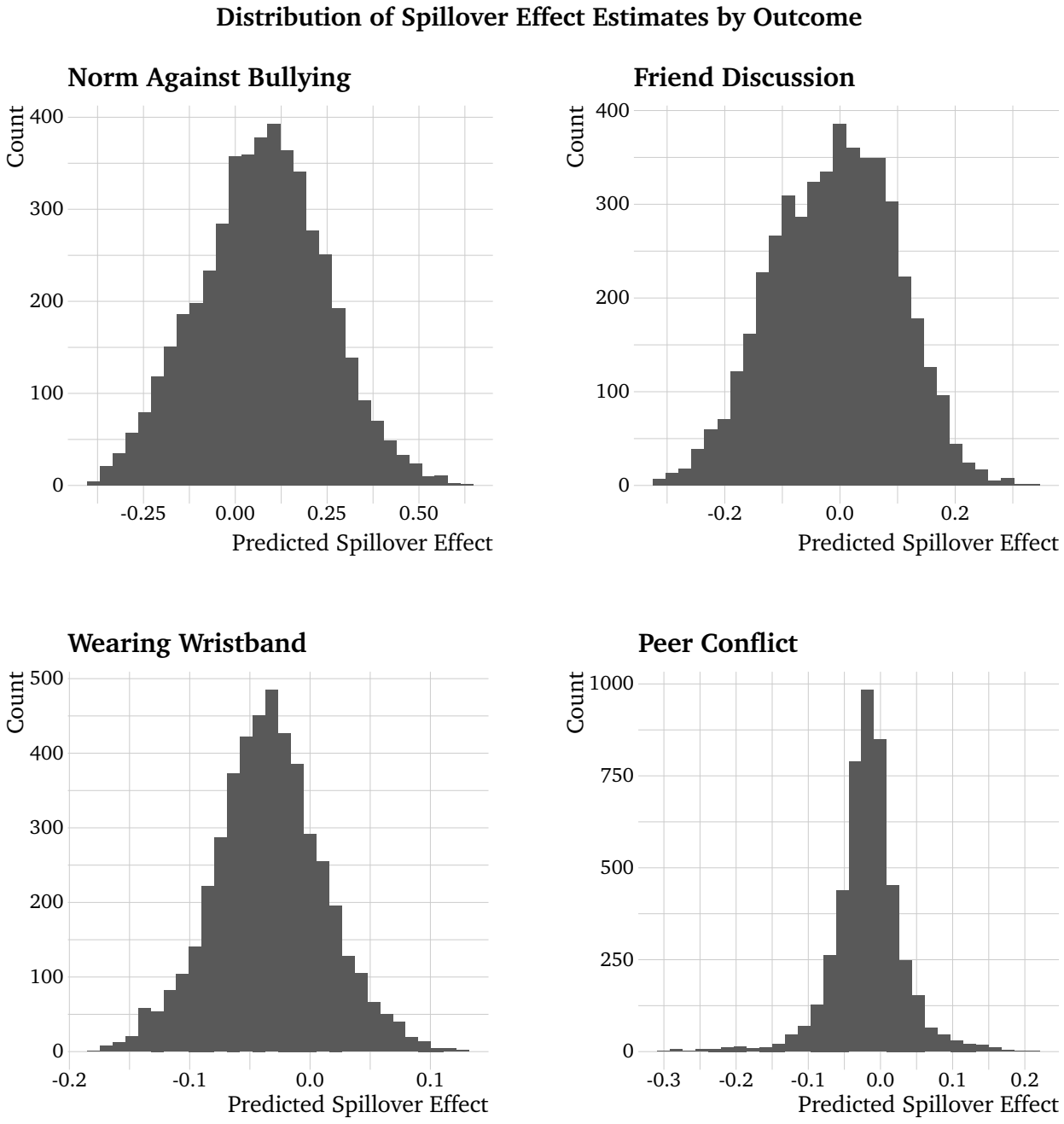

Figure 4: Histogram of predicted spillover effects for each dependent variable.

of spillover effects, which I briefly examine here, is to compare treated students who are and are not friends with other treated students. Figure 4 includes histograms with the distribution of predicted spillover effects for each student, conditional on their covariate values. For these dependent variables, wearing a wristband is a binary outcome, whereas the others are 3 or

4 point scales. This means the spillover effects are of a relatively small magnitude, but that is reasonable given the size of the program and study. However, we do see heterogeneity in the direction of the spillover effect, which is an important distinction. Across all outcomes, a substantial portion of students fall under both positive and negative spillover effects. This suggests that if the program were reimplemented at a larger scale, we should expect that, in terms of spillover effects, even if the program is more effective than not, in some schools the program has a real possibility of backfiring.

Turning to how spillover effects vary across schools, I cluster students by schools and compare the within-school distributions of predicted spillover effects. As an aside, these results are not identical to running a regression with interaction effects for each school. Rather, Figure 5 takes the causal forest's predictions, which are a function of each observation's entire set of covariate values, and then groups the predictions by schools. So while the output may appear as if only outcomes and school ID's are considered in model fitting, the results actually reflect the entirety of the covariates for each observation. Looking at the within-school predictions for each outcome, we see where the effects vary in magnitude and direction. Without more information about the schools it is difficult to say why we see the variation we do across these schools, but we can see that some schools are far more receptive than others to the anti-bullying program, with a handful of schools having the opposite effect. The heterogeneity of these results demonstrates how taking evidence of generally effective spillovers can be dangerous and mislead to applying a treatment to contexts where it will likely have the opposite effect, compared to the intended effect.

Curiously, while three out of the four dependent variables demonstrate spillovers associated with the desired outcome (more of a norm against bullying, more discussions about bullying among friends, and less subsequent conflicts), we see the opposite with wearing wristbands. Instead, the causal forest predicts that the spillover effects are generally negative – causing less wristband wearing than would otherwise be the case. This clashes with the expectation from the original paper, where the anti-bullying program is associated with wearing more

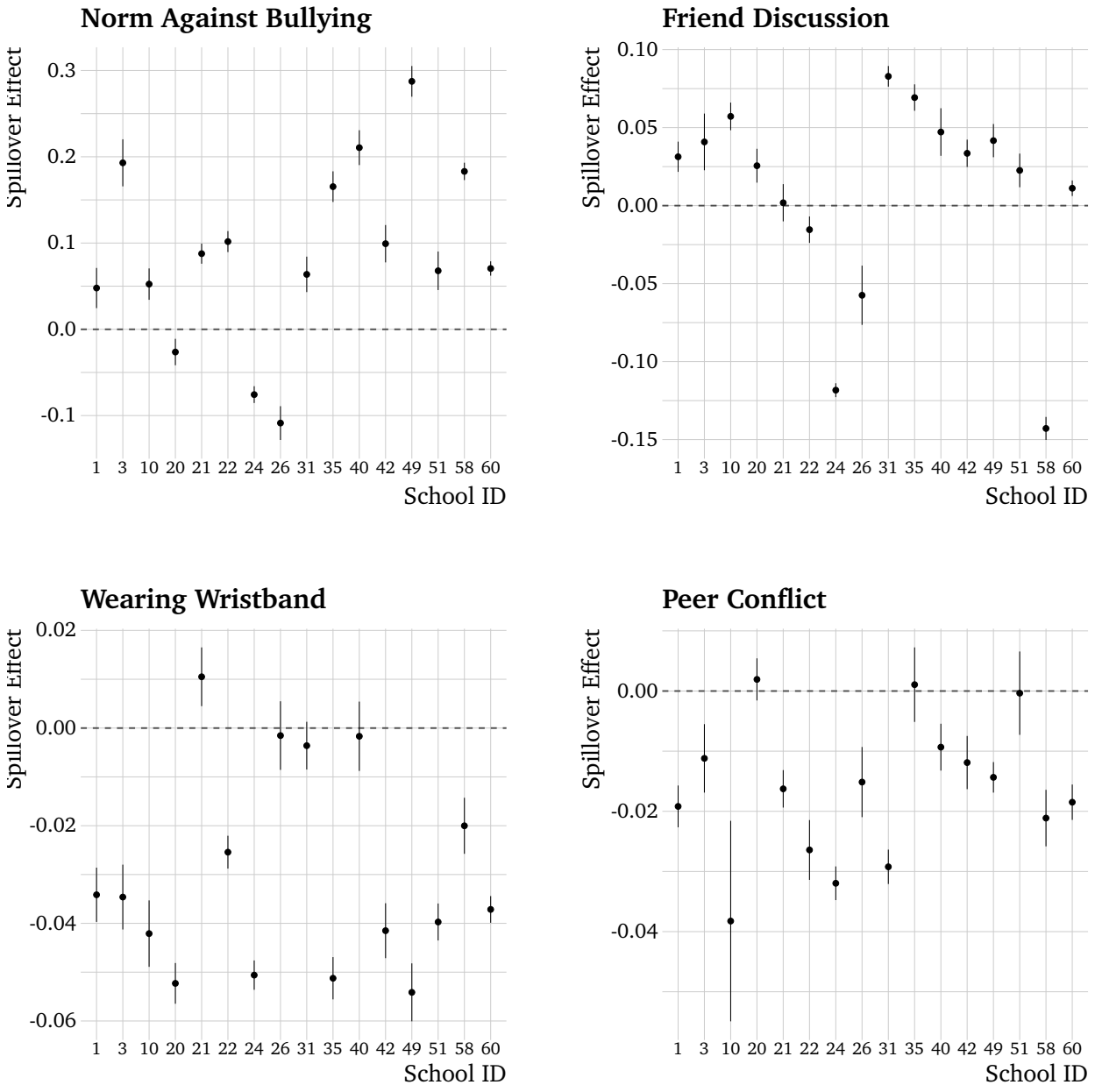

Figure 5: Spillover effects across dependent variables clustered by schools. Effects are calculated at the individual level and then grouped and summarized for each school.

wristbands across the schools. In response, in Figure 6, I compare treated students with a tie to other treated students to treated students with no indirect exposure – $\tau(d_{11}, d_{10})$ – and see consistently large positive spillover effects. This suggests that the results around wristband wearing are driven largely by the popular students selected into the anti-bullying program and not the other students. This spillover effect is generally larger and in the expected direction,

suggesting that not all spillovers work the same way, another interesting topic for discussion.

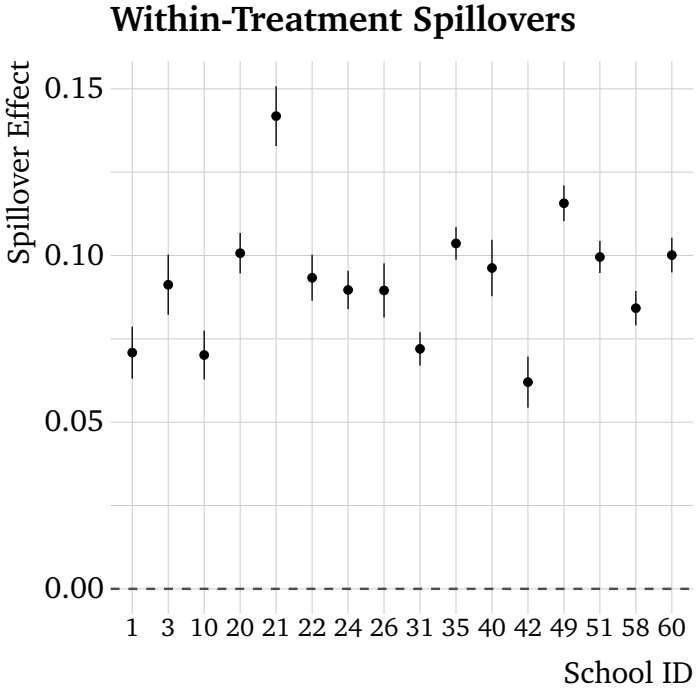

Figure 6: Spillover effects for wearing the anti-conflict wristband when observations in the treatment group that share a tie other treated units are compared to observations in the treatment group without a tie to treated units.

6 Conclusion

This article develops a procedure for estimating heterogeneous spillover effects. The proposed methodology combines techniques for estimating spillover effects with causal random forests. Each observation's potential outcomes are decomposed to not only differentiate by treatment status, but also whether an observation shares a network tie with a treated unit. Then weights are estimated for the probability that each observation ends up in each of the decomposed potential outcomes. Lastly, the weights are included in a causal random forest that is built to flexibly model heterogeneous treatment effects, but instead of comparing observations by their treatment status, observations are compared based on whether they share a network tie

with treated units. After verifying the method's efficacy in simulations, I apply it to a study of spillover effects in anti-bullying programs at schools. I find that certain schools do tend to respond favorably, whereas others do not. This example shows how a pilot study, like the anti-bullying program, can inform where subsequent efforts can be targeted to maximize their effectiveness.

References

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

P. M. Aronow, D. Eckles, C. Samii, and S. Zonszein. Spillover effects in experimental data. *arXiv preprint arXiv:2001.05444*, 2020.

S. Athey and G. Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

K. Baicker. The spillover effects of state spending. *Journal of public economics*, 89(2-3):529–544, 2005.

T. Brambor, W. R. Clark, and M. Golder. Understanding interaction models: Improving empirical analyses. *Political analysis*, 14(1):63–82, 2006.

B. F. Braumoeller. Hypothesis testing and multiplicative interaction terms. *International organization*, 58(4):807–820, 2004.

K.-y. Cheung and L. Ping. Spillover effects of fdi on innovation in china: Evidence from the provincial data. *China economic review*, 15(1):25–44, 2004.

S. J. Cranmer, P. Leifeld, S. D. McClurg, and M. Rolfe. Navigating the range of statistical tools for inferential network analysis. *American Journal of Political Science*, 61(1):237–251, 2017.

J. Esarey and J. L. Sumner. Marginal effects in interaction models: Determining and controlling the false positive rate. *Comparative Political Studies*, 51(9):1144–1176, 2018.

D. P. Green and H. L. Kern. Modeling heterogeneous treatment effects in survey experiments with bayesian additive regression trees. *Public opinion quarterly*, 76(3):491–511, 2012.

J. Grimmer, S. Messing, and S. J. Westwood. Estimating heterogeneous treatment effects and the effects of heterogeneous treatments with ensemble methods. *Political Analysis*, 25(4): 413–434, 2017.

J. Haushofer and J. Shapiro. The long-term impact of unconditional cash transfers: experimental evidence from kenya. *Busara Center for Behavioral Economics, Nairobi, Kenya*, 2018.

J. Hill and Y.-S. Su. Assessing lack of common support in causal inference using bayesian nonparametrics: Implications for evaluating the effect of breastfeeding on children's cognitive outcomes. *The Annals of Applied Statistics*, pages 1386–1420, 2013.

J. Hill, A. Linero, and J. Murray. Bayesian additive regression trees: A review and look forward. *Annual Review of Statistics and Its Application*, 7, 2020.

J. L. Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

M. G. Hudgens and M. E. Halloran. Toward causal inference with interference. *Journal of the American Statistical Association*, 103(482):832–842, 2008.

D. R. Hunter, M. S. Handcock, C. T. Butts, S. M. Goodreau, and M. Morris. ergm: A package to fit, simulate and diagnose exponential-family models for networks. *Journal of statistical software*, 24(3):nihpa54860, 2008.

K. Imai and M. Ratkovic. Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7(1):443–470, 2013.

L. R. James and J. M. Brett. Mediators, moderators, and tests for mediation. *Journal of applied psychology*, 69(2):307, 1984.

J. J. Jones, R. M. Bond, E. Bakshy, D. Eckles, and J. H. Fowler. Social influence and political mobilization: Further evidence from a randomized experiment in the 2012 us presidential election. *PloS one*, 12(4), 2017.

G. King and R. Nielsen. Why propensity scores should not be used for matching. *Political Analysis*, 27(4):435–454, 2019.

S. Minhas, P. D. Hoff, and M. D. Ward. Inferential approaches for network analysis: Amen for latent factor models. *Political Analysis*, 27(2):208–222, 2019.

A. Ng. Volatility spillover effects from japan and the us to the pacific–basin. *Journal of international money and finance*, 19(2):207–233, 2000.

D. W. Nickerson. Is voting contagious? evidence from two field experiments. *American political Science review*, 102(1):49–57, 2008.

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

D. B. Rubin. Formal models of statistical inference for causal effects. *Journal of statistical planning and inference*, 25(3):279–292, 1990.

S. J. Taylor and D. Eckles. Randomized experiments to detect and estimate social influence in networks. In *Complex Spreading Phenomena in Social Systems*, pages 289–322. Springer, 2018.

J. Ugander, B. Karrer, L. Backstrom, and J. Kleinberg. Graph cluster randomization: Network exposure to multiple universes. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 329–337, 2013.

T. J. VanderWeele and E. J. T. Tchetgen. Effect partitioning under interference in two-stage randomized vaccine trials. *Statistics & probability letters*, 81(7):861–869, 2011.

J. N. Victor, A. H. Montgomery, and M. Lubell. *The Oxford Handbook of Political Networks*. Oxford University Press, 2017.

S. Wager and S. Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

J. D. Wilson, M. J. Denny, S. Bhamidi, S. J. Cranmer, and B. A. Desmarais. Stochastic weighted graphs: Flexible model specification and simulation. *Social Networks*, 49:37–47, 2017.