Lecture 4 - Interference, Spillovers and Dynamics

Paul Goldsmith-Pinkham

January 21, 2026

This lecture note will discuss what it means to relax the following assumptions from the previous lecture:

1. Binary scalar treatment
2. Single time period (e.g. one treatment within the person)
3. SUTVA – Stable Unit Treatment Value Assignment

Multivalued treatments

So far, our discussion of treatment effects has focused on single binary treatments. This made life very easy, but we have a lot of other more complex settings. We'll consider a few different cases. First, a multi-valued treatment. Then, we'll consider a continuous treatment. Finally, we'll consider an unordered multi-valued treatment.

Discrete multi-valued treatment

Let's start with a discrete, multi-valued treatment to start: $D_i \in \mathcal{D} = \{0, 1, \ldots, d\}$. This captures a simple setting like "what is the impact of 0, 1, 2, or 3 children on labor force participation?" In this setting, we can easily shift the scale up and down ("what is the impact of 5, 10, 15, or 20 minutes on a task") but the order and spacing may matter, depending on how we choose to parameterize and estimate the treatment effect.

In this setting, how should we consider the treatment effects? First, what should we consider the "control"? We could consider the control to be the lowest value of the treatment, but that depends a bit on the context. For example, if we are considering the impact of 0, 1, 2, or 3 children on labor force participation, we might consider the control to be 0 children. However, if we are considering the impact of 5, 10, 15, or 20 minutes on a task, we might consider the control to be whatever the status quo was.

Formally, we define the potential outcome for any $d \in \mathcal{D}$ as $Y_i(d)$, and we consider the individual and average treatment effect difference between d and d' as:

$$\begin{aligned} \tau_i(d, d') &= Y_i(d) - Y_i(d') \\ E(\tau_i(d, d')) &= E(Y_i(d) - Y_i(d')). \end{aligned}$$

If strong ignorability holds, then this is also identified by simply conditioning on each observed value:[1]

$$E(\tau_i(d, d')) = E(Y_i|D_i = d) - E(Y_i|D_i = d').$$

This type of estimation is non-parametric in nature: we've assumed no functional form between the treatment and the potential outcome. A consequence of that, just like in the case with many covariates $\mathbf{X}$, is that it requires a lot more data to provide precise estimates. If we wanted to consider the CATE:

$$E(\tau_i(d, d')|\mathbf{X} = x) = E(Y_i|D_i = d, \mathbf{X} = x) - E(Y_i|D_i = d', \mathbf{X} = x),$$

then we'll need to condition on treatment categories within each cell, which can be very data hungry, and less precise.

Often, instead of estimating the effect for every point separately, we will postulate a model for the potential outcomes:

$$Y_i(d) = Y_i(0) + \tau_i d.$$

Notice that in this case, this implies that for all d, d' pairs

$$\tau_i(d, d') = \tau_i,$$

which is the slope parameter. Hence, estimation can be made more precise by pooling all of these estimands together into a single estimand $E(\tau_i)$.[2]

Example 1

Consider the following simulated data, where the true effect is linear and simulated such that $E(\tau_i(d, d') = d' - d$ and strong ignorability holds.

Each dot in Figure 1 is the estimated mean at the point, and we find a positive treatment effect. Imposing the model helps a lot *compared to non-parametric form. To see this, consider the treatment effect comparing d to $d - 1$ in Figure 2*

The direction of the effect is much more ambiguous. This is a common trade-off in estimation: imposing a model can help with precision, but can also lead to bias if the model is misspecified.

How should we consider these functional forms once we include controls? To see what I mean, consider the same context, but we now assume strong ignorability conditional on $\mathbf{X}$. Then, we would need to estimate the slope τ_i for each value of $\mathbf{X}$. How would that map over to a linear regression model? The simplest version would be one where the heterogeneity, τ_i, is uncorrelated with $\mathbf{X}$. Then, one

[1] The overlap condition in strong ignorability with multiple treatments is more complicated, but effectively entails that for any $\mathbf{X}$, there are observations for every d in $\mathcal{D}$.

[2] Other, more flexible parametric forms for $Y_i(d)$ could be chosen as well. The insights will carry through so long as the functional form is finite-dimensional.

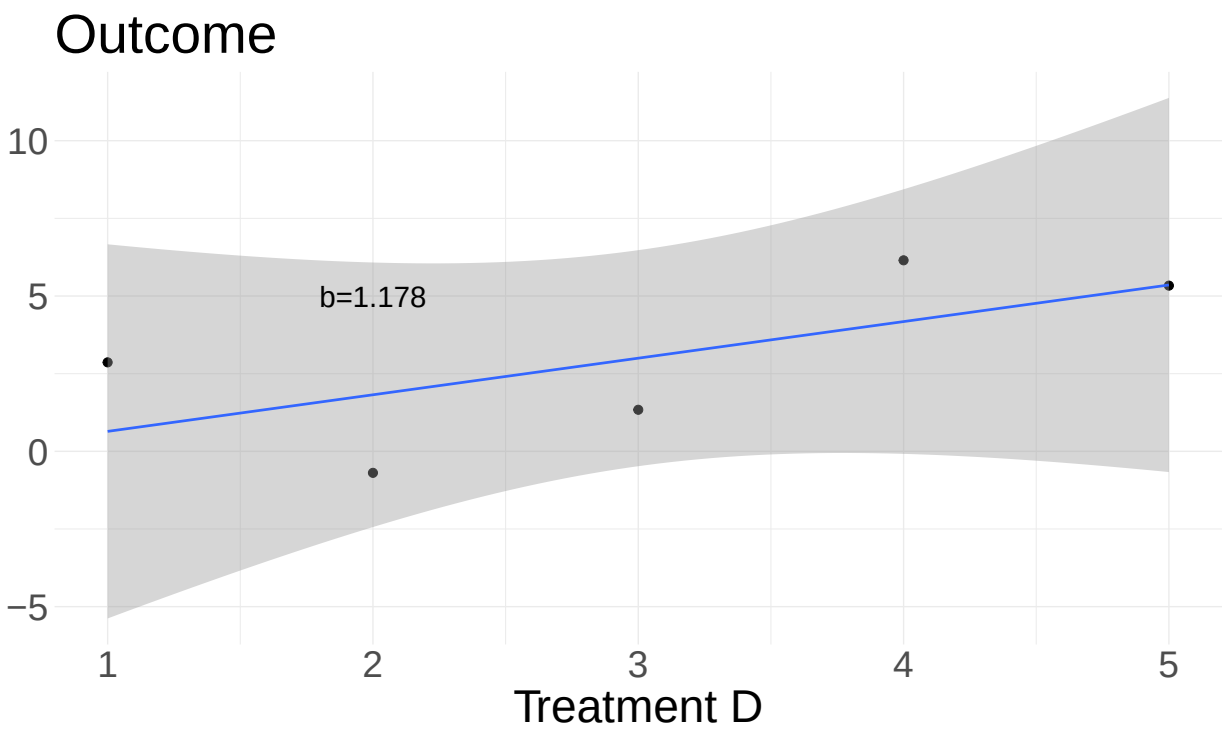

Figure 1: Linear effects estimated in simulated data with a true linear model for Example 1

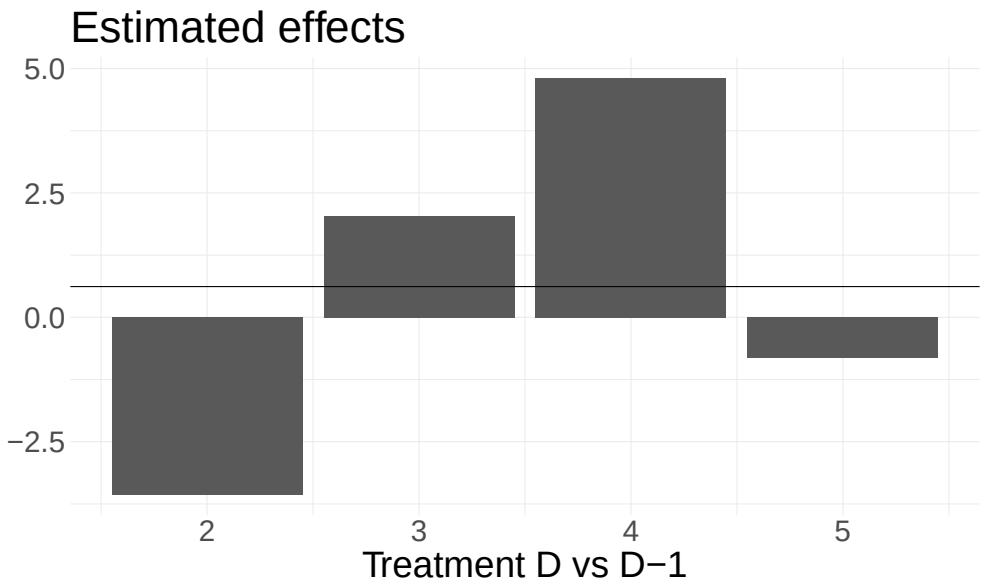

Figure 2: Non-parametric effects estimated in simulated data with a true linear model for Example 1

way to estimate the ATE is to assume that $E(D_i|\mathbf{X}) = \mathbf{X}_i\beta$ (e.g. the propensity score is linear in $\mathbf{X}$), we could estimate $E(\tau_i)$ by simply running the following regression:

$$Y_i = \alpha + D_i\tau + \mathbf{X}_i\beta + \epsilon_i \tag{1}$$

and using τ as our estimate of $E(\tau_i)$. But, if τ_i is not uncorrelated with $\mathbf{X}$, then to estimate the ATE we would need to estimate the slope for each value of $\mathbf{X}$, and pool separately. See Comment 1 for more details.

Comment 1

It is worth thinking about why Equation (1) will correctly estimate the ATE in this setting. To do this, let $E^*(D_i|\mathbf{X}_i)$ denote the best linear predictor of D_i conditional on $\mathbf{X}_i$. Now, note that by Frisch-Waugh-Lovell,

$$\tilde{Y}_i = \tilde{\alpha} + \tilde{D}_i\tau + u_i, \tag{2}$$

where $\tilde{Y}_i = Y_i - E^*(Y_i|\mathbf{X}_i)$ and $\tilde{D}_i = D_i - E^*(D_i|\mathbf{X}_i)$. Then,

$$\begin{aligned}\tau &= \frac{E(\tilde{D}_i Y_i)}{E(\tilde{D}_i^2)} \\ &= \frac{E(\tilde{D}_i Y_i(0))}{E(\tilde{D}_i^2)} + \frac{E(\tilde{D}_i D_i \tau)}{E(\tilde{D}_i^2)}\end{aligned}$$

The first term is zero because the residual $\tilde{D}_i$ is mean independent of $\mathbf{X}$ by the linearity of $E(D_i|\mathbf{X})$. Therefore,

$$E(\tilde{D}_i Y_i(0)) = E(E(\tilde{D}_i Y_i(0)|\mathbf{X})) = E(E(\tilde{D}_i|\mathbf{X})E(Y_i(0)|\mathbf{X})) = 0.$$

Now note the second term by similar arugments:

$$\tau = \frac{E(\tilde{D}_i D_i \tau)}{E(\tilde{D}_i^2)} = \frac{E(Var(D_i|\mathbf{X})E(\tau|\mathbf{X}))}{E(Var(D_i|\mathbf{X}))}.$$

But, since we assumed $E(\tau|\mathbf{X}) = E(\tau)$, this is just the ATE. If there is correlation of τ with $\mathbf{X}$, we will get a different estimand.

Discussion Questions 1

Under what assumptions about $E(Y_i(0)|\mathbf{X}_i)$, instead of $E(D_i|\mathbf{X}_i)$, could we estimate the ATE using Equation (1)?

1. *Hint: think about a linear model for $E(Y_i(0))$ that is linear in $\mathbf{X}_i$.*
2. *Second hint: the estimand could also be written as*

$$\tau = \frac{E(\tilde{D}_i\tilde{Y}_i)}{E(\tilde{D}_i^2)}$$

Continuous valued treatment

In many cases, the jump from discrete ordered treatments to continuous valued treatments is not large. Often, it just has to do with how many repeated observations we have of the same treatment; if each treatment value is unique, we're more likely to treat it as continuous. None of what we discussed above changes, except that direct non-parametric estimation becomes infeasible.

Instead, to do non-parametric estimation we'll need to make other assumptions and use other methods, like kernel regression or local linear regression. We'll discuss these in future classes, but the key point is that we will want to make some amount of smoothness assumptions the effect of the treatment on the potential outcome.

Instead of non-parametric estimation, it is also reasonable to assume a functional form, as above, and proceed from there. Then everything is exactly the same.

Unordered multi-valued treatment

Finally, we might have a setting where the treatment is unordered. For example, we might consider the impact of different CEOs on firms' performance. In this case, we can't assume any particular ordering of the treatment, and it's not clear how to presume a functional form. Instead, there is a set of K treatments in $\mathcal{D}$, and we can consider a set of different contrasts between them: $E(\tau_i(d, d'))$ for all $d, d' \in \mathcal{D}$.

A straightforward special case would be to consider a *factorial* design: a randomized treatment where two treatments are cross-randomized, such that an individual can receive either no treatment, treatment 1, treatment 2, or both. Then, our potential outcomes look like Table 1.

Given this, we have a number of potential estimands to consider. For example, we could consider the average treatment effect of treatment 1, but we would need to make a decision on what to

do about the individuals who received both treatments. If the treatments interact in some way, then the average treatment effect of treatment 1 is not well-defined. Instead, we might consider the average treatment effect of treatment 1 for those who received treatment 2 ($E(Y_i(1,1) - Y_i(0,1))$), and the average treatment effect of treatment 1 for those who did not receive treatment 2 ($E(Y_i(1,0) - Y_i(0,0))$).

$Y_i(\mathbf{D_i})$	$D_{1i} = 0$	$D_{1i} = 1$
$D_{2i} = 0$	$Y_i(0,0)$	$Y_i(1,0)$
$D_{2i} = 1$	$Y_i(0,1)$	$Y_i(1,1)$

Table 1: Potential outcomes for a factorial design

Of course, when data is sparse, we might want to do more with this, and just pool all of the data together: $E(Y_i(1, D_{i2}) - Y_i(0, D_{i2}))$. This is a reasonable estimand if we believe that the treatment effects are constant across the different levels of the other treatment, but if there are interactions, the external validity of this estimate will be suspect.[3] See **?** for a very interesting discussion on how to think about these types of estimands in the context of factorial designs when there are many treatments.

[3] Think about why this is the case, if it's not clear.

Factorial designs are the simplest case to consider, because the treatments are cross-randomized for many binary treatments. Often, we have just an ordered set of treatments. In this case, the same logic applies, but we have to be more careful about how we define the estimands. For example, if we consider the impact of different CEOs on firm performance, how do we define a "control"? It is often not obvious, and we need to be careful about how we define the estimands. We may instead focus on the conditional means for each treatment, and then consider the full distribution of effects. This is the type of consideration in work thinking about place-based effects, for example, such as **?**.[4]

[4] It is interesting to think about these unordered treatments can sometimes be projected into continuous scalar measures. For example, when considering the impact of a CEO on a firm, you might measure a CEO's experience, and project the overall CEO's effect onto experience to capture a continuous measure of the CEO's effect. If you are additionally willing to assume that the effect of experience is the *only* channel controlling the CEO's effect, then a more efficient procedure would use experience as the treatment effect, instead of the CEO. But that may not be a reasonable assumption. This issue arises when considering the effect of judges as in **?**.

Estimating these effects seem like they should be straightforward extensions of the binary treatment case. However, the partial linear regression model fails us in this case. The variation in the propensity score across strata (controls) combined with heterogeneity in the treatment effects across strata will lead to contamination bias in the linear regression model. See **?** for more details, which we will revisit in the linear regression lectures.

Another issue that can arise is when the treatments are not cross-randomized, but instead are correlated on one another. A simple example of this is sequenced treatments: e.g. treatment 2 is only given after treatment 1, and only a subset of individuals receive treatment 1. See Figure 3 for an example of this. In this case, it is not possible to identify the effect of D_2 separately from D_1: $E(Y_i(0,1) - Y_i(0,0))$ is not identified because $E(Y_i(0,1))$ is never observed. This rarely

happens in many cross-sectional settings, but is quite common in dynamic settings (our next topic).

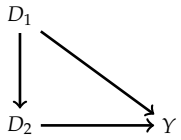

Figure 3: Correlated treatments

Treatment dynamics

We will briefly discuss the impact of treatments over time to set the stage for our study of panel data later in the class. Consider a setting where we now observe T time periods for a unit: $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{iT})$. Now, for each time period, there is a treatment D_{it}. It would be convenient to simply consider $Y_{it}(D_{it})$ as the potential outcome for an individual i in period t, but that would be very restrictive: it would assume that only the period t treatment affects the outcome in period t. A more general form would define a vector $\mathbf{D}_i = (D_{i1}, D_{i2}, \ldots, D_{iT})$, and define the potential outcome in period t as $Y_{it}(\mathbf{D}_i)$. In this case, however, we are perhaps too general: this allows for treatments in the future to affect current outcomes, which may be too strong.[5]

There are a large number of ways to simplify these potential outcomes. One simple way would be to restrict treatments to only affect outcomes in the future, and not the past. This is often referred to as the "no anticipation" assumption. The second is to assume that treatments will only turn on once: this allows the researcher to only consider the adoption date as the relevant period.[6]

As you can see, things become much more complex as soon as you allow for dynamic effects. In order to make progress, it will often be necessary to make restrictions on the dynamics to make the estimands identified. We will discuss these in more detail when we discuss difference-in-differences.

[5] This is often referred to as "anticipatory effects" in the literature.

[6] This is common in the staggered difference-in-difference setting.

The SUTVA hits the fan

In the discussion so far, the "interference" between treatments just comes from having multiple treatments to worry about, or from spillover across time. However, there are many other ways that treatments can interfere with one another. For example, what if treatments spill across units? What if the treatment of one unit affects the potential outcomes of another unit?

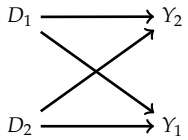

Figure 4: Interference between units

Recall the key assumption of Stable Unit Treatment Value Assumption (SUTVA): the potential outcomes of a unit do not vary with the treatment of other units. When could this be violated?

So many places

Why does failure of SUTVA create an issue? Recall our discussion regarding marginal estimands when there were multiple treatments:

even with random assignment, the estimates effect will be contaminated by others' treatment status, thereby leading to estimates that are not informative for the policy maker.

This type of problem is generally referred to as "interference." It is challenging for identification, estimation and inference. For now, we'll focus on identification. I will flag three versions of this problem:

1. Social interactons and peer effects
2. Spatial spillovers
3. Economic interactions – budget constraints, etc.

All these problems are versions of violation of SUTVA. With a clean, well-identified experiment, it is still possible to identify interesting estimands, but we may have to substantially modify our traditional estimators or make strong assumptions to make progress. One way to view this fact is that our original setting — SUTVA, binary treatment and a single time period — is a very special (and somewhat unrealistic) case.

Social interactions and peer effects

> A variety of terms in common use connote endogenous social effects, wherein the propensity of an individual to behave in some way varies with the prevalence of that behaviour in some reference group containing the individual. These effects may, depending on the context, be called "social norms", "peer influences", "neighbourhood effects", "conformity", "imitation", "contagion", "epidemics", "bandwagons", "herd behaviour", "social interactions", or "interdependent preferences".
>
> — **?**

Manski (1993) spawned a huge literature, much of which focused on the linear-in-means model.[7]. An inherent issue, in my view, is that many empirical papers jumped to this construction immediately. They did not have a structural interpretation in mind, but were instead interested in testing for the *statistical* presence of spillovers across individuals.

[7] There are theoretical models microfounding a linear-in-means outcome model, which typically involve some kind of quadratic cost to deviating from the group. See **?** for an example

Comment 2 (Historical context on peer effects)
? *focused on a linear-in-means structural equation*

$$Y = \underbrace{\beta E(Y|g)}_{\text{endogenous}} + \underbrace{\gamma_1 E(X|g)}_{\text{exogenous}} + \gamma_2 X + \underbrace{\gamma_3 g}_{\text{contextual}} + u.$$

Peers were not well-defined in the model, but empirically, were usually groups like classrooms or clubs. What is important to note about this model is that it is a structural *model of the outcome, Y. The reduced form is*

$$Y = \gamma_1/(1-\beta)E(X|g) + (\gamma_2/(1-\beta))X + (\gamma_3/(1-\beta))g + \tilde{u},$$

which is estimable under special exogeneity assumptions on X, g and $E(X|g)$.

An innovation in this literature was to start using network data to define the group structure. **?** *was a key paper in this literature, that reframed the Manski linear-in-means model to*

$$Y = \beta AY + \gamma_1 AX + \gamma_2 X + \epsilon_i,$$
$$Y = (I - \beta A)^{-1}\gamma_1 AX + (I - \beta A)^{-1}\gamma_2 X + (I - \beta A)^{-1}\epsilon_i$$

where A was an $n \times n$ matrix of individuals' connections. This was still a structural model, but allowed for richer data and more easily identified the effect of peers.

Given that most researchers studying peer effects were not initially motivated by a structural model, it seems more natural to initially take a statistical approach to the problem. Namely, we would like to identify the effect of spillovers across units. How can we approach this problem using the tools we've developed so far?

Given n individuals, for person i, how much interference can we allow? What types?

$$Y_i(D_1, D_2, \ldots, D_n)$$

is far more extreme than

$$Y_i(D_i, A\mathbf{D}_n).$$

This question is analogous to our setting with treatment dynamics: how much spillover should we allow? SUTVA is complicated by the fact that there is no natural "no anticipation" condition due to the

natural flow of time. As you might expect, there is no "one solution" in this setting. Certain restrictions need to be made to identify some estimands.

? is a very nice discussion of this in a *very* high-level way. One key assumption he highlights is that "anonymity" of treatment spillover is a very important assumption. This implies that if I have peers who are treated, it does not matter *which* of those peers are treated – the impact on me is identical. This is a very strong assumption, but it is a necessary one to identify the effect of peers. If each peer's effect is allowed to be unique, then the effect of peers is not identified, since there is no way to separate out the treatment's spillover effect from the effect of the peer itself.[8]

[8] It might be doable in some networks, but it would be very challenging to do so, and require exogenous network connections.

A key question to keep in mind when considering spillovers: are you attempting to estimate the *spillover* effect, or are you attempting to identify individual ATE in the presence of spillovers? These are very different estimands, and require different assumptions to identify. For the purposes of external validity, the latter is really only relevant if the context you apply the treatment in would have limited spillovers as well.

We now briefly discuss two papers in this space to give intuition.

? is a lynchpin paper in this setting that provides a framework for thinking about estimation and identification under general forms of interference. They use design-based inference, and consider the following generalized mapping.

Definition 1

For any generalized vector of interventions, $\mathbf{D}_n$, there's an experimental design which assigns probabilities over $\mathbf{D}_n$. There is then an exposure *mapping $f(\mathbf{D}_n, \theta_i)$ from these vectors to a treatment for an individual, which includes traits of an individual, θ_i (e.g. their network location) and the treatment vector, and maps it to an exposure outcome.*

This exposure mapping does two things. First, it makes restrictions on types of interactions (e.g. who can affect you and what type of effect it is).[9] Second, it maps the experimental design to a propensity score of the exposure treatment. This allows the use of Horvitz-Thompson estimators.

[9] Concretely: consider a network of peers affecting you. Is it the sum of your connected individuals in your network? Any exposure at all? Does it matter who in your network exposes you?

So where are the bodies buried in this method? You have to have a correctly defined exposure mapping, and you have to have a correctly defined experimental design. In the case of a randomized experiment, the latter is straightforward, but the former is not, and typically needs to be motivated by theory, or assessed for robustness. This is an active literature.

? studies null hypothesis tests in networks under interference. A key feature that this paper adds: testing specific types of analysis by creating "artificial" experiments. This paper is particularly powerful because it allows for testing in settings where there is uncertainty about the exposure mapping, and gives a framework for thinking about testing in a single network.

Comment 3

When thinking about experiments in networks (and other settings), the structure of spillovers is very important. It is extraordinarily helpful to identify settings where there are zero *spillovers. Having units (such as villages, roommate pairs, etc.) that are isolated from one another is a very helpful way to identify the effect of the treatments. If we permute the treatments across these groups, then we can assess the spillover effects in a very clean way.*

If, instead, we have only a single network, then we need to make strong assumptions about the structure of the network to identify the spillover effects. Namely, we need to have a well-defined exposure mapping that asserts that some units are sufficiently independent from others to serve as control units.

Figure 5: My views on social interactions summed up

It is already very hard to do research on spillovers. Make sure to not ignore the difficult identification challenges and assumptions that you'll need to make. If you need a model, that's great! But often you are just interested in starting from a statistical perspective, which suggests you should focus on a design-based approach as in **?**.

Spatial Spillovers

Much of the spatial literature has sat in the same literature as social interactions. Distance on a network graph can be viewed as a similar distance metric to geographic (or economic) distance.Similar A matrix, and consequentially similar structural models are proposed.

The **?** setting allows for this as well. From an identification standpoint, there is nothing deeply different here relative to networks, except that distance is potentially more continuous / complex. When we revisit simulated instruments, we will discuss some interesting implications raised by **?**.

Economic interactions

Consider the following simple experiment – I give one half of people in the economy checks for $2000 dollars. I then study the impact of these checks on their consumption. Why might the effects be differ-

ent than if I had run this experiment on a small share of individuals?

The economic spillovers coming through budget constraints are hugely important, but also deeply challenging as well. They require, often, modeling assumptions about spillovers. I will discuss two examples from the literature to give a flavor of the issues and solutions.

? studies the impact of fiscal stimulus on local employment. The key identification strategy is to use cross-region incidence of fiscal stimulus to identify multipliers on local employment. The paper argues that cross-region evidence bounds the estimand of interest, the impact of a *national* stimulus, from below.

> Drawing on theoretical explorations, I argue that the typical empirical cross-sectional multiplier study provides a rough lower bound for a particular, policy-relevant type of national multiplier, the closed economy, no-monetary-policy-response, deficit-financed multiplier. The lower bound reflects the high openness of local regions, while the "rough" accounts for the small effects of outside financing common in cross-sectional studies.

? use cross-firm experiment to influence the allocation of credit. Some firms got lots more credit! Some did not. How to aggregate up this affect? E.g. the policy effect is estimated by differencing the impact of the change on those who were more directly exposed vs. not – however, this doesn't tell us about the aggregate impact on the economy. The paper argues, using economic theory, that these issues can be safely ignored under certain assumptions.

> Our paper bridges these two approaches. We offer a method to measure allocative efficiency in a (quasi-) experimental settings. This method works as follows. An econometrician observes firm-level data in an economy where a (quasi-) natural experiment has taken place. This experiment changes the set of frictions faced by treated firms while leaving control firms unaffected. Under the appropriate identifying assumption, the econometrician can estimate the causal effect of the experimental firm-level outcomes, using classic difference-in-difference estimators. Standard policy evaluations typically estimate treatment effects on firm size or employment. However, these treatment effects alone cannot speak to allocative efficiency. To do so, we show that the econometrician needs to estimate treatment effects on the distribution of log marginal products of capital (lMRPKs). These estimates can then be injected in a simple aggregation formula to answer two simple questions: (i) how much did the actual policy change contribute to changes in aggregate efficiency (ex post evaluation)? (ii) how would aggregate efficiency have changed if the policy had been extended to all firms in the economy (scale-up)?