Estimating Geographic Spillover Effects of COVID-19 Policies From Large-Scale Mobility Networks*

Serina Chang[1], Damir Vrabac[1], Jure Leskovec[1], Johan Ugander[2]

[1]Department of Computer Science, Stanford University
[2]Department of Management Science & Engineering, Stanford University
{serinac, dvrabac, jure, jugander}@stanford.edu

Abstract

Many policies in the US are determined locally, e.g., at the county-level. Local policy regimes provide flexibility between regions, but may become less effective in the presence of geographic spillovers, where populations circumvent local restrictions by traveling to less restricted regions nearby. Due to the endogenous nature of policymaking, there have been few opportunities to reliably estimate causal spillover effects or evaluate their impact on local policies. In this work, we identify a novel setting and develop a suitable methodology that allow us to make unconfounded estimates of spillover effects of local policies. Focusing on California's Blueprint for a Safer Economy, we leverage how county-level mobility restrictions were deterministically set by public COVID-19 severity statistics, enabling a regression discontinuity design framework to estimate spillovers between counties. We estimate these effects using a mobility network with billions of timestamped edges and find significant spillover movement, with larger effects in retail, eating places, and gyms. Contrasting local and global policy regimes, our spillover estimates suggest that county-level restrictions are only 54% as effective as statewide restrictions at reducing mobility. However, an intermediate strategy of macro-county restrictions—where we optimize county partitions by solving a minimum k-cut problem on a graph weighted by our spillover estimates—can recover over 90% of statewide mobility reductions, while maintaining substantial flexibility between counties.

1 Introduction

Many policies in the United States—COVID-19 restrictions, environmental regulations, and laws controlling the sales of e-cigarettes, firearms, and controlled substances—are determined at the state- or county-level. Local policy regimes provide flexibility between regions, allowing policymakers to set regulations depending on local circumstances (e.g., COVID-19 severity) and the preferences of their constituents (e.g., on gun control). On the other hand, allowing policies to be set locally often results in differing levels of restrictiveness between neighboring regions. These differences can lead to *geographic spillovers*, where populations circumvent restrictions by traveling to less restricted regions nearby. Spillovers risk undermining the efficacy of local policies; for example, if banned goods are imported across state borders or if, during the pandemic, individuals in counties under lockdown continue to visit places in neighboring counties. Furthermore, spillovers can affect important downstream consequences. For example, the movement of individuals from more restricted (and possibly more infected) regions to less restricted (and possibly less infected) regions during the pandemic could result in greater overall spread of the virus.

However, there are few opportunities to reliably estimate causal spillover effects. Researchers cannot run experiments to randomly assign policies to states and counties, and causal identification is difficult in most observational studies, due to the presence of confounders. For example, attempting to study the effects of COVID-19 restrictions (e.g., closing restaurants) on mobility patterns introduces potential confounding covariates that predict both the treatment and the outcome, such as current COVID-19 severity in the region and the population's demographics. Prior work has attempted to address these confounders by controlling for them, but there could always be unobserved or unknown confounders that bias causal estimates. Furthermore, the decentralized nature of policymaking that gives rise to potential spillovers also often results in varying policy definitions and implementations across regions. This heterogeneity makes it difficult to define a consistent treatment whose effects we can measure.

In this work, we introduce a setting in which we can make unconfounded estimates of the spillover effects of consistent policies. We focus on California's Blueprint for a Safer Economy, a statewide policy framework that determined weekly county-level mobility restrictions for all 58 counties in California from August 2020 to June 2021. The Blueprint consisted of four tiers that corresponded to policies of decreasing restrictiveness. At the start of each week, each county's tier was determined based on that county's COVID-19 metrics (case rate and test positivity) in the preceding weeks. The California Blueprint presents a unique opportunity for studying spillover for three reasons: (1) neighboring counties were frequently in differing tiers, enabling the analysis of spillovers from more restricted to less restricted counties; (2) tiers were defined in the same way across counties, yielding a consistent treatment; (3) tiers were deterministically assigned at the thresholds of COVID-19 metrics. These three ingredients allow us to develop a causal inference framework based on regression discontinuity design to

*This is the extended version of a paper accepted to AAAI'23.

make unconfounded estimates of spillover effects.

To capture spillover, we focus on cross-county mobility in a large-scale mobility network. Our network is a dynamic bipartite graph that represents the weekly movements of individuals from census block groups (CBGs) to specific points-of-interest (POIs) such as restaurants and grocery stores. Our objective is to estimate the effect of pairwise county tiers on the number of visits from each CBG to POI. The mobility network for California contains around 23,000 CBGs and 130,000 POIs, with nearly 3 billion edges per week. We use stochastic gradient descent, with loss-corrected negative sampling, to make estimation computationally feasible in this large-scale setting. Studying mobility patterns at the POI-level enables us to estimate heterogeneous treatment effects for POI categories; this ability is particularly relevant since tier restrictions were often industry-specific.

Finally, our spillover estimates allow us to quantify the cost of spillovers on policies across spatial scales. In the presence of spillovers, we find that county-level restrictions are, on average, only 54% as effective as statewide restrictions at reducing mobility. However, intermediate strategies of macro-county restrictions—when counties are grouped intelligently—can balance the trade-off between the policy flexibility and efficacy. We show that finding the most effective county partition for a given spatial granularity is equivalent to solving a minimum k-cut problem on an undirected county graph weighted by our spillover estimates. Using this approach, we identify macro-county restrictions that recover over 90% of statewide mobility reductions, while maintaining substantial flexibility between counties.

In summary, our contributions are as follows:

- **Setting:** we identify a novel setting for studying spillovers where the same set of policies was applied with the same thresholds to many areas;
- **Methods:** we develop a regression discontinuity (RD) design framework that allows us to make unconfounded estimates of heterogeneous spillover effects in this setting, estimated over a large-scale mobility network containing billions of edges;
- **Analyses:** we demonstrate significant spillover effects in many POI groups and evaluate the costs of these spillovers on policies across spatial scales.

In a complex, interconnected world with few opportunities to reliably estimate policy effects, our work is among the first to identify a setting where spillovers can be rigorously estimated and to develop an appropriate methodology to estimate and evaluate the effects of spillovers.[1]

2 Related Work

Spillovers often arise from decentralized policymaking for interconnected regions. For example, Sigman (2005) finds that water quality is lower at stations downstream of states that are authorized to control their own water programs, since they "free-ride." Coates and Pearson-Merkowitz (2017) show that in states with stronger gun laws, there is an increased likelihood of gun imports from states with weaker gun laws. Bronars and Lott (1998) show that while a concealed handgun law led to a reduction in crime in the state, it also led to an increase in crimes in neighboring states, suggesting that criminals were crossing borders. Hao and Cowan (2017) find that legalization of recreational marijuana in a state leads to an increase in marijuana-related arrests in bordering states. Spillovers also arise in online contexts, where instead of crossing geographic borders, users can migrate across platforms if they are banned on one platform; furthermore, levels of toxicity and radicalization are sometimes higher on the new, often less regulated platforms, compromising the efficacy of the original content moderation (Ribeiro et al. 2021; Ali et al. 2021).

In the context of COVID-19, prior research has mostly focused on the direct effects of policies on population health or behavior, without explicitly modeling spillovers (Chernozhukov, Kasahara, and Schrimpf 2021; Nguyen et al. 2020; Brauner et al. 2020). Chandrasekhar et al. (2021) investigate disease spillovers between interconnected regions in a model-based setting and Holtz et al. (2020) provide early evidence of mobility spillovers, showing that a state's population reduced its own mobility when neighboring states implemented shelter-in-place policies. Most related to our work is Zhao, Holtz, and Aral (2021), who use a difference-in-difference approach to estimate the effects of COVID-19 policies on mobility and provide evidence of spillovers in cross-state travel. We build on this work by addressing two primary limitations of their study: first, the authors note that their estimates could be confounded by unobserved, time-varying factors; other research on spillovers also suffers from potential confounding, using difference-in-difference approaches (Hao and Cowan 2017; Holtz et al. 2020) or regressions (Coates and Pearson-Merkowitz 2017; Sigman 2005; Bronars and Lott 1998). Second, in order "to create sufficient statistical power to identify causal effects," the authors collapse different policy interventions into general policy "types" (e.g., resuming dine-in and lifting gathering restrictions are both counted as reopening), which violates assumptions of consistent treatment.

In contrast with prior work, we are able to identify unconfounded spillover estimates for a single set of policies by applying our RD-based framework to California's Blueprint for a Safer Economy. Furthermore, by estimating effects on the CBG-POI network, our model enables the analysis of counterfactual fine-grained mobility patterns under different pandemic policies. Understanding mobility patterns has been essential to controlling the spread of COVID-19 (Buckee et al. 2020), and many researchers rely on fine-grained mobility data to model the effect of mobility on the spread of the virus (Badr et al. 2020; Chinazzi et al. 2020; Kraemer et al. 2020; Chang et al. 2020, 2021; Nouvellet et al. 2021). Our model furthers such analyses by investigating the complex effects of policy interventions on mobility, closing the gap from policy to behavior to COVID-19 outcomes.

[1]The code to run our experiments and regenerate figures is available at https://github.com/snap-stanford/covid-spillovers. We also provide our constructed Z variables (Section 4) that can be used with RD design to estimate the effects of the California Blueprint tiers on spillovers and other outcomes.

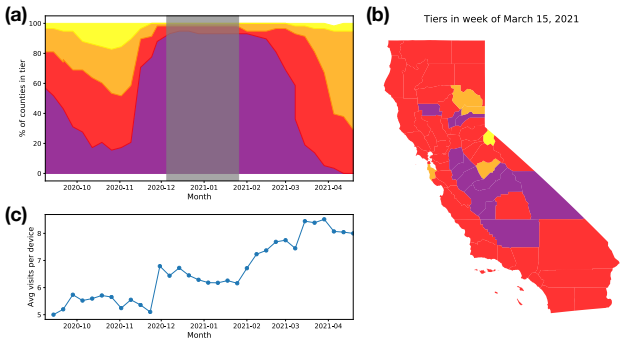


Figure 1: Primary data sources. (a) Percentage of California counties in Blueprint tiers—purple, red, orange, and yellow—over time (greyed-out period represents Regional Stay-At-Home Order); (b) Tiers in the week of March 15, 2021; (c) Average weekly visits per device over time.

3 Data

California Blueprint for a Safer Economy. The Blueprint was implemented for all 58 counties in California from August 30, 2020 to June 15, 2021. It consisted of four tiers: purple ("widespread"), red ("substantial"), orange ("moderate"), and yellow ("minimal"). These tiers corresponded to mobility policies of decreasing restrictiveness; for example, in the purple tier, most non-essential indoor businesses were closed, while in yellow, they could be open with modifications. We use the archived data sheets from the California Department of Public Health (CDPH), which provide detailed documentation of every county's weekly tier assignment and the COVID-19 metrics used to make those assignments.[2] In Figure 1a, we visualize the progression of counties through tiers over time; we grey out the period from December 5, 2020 to January 25, 2021, during which most of the state was under a Regional Stay-At-Home Order (CDPH 2020). We can see that counties generally moved through similar tiers at similar times, which is expected, since COVID-19 severity was correlated across counties. However, in many weeks, we also see substantial representation from at least two different tiers. For example, in the week of March 15, 2021, there were 11 counties in the purple tier, 42 in the red tier, 4 in the orange tier, and 1 in the yellow tier (Figure 1b). Many of these differing tiers appeared between adjacent counties, enabling the analysis of spillovers across county borders.

Mobility network. We use data from SafeGraph, a company that anonymizes and aggregates location data from mobile apps. For each POI, SafeGraph provides weekly estimates of where visitors are coming from, aggregated over CBGs.[3] This creates a dynamic, bipartite graph between CBGs and POIs, where an edge weight Y_{ijw} represents the number of visits recorded by SafeGraph from CBG c_i to POI p_j in week w. SafeGraph also reports how many devices they recorded in each CBG and week. Incorporating device counts into our model allows us to account for varying coverage across CBGs and over time.

In Figure 1c, we show the average number of weekly visits recorded per device over time, aggregated over the entire CBG-POI network for California. We see that visits increased post-Regional Stay-at-Home as Blueprint tiers decreased in restrictiveness. However, various latent variables could explain this correlation, such as reduced COVID-19 severity leading to less restrictive tiers and less fear of visiting places. Thus, it is necessary to develop a robust causal framework that allows us to disentangle tier effects from confounders, which we describe in the following section.

4 Causal framework

To capture spillovers, our objective is to estimate the effect of *pairwise* tiers on cross-county mobility. The key to our causal framework is that we can utilize RD design, which is widely recognized as "one of the most credible non-experimental strategies for the analysis of causal effects" (Cattaneo, Idrobo, and Titiunik 2020). In a typical RD design, units are assigned to the treatment or control condition according to an exogenously determined threshold of a single continuous variable, known as the assignment variable (or running or forcing variable). Researchers can then compare the outcomes for units just below the threshold to units just above the threshold to estimate the *local* causal effect of treatment. A primary advantage of RD design is that it achieves unconfoundedness, without needing to control for all possible confounders. This is because the unconfoundedness assumption is met: treatment assignment is conditionally independent of potential outcomes, given covariates (Imbens and Lemieux 2008). This assumption is clearly met in RD design, since treatment assignment is determined by the assignment variable, and so, conditioned on covariates, there is no variation in treatment.

Our problem generally fits RD design, since Blueprint tiers were assigned at the thresholds of continuous COVID-19 metrics. We focus on the threshold between the purple and red tiers, since they were the adjacent pair with the most support. However, we need to extend generic RD design in two ways: (1) to account for multiple assignment variables, since tiers were assigned based on numerous COVID-19 metrics, (2) to account for multiple treatment conditions, since we are considering pairwise tiers as our treatment. We describe our approach in the following sections.

Assigning Blueprint tiers. First, let us focus on the problem of determining a single county's tier, T_{iw}, from its COVID-19 metrics. Tier assignments depended on three metrics: adjusted case rate, test positivity rate, and a health equity metric, which was the test positivity rate in the most disadvantaged quartile of neighborhoods (CDPH 2021b). To advance to a less restricted tier, counties needed to meet the criteria for movement for two consecutive weeks (CDPH 2021a). For a large county (population over 106,000), the criteria to move from purple to red could be met in two ways: (1) by meeting the thresholds for the red tier for all three metrics, (2) by meeting the thresholds for test positiv-

[2]https://www.cdph.ca.gov/Programs/CID/DCDC/Pages/COVID-19/CaliforniaBlueprintDataCharts.aspx

[3]https://docs.safegraph.com/docs/weekly-patterns

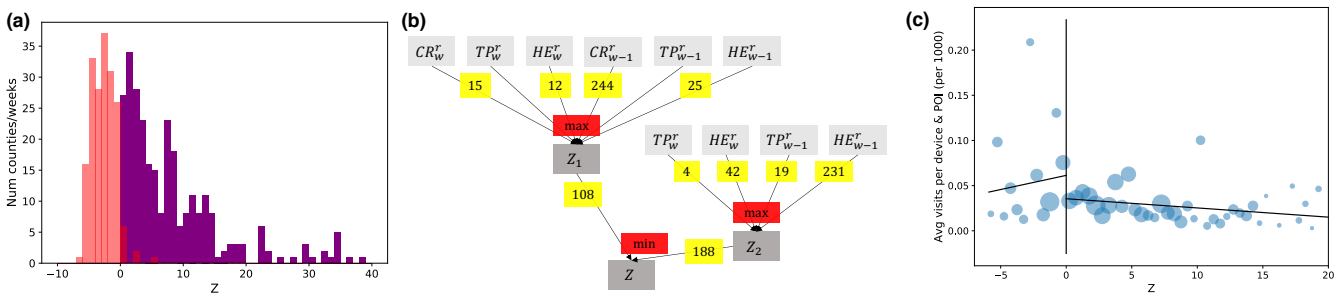


Figure 2: Visualizing our Z variable. (a) Z almost perfectly separates counties in the purple and red tiers. (b) Triggering patterns for Z (among large counties compliers). (c) Cross-county mobility vs Z. All source counties are in purple ($0 \leq Z \leq 5$) and the x-axis represents the target county's Z. Black lines represent linear fits and dots are average outcomes per bin (size represents the number of data points in the bin). We observe a discontuinity at $Z = 0$, when the target county changes from purple to red.

ity and health equity for the *orange* tier, thus exchanging adjusted case rate for more stringent thresholds on the other two. Small counties (population under 106,000) only had one possible path, which was to meet the adjusted case rate and test positivity thresholds for the red tier. Small counties were not required to meet the health equity thresholds, but needed to demonstrate their commitment to equity through other plans (CDPH 2021b). For most of the duration of the Blueprint, the purple-red threshold for adjusted case rate was 7 per 100,000 and 8% for test positivity and health equity (and 5% for the red-orange threshold). The purple-red threshold for adjusted case rate was increased to 10 per 100,000 on March 12, 2021, after 2 million vaccines had been administered statewide (Ibarra and Becker 2021).

Constructing a single assignment variable Z. We take a *centering* approach to RD design with multiple assignment variables (Wong, Steiner, and Cook 2013). That is, we first center each of the assignment variables by subtracting their respective thresholds, then apply min/max aggregations to the centered variables in order to construct a new assignment variable Z that can singly determine a unit's treatment. More formally, we design a function $f : \mathbb{R}^m \to \mathbb{R}$ that maps a county's m COVID-19 metrics to a single continuous variable, Z_{iw}. For a large county, the m metrics include the county's adjusted case rate (CR), test positivity (TP), and health equity metric (HE) from the preceding two weeks; for a small county, only adjusted case rate and test positivity. Our mapping f satisfies the key property that $Z_{iw} < 0$ if and only if the county was assigned to the red tier.

Let CR^{red}_{iw} represent the adjusted case rate for the county in week w with the purple-red threshold subtracted, and let us define other terms similarly. We construct Z_{iw} for large counties as follows:

$$Z_{1iw} = \max(CR^{\text{red}}_{iw}, TP^{\text{red}}_{iw}, HE^{\text{red}}_{iw}, CR^{\text{red}}_{iw-1}, TP^{\text{red}}_{iw-1}, HE^{\text{red}}_{iw-1}) \quad (1)$$

$$Z_{2iw} = \max(TP^{\text{orange}}_{iw}, HE^{\text{orange}}_{iw}, TP^{\text{orange}}_{iw-1}, HE^{\text{orange}}_{iw-1}) \quad (2)$$

$$Z_{iw} = \min(Z_{1iw}, Z_{2iw}). \quad (3)$$

For small counties, we only have

$$Z_{iw} = \max(CR^{\text{red}}_{iw}, TP^{\text{red}}_{iw}, CR^{\text{red}}_{iw-1}, TP^{\text{red}}_{iw-1}). \quad (4)$$

In Figure 2a, we show that our new Z variable almost perfectly separates the counties in the purple and red tiers. Over the 9-week period from February 1 to March 29, 2021, there were 480 counties/weeks in the purple or red tier, and 471 of them follow that $Z_{iw} < 0$ if and only if the county is in the red tier. We manually check the non-compliers and find that they were cases of counties, mostly small, that were allowed to remain in the red tier upon special request, as noted in the CDPH documentation.

To interpret our new Z variable, we also analyze its "triggering" patterns; that is, for each min/max aggregation, which input is the minimum or maximum (Figure 2b). Since $Z < 0$ moves the county into the red (less restricted) tier, a maximum can be interpreted as holding the county back and a minimum as improving the county's prospects. For large counties, we find that the most frequent maximum for the first criteria Z_1 is the adjusted case rate from week $w - 1$. For the second criteria Z_2, the most frequent maximum is the health equity metric from week $w - 1$. This reflects trends from this time period: COVID-19 severity was improving over time, so week $w - 1$ tended to have higher rates than week w, and health equity (i.e., test positivity in the most disadvantaged quartile) tended to be worse than the overall test positivity. Interestingly, we also find that Z_2 triggers more often than Z_1, when taking the minimum between them. This indicates that this alternative path—meeting more stringent test positivity and health equity thresholds and dropping adjusted case rate—substantially helped counties move toward less restricted tiers.

RD design with pairwise treatments. We can now formulate an RD design problem where treatment (purple/red tier) is assigned at the threshold of a single continuous variable (Z). Since we are interested in spillover effects in this work, we use *cross-county* mobility as our outcome. However, our RD framework is general and could be applied to study the effects of Blueprint tiers on a variety of outcomes, such as mask-wearing rates, vaccination rates, and COVID-19 cases and deaths.

With cross-county mobility as our outcome, our treatment becomes pairwise to capture the tier of each county, and we have four treatment conditions: PP, PR, RP and RR,

where P and R represent the purple and red tiers, respectively. We are particularly interested in the difference between PP and PR, since this difference indicates whether individuals from a restricted county will increase their visits to another county when that other county becomes less restricted. In Figure 2c, we illustrate this comparison. We consider all source counties that were in the purple tier and plot their mobility to target counties that were either in the purple or red tier. The x-axis represents Z for the target county, so that the region to the left of $Z = 0$ represents the PR condition and the region to the right represents PP. We see a discontinuity in visits at $Z = 0$, indicating that there is indeed a local effect on cross-county visits when a neighboring county changes from more to less restricted. In the following section, we estimate this effect more precisely by defining a zero-inflated Poisson regression model that we fit to the rich CBG-POI mobility network with covariates.

Poisson regression model. We define a Poisson regression model to describe visits from CBGs to POIs. For a given CBG c_i, POI p_j, and week w, the Poisson rate λ_{ijw} is

$$\lambda_{ijw} = \exp(\beta_0 + \beta_1 Z_{iw} + \beta_2 Z_{jw} + \boldsymbol{\beta_3}^T \mathbf{X}_{ijw} + \beta_{T_{iw},T_{jw}}), \tag{5}$$

where the β terms are model parameters, Z_{iw} and Z_{jw} represent the Z variables for c_i's and p_j's counties in this week, T_{iw} and T_{jw} describe their respective tiers, and $\mathbf{X}_{ij}$ contains other covariates. Those covariates include the distance between the POI and CBG, SafeGraph's CBG device count in that week, CBG demographics from US Census, and POI attributes (area in square feet, NAICS code). Spillover effects are captured in the difference between the $\beta_{T_{iw},T_{jw}}$ terms: for example, $\exp(\beta_{PR} - \beta_{PP})$ represents the multiplicative increase in visits when a POI changes from the purple to red tier, while the CBG remains in purple. To capture heterogeneous treatment effects, we learn separate $\beta_{T_{iw},T_{jw}}$'s for different POI groups (Table A2). We also learn separate β_1's and β_2's for our four different constructions of Z that reflect two binary dimensions: 1) large vs. small county, 2) before vs. after March 12, 2021, when the statewide vaccine goal was met and the adjusted case rate threshold was increased.

The CBG-POI network is very large, with billions of edges, but over 99% of the edges represent zero visits. Thus, we zero-inflate our Poisson model, based on the notion that observed zeros in zero-heavy data may represent actual preferences, but could also reflect lack of exposure (Liu and Blei 2017), i.e., the CBG had never heard of the POI. We represent each number of visits Y_{ijw} as drawn from a mixture of a Poisson(λ_{ijw}) and δ_0 (a point mass on 0), with mixing parameter π_{ij}. We assume the likelihood of exposure is inversely proportional to the distance d_{ij} between the CBG and POI and define $\pi_{ij} = \frac{1}{1+\alpha_1 d_{ij}^{\alpha_2}}$, where the α terms are learned. Then, our generative model for Y_{ijw} is

$$b_{ijw} \sim \text{Bern}(\pi_{ij}) \tag{6}$$

$$Y_{ijw} \sim \begin{cases} \delta_0 \text{ if } b_{ijw} = 0, \\ \text{Poisson}(\lambda_{ijq}), \text{ otherwise.} \end{cases} \tag{7}$$

In this mixture, the likelihood of a single data point given model parameters θ is

$$\Pr(Y_{ijw} = y|\theta) = \begin{cases} (1 - \pi_{ij}) + \pi_{ij} e^{-\lambda_{ijw}}, \text{ if } y = 0 \\ \pi_{ij} \frac{\lambda_{ijw}^y e^{-\lambda_{ijw}}}{y!}, \text{ otherwise.} \end{cases} \tag{8}$$

We fit our model using gradient descent, with negative log likelihood as our model loss.

Data filtering and bandwidth selection. We focus our experiments on the 9-week period following the Regional Stay-At-Home Order, during which we could almost perfectly separate the purple and red tiers with our Z variable (Figure 2a). We keep all CBGs with at least 50 non-zero visits (to any POIs) during this period and POIs with at least 30 non-zero visits (from any CBGs), so that we focus on CBGs and POIs for which SafeGraph has more reliable coverage; this filtering leaves 22,972 CBGs and 128,655 POIs. Due to the specifics of our RD-based analysis, we cannot keep every CBG-POI pair from every week. First, we do not fit the model on data from the week of March 8, 2021, since the purple-red threshold for adjusted case rate was changed in the middle of the week (due to the statewide vaccine goal being met). In the remaining 8 weeks, we keep all data points that meet the following criteria:

- CBG c_i and POI p_j lie in adjacent counties,
- T_{iw} and T_{jw} are both in the purple or red tier,
- Both are compliers, i.e., T_{iw} is red if and only if $Z_{iw} < 0$, and likewise for T_{jw},
- Z_{iw} and Z_{jw} both lie within a bandwidth h of 0.

In total over the 8 weeks, we keep 1.4 billion data points after filtering (Table A3).

We only keep data points that fall within the bandwidth since our goal is to estimate the *local* effect of changing tier pairs at the purple-red threshold ($Z = 0$). By requiring both Z_{iw} and Z_{jw} to fall within the bandwidth, we interpret our resulting parameters as estimated effects at the *joint* cutoff, when both the CBG and the POI are at the threshold.[4] Bandwidth selection introduces a bias-variance tradeoff, with larger bandwidths corresponding to greater bias but reduced variance. We err on the side of larger bandwidths in this work, out of concern for variance. Even though we have over a billion data points, our assignment variable Z only varies at the level of counties and, thus, bandwidths that are too small could lead to very few counties represented, particularly for the PR or RP treatment conditions, which appear less often. We choose $h = 5$, which keeps most of the counties in the red tier, but drops many of the counties in purple (Figure 2a). We show in the Appendix that each treatment condition is well-represented at this bandwidth, with a diversity of county pairs (Table A4). Furthermore, we conduct sensitivity analyses with $h = 4$ and $h = 6$ and show that results remain highly similar (Figure A1).

[4] Alternatively, RD design with multiple assignment variables can estimate effects along the threshold frontiers, i.e., varying one assignment variable while fixing the other one at its threshold (Papay, Willett, and Murnane 2011). For simplicity, we focus on effects at the joint threshold.

Loss-corrected negative sampling. To make estimation computationally feasible in this large-scale setting, we perform negative sampling. Specifically, for each zero data point (i, j, w), we define its sampling probability s_{ijw} as inversely proportional to the distance between the CBG and POI ($s_{ijw} \propto \frac{1}{1+d_{ij}}$). We do this to upweight "hard" negative samples; that is, since far-apart CBGs and POIs are highly unlikely to have any visits, the model learns more from nearby CBGs and POIs with zero visits. However, a unique aspect of our problem—which does not typically appear in other machine learning prediction problems where negative sampling might be used, such as link prediction or learning word embeddings—is that because we seek to interpret the model parameters as effect sizes, our learned model parameters need to be unbiased estimates of the model parameters when learned on the full data. Left uncorrected, negative sampling biases our model parameters by greatly reducing the number of zeros in the training data.

In the Appendix we show that by weighting each sampled zero data point by $\frac{1}{s_{ijw}}$ when computing the overall loss (negative log likelihood), our stochastic gradient (which is stochastic from sampling) forms an unbiased estimate of the true gradient, which ultimately guarantees unbiased parameter estimates assuming proper model specification. We also show that upweighting harder negative samples, as well as increasing the size of the sample, decreases the variance of the stochastic gradient, providing formal validation of these techniques. In our experiments, we retain 2% of the zero data points, with sampling probabilities inversely weighted by distance. We verify that after incorporating our loss corrections, different negative sampling schemes arrive at the same average parameters, but distance-weighting and larger samples decrease variance. The agreement between the estimates from different negative sampling schemes is consistent with the underlying model being properly specified.

Uncertainty quantification with bootstrapping. We run 30 trials, where in each trial, we perform negative sampling on the zero data points and we sample N_{nnz} non-zero data points with replacement, where N_{nnz} is our total number of non-zero data points. For a given estimand, such as $\tau_{PR} = \exp(\beta_{PR} - \beta_{PP})$, we compute its 95% confidence interval as $\bar{\tau}_{PR} \pm 1.96 \cdot \hat{\sigma}_{\tau_{PR}}$, where $\bar{\tau}_{PR}$ and $\hat{\sigma}_{\tau_{PR}}$ are its sample mean and standard deviation over trials, respectively. This procedure captures uncertainty from the data and from negative sampling, although we show that, given our chosen negative sampling scheme, the former accounts for the vast majority of the variance (Figure A2).

5 Results

Spillover estimates. We learn heterogeneous effects for different POI groups, where we consider all "top"-categories (first 4 digits of the POI's NAICS code) with at least 1000 POIs.[5] We also include separate effects for 4 "sub"-categories (first 6 digits) of interest, all of which have over 1000 POIs. In Table A2, we provide the NAICS codes, descriptions, and number of POIs in each POI group.

[5]Following prior work using SafeGraph data (Chang et al. 2020), we drop the category "Elementary and Secondary Schools" due to poor coverage of children from cell phone data.

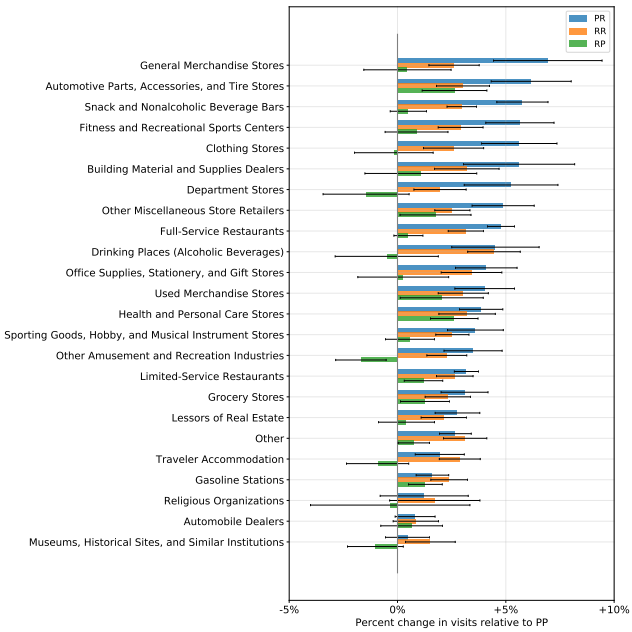


Figure 3: Estimated spillover effects across POI groups, with 95% confidence intervals.

We present our spillover results in Figure 3. First, we find significant positive PR effects in 21 out of 24 groups (all results remain significant with Bonferroni correction). That is, visits from the CBG increase significantly when the POI's county changes from purple to red, while the CBG's county remains in purple. This indicates spillovers, as people from more restricted counties spill over in less restricted, adjacent counties. Furthermore, we observe varying effect sizes; for example, with larger effects in retail (General Merchandise Stores, Automotive Stores, Clothing Stores, Building Material and Supplies Dealers, Department Stores), most eating places (Snack Bars, Full-Service Restaurants, Drinking Places), and gyms (Fitness and Recreational Sports Centers). Smaller effects are in essential retail (Grocery Stores, Gas Stations), hotels (Traveler Accommodation), malls (Lessors of Real Estate), museums, historical sites, and nature parks (Museums, Historical Sites, and Similar Institutions). This heterogeneity in effect size may partially reflect differences across sectors in tier restrictions. For example, essential retail, hotels, and malls remained open indoors with modifications under both tiers, while restaurants and gyms—which have larger estimated spillovers—were outdoor only under the purple tier and open indoors with modifications under the red tier (Table A1).

We also observe significant positive RR effects in 22 POI groups (21 with Bonferroni correction), as in, visits increase significantly when both the CBG and POI are in red, compared to when they are both in purple. Furthermore, in most POI groups, the PR effect is *larger* than the RR effect (although not always significantly so). This suggests an interac-

tion effect: individuals not only spill over into adjacent counties when those counties become less restrictive, but also the spillover is larger if their home counties are more restrictive. Finally, we observe a varying effect of RP, which represents when the CBG changes from purple to red, while the POI remains in purple. The effect is slightly positive or negative for some POI groups, but significant in neither direction for most. We hypothesize that two mechanisms take place here: on one hand, since the POI is in a more restricted tier than the CBG, it becomes less appealing; on the other hand, since the CBG opened up, its population is more willing to travel. These counteracting mechanisms may explain the varying and weak RP effects across POI groups.

Local vs. global restrictions. To contrast local and global approaches to policymaking, we use our fitted model to compare counterfactual mobility reductions under county-level vs. hypothetical statewide restrictions. Formally, let $\mathbf{T} \in \mathbb{R}^{58}$ represent the treatment vector for all counties. For each county A, we estimate this county's expected mobility (out-degree in the mobility network) under three treatment conditions: when the entire state is in the red tier ($\mathbf{T}_R$), when the entire state is in the purple tier ($\mathbf{T}_P$), and when only this county is in purple while the rest of the state remains in red ($\mathbf{T}_A$). We then compare the mobility reduction that a county would experience by going to purple on its own, relative to the statewide shutdown, where all counties go to purple:

$$r(A) = \frac{\mathbb{E}[out(A)|\mathbf{T}_R] - \mathbb{E}[out(A)|\mathbf{T}_A]}{\mathbb{E}[out(A)|\mathbf{T}_R] - \mathbb{E}[out(A)|\mathbf{T}_P]}. \quad (9)$$

We calculate $\mathbb{E}[out(A)|\mathbf{T}]$ as the sum over within-county visits and out-of-county visits:

$$\mathbb{E}[out(A)|\mathbf{T}] = \mathbb{E}[Y_{AA}|T_A] + \sum_{B \in N(A)} \mathbb{E}[Y_{AB}|T_A, T_B], \quad (10)$$

where Y_{AB} represents the total number of visits from any CBG in county A to any POI in county B. When we use our fitted model to compute the conditional expectation of Y_{AB} given tiers, we assume $Z = 0$ for all CBGs and POIs, since our RD-based framework estimated tier effects at the joint cutoff. We also marginalize over the remaining dynamic covariate, the CBG's weekly device count, by taking each CBG's average device count over the 9-week period that we study. In the Appendix, we describe how to efficiently compute Y_{AB} for all pairs of counties and possible tier pairs.

We estimate that counties applying local restrictions can only achieve, on average, 54.0% (46.4%–61.7%) of the reduction in mobility that they would experience under a statewide shutdown. Small counties are particularly affected, keeping only 41.7% (31.0%–52.4%) of their statewide mobility reduction, while large counties retain 62.1% (56.3%–67.9%). While we assume that the reduction in mobility *within* the county stays the same between local and global regimes, the difference arises from the increase in *out-of-county visits* when all surrounding counties are less restricted in the red tier; this is why smaller counties are especially hard-hit, since a larger fraction of their mobility tends to be out-of-county. We also consider a less

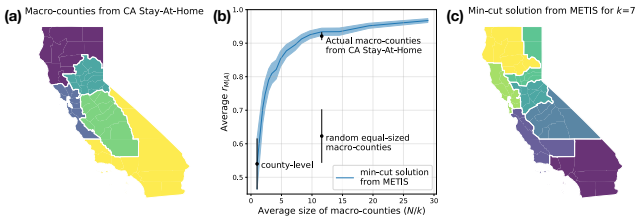


Figure 4: Macro-county restrictions. (a) Actual macro-counties from California's Stay-At-Home Order. (b) Trade-off between flexibility (lower macro-county size) versus efficacy (higher $r_M(A)$), with 95% confidence intervals. (c) Macro-county partition for $k = 7$, computed by METIS.

extreme case, where instead of having all surrounding counties in red, we use the actual configuration of tiers from the Blueprint. We still observe serious costs to efficacy in these more realistic settings: over the course of our study period, as the number of counties in purple fell from 40 to 11 to 3, the average percent of mobility reduction kept for counties in purple (compared to statewide shutdown) fell from 94% to 75% to 65% (Figure A4). These substantial decreases in efficacy demonstrate the cost of spillovers on local policies.

Balancing efficacy and flexibility. Although local policies are less effective in the presence of spillovers, global policies are often too blunt and inflexible. In our final analysis, we explore this trade-off between efficacy and flexibility across policies at different spatial scales. Instead of being entirely local (county-level) or global (statewide), intermediate strategies could be implemented at the macro-county level. California in fact pursued such a strategy with its Regional Stay-At-Home Order (Newsom 2020) that grouped counties into 5 macro-counties, each containing 11–13 counties (Figure 4a). Given a county partition M, we extend our analysis to compute $r_M(A)$, the ratio of mobility reduction that each county A would experience if only its *macro-county* went to purple, compared to statewide shutdown. When we use the county partition from California's Regional Stay-At-Home Order, we find that macro-county restrictions can achieve 92.1% (90.9%–93.3%) of statewide mobility reductions. In contrast, if we use a random partitioning of counties into equal-sized segments, such restrictions only reach 62.3% (54.3%–70.4%, 95% CI includes randomness in partitioning) of statewide reductions. Thus, policies of intermediate scale are promising in their ability to balance efficacy and flexibility, but achieving that balance relies on optimizing how macro-counties are defined.

Given a desired number of macro-counties k, we show in the Appendix that we can find the optimal county partition that maximizes the average $r_M(A)$ over counties by solving a minimum k-cut problem, which seeks to partition the nodes of an undirected graph into k disjoint sets while minimizing the total weight of edges between nodes in different sets. We define our undirected graph as one between counties, where the edge weight w_{AB} between two adjacent

counties A and B is

$$w_{AB} = \frac{\mathbb{E}[Y_{AB}|P,R] - \mathbb{E}[Y_{AB}|P,P]}{\mathbb{E}[out(A)|\mathbf{T}_R] - \mathbb{E}[out(A)|\mathbf{T}_P]} + \frac{\mathbb{E}[Y_{BA}|P,R] - \mathbb{E}[Y_{BA}|P,P]}{\mathbb{E}[out(B)|\mathbf{T}_R] - \mathbb{E}[out(B)|\mathbf{T}_P]}. \quad (11)$$

To achieve evenly sized macro-counties, we impose an additional constraint (common in balanced graph partitioning) that each set is no larger than $1.05 \cdot \frac{N}{k}$, where $N = 58$ is the total number of counties. While this problem is NP-hard, we can approximate the solution using METIS (Karypis and Kumar 1997). In Figure 4b, we display our solutions over a range of k. Smaller macro-county sizes are preferred for flexibility (x-axis), while higher $r_M(A)$ represents better efficacy (y-axis). We observe a clear trade-off between the two objectives; however, even small macro-counties—when grouped intelligently—yield large improvements in efficacy over county-level restrictions. For example, by just increasing the average macro-county size to 8 (still $1/7^{th}$ the total number of counties), we reach over 90% of the full efficacy of the much more drastic statewide shutdown (Figure 4c).

6 Conclusion

Geographic spillovers arise in many domains, but there are few opportunities to reliably estimate spillover effects. In this work, we identify a novel setting that is uniquely suitable for spillover analysis, California's Blueprint for a Safer Economy, which defined a set of policies applied with the same deterministic thresholds across 58 counties. We leverage these properties to develop a causal inference framework that allows us to make unconfounded estimates of spillover movement between counties and we observe significant spillovers in many POI groups. Finally, we evaluate the cost of spillovers on policies across spatial scales, analyzing the trade-off between efficacy and flexibility.

Our work is not without limitations. First, SafeGraph's data does not cover all POIs or populations uniformly. To mitigate this issue, we control for CBG weekly device count, only estimate effects for the largest POI categories, and drop categories such as elementary schools that have unreliable coverage from cell phone apps. Second, our causal inference framework may not entirely satisfy SUTVA, the assumption that a unit's outcome is only influenced by its own treatment. In this work, we attempt to better satisfy SUTVA by modeling the effect of pairwise policies on cross-county movement, instead of only modeling the effect of a single county's policies on its population's mobility, as prior work has done. However, future work should explore interference beyond pairs; for example, mobility from county A to B may depend not only on A and B's policies but also on the policies of A's other neighbors. We also hope that future work will dive deeper into the complex trade-offs of policymaking for interconnected regions. In this work, we explored efficacy and flexibility, but other dimensions should be considered, such as equity in the context of certain regions bearing disproportionate risks and unequal resources (e.g., with resourced areas better able to handle spikes in COVID-19 cases).

Acknowledgements

S. C. was supported in part by an NSF Graduate Research Fellowship and the Meta PhD Fellowship. The authors thank Emma Pierson, Martin Saveski, Hamed Nilforoshan, and anonymous reviewers for helpful comments and discussions.

References

Ali, S.; Saeed, M. H.; Aldreabi, E.; Blackburn, J.; De Cristofaro, E.; Zannettou, S.; and Stringhini, G. 2021. Understanding the Effect of Deplatforming on Social Networks. In *Proceedings of the 13th ACM Web Science Conference*.

Badr, H.; et al. 2020. Association between mobility patterns and COVID-19 transmission in the USA: a mathematical modelling study. *The Lancet Infectious Diseases*.

Brauner, J. M.; Mindermann, S.; Sharma, M.; Johnston, D.; Salvatier, J.; et al. 2020. Inferring the effectiveness of government interventions against COVID-19. *Science*, 371(6531).

Bronars, S. G.; and Lott, J. R. 1998. Criminal Deterrence, Geographic Spillovers, and the Right to Carry Concealed Handguns. *The American Economic Review*, 88(2): 475–479.

Buckee, C. O.; et al. 2020. Aggregated mobility data could help fight COVID-19. *Science*, 368(6487): 145.

Cattaneo, M. D.; Idrobo, N.; and Titiunik, R. 2020. *A Practical Introduction to Regression Discontinuity Designs: Foundations*. Elements in Quantitative and Computational Methods for the Social Sciences. Cambridge University Press.

CDPH. 2020. Regional Stay at Home Order. Available at https://www.cdph.ca.gov/Programs/CID/DCDC/Pages/COVID-19/Regional-Stay-at-Home-Order-.aspx.

CDPH. 2021a. Blueprint for a Safer Economy. Available at https://www.cdph.ca.gov/Programs/CID/DCDC/Pages/COVID-19/COVID19CountyMonitoringOverview.aspx.

CDPH. 2021b. Blueprint For a Safer Economy: Equity Focus. Available at https://www.cdph.ca.gov/Programs/CID/DCDC/Pages/COVID-19/CaliforniaHealthEquityMetric.aspx.

Chandrasekhar, A. G.; Goldsmith-Pinkham, P.; Jackson, M. O.; and Thau, S. 2021. Interacting regional policies in containing a disease. *PNAS*, 118(19).

Chang, S.; Pierson, E.; Koh, P.; et al. 2020. Mobility network models of COVID-19 explain inequities and inform reopening. *Nature*, 589(82–87).

Chang, S.; Wilson, M. L.; Lewis, B.; Mehrab, Z.; Dudakiya, K. K.; Pierson, E.; Koh, P. W.; Gerardin, J.; Redbird, B.; Grusky, D.; et al. 2021. Supporting COVID-19 Policy Response with Large-scale Mobility-based Modeling. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*.

Chernozhukov, V.; Kasahara, H.; and Schrimpf, P. 2021. Causal impact of masks, policies, behavior on early covid-19 pandemic in the U.S. *Journal of Econometrics*, 220(1).

Chinazzi, M.; et al. 2020. The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. *Science*, 368(6489): 395–400.

Coates, M.; and Pearson-Merkowitz, S. 2017. Policy Spillover and Gun Migration: The Interstate Dynamics of State Gun Control Policies. *Social Science Quarterly*, 98: 500–512.

Hao, Z.; and Cowan, B. 2017. The Cross-Border Spillover Effects of Recreational Marijuana Legalization. *NBER Working Papers*, 23426.

Holtz, D.; Zhao, M.; Benzell, S. G.; Cao, C. Y.; Rahimian, M. A.; et al. 2020. Interdependence and the cost of uncoordinated responses to COVID-19. *PNAS*, 117(33).

Ibarra, A. B.; and Becker, R. 2021. California hit critical milestone for reopening today. *Cal Matters*. Available at https://calmatters.org/health/coronavirus/2021/03/california-milestone-reopening-today/.

Imbens, G. W.; and Lemieux, T. 2008. Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, 142(2): 615–635.

Karypis, G.; and Kumar, V. 1997. METIS: A Software Package for Partitioning Unstructured Graphs, Partitioning Meshes, and Computing Fill-Reducing Orderings of Sparse Matrices. In *University of Minnesota Computer Science & Engineering Technical Reports*.

Kraemer, M. U. G.; Yang, C.-H.; Gutierrez, B.; Wu, C.-H.; Klein, B.; Pigott, D. M.; Group, O. C.-. D. W.; du Plessis, L.; Faria, N. R.; Li, R.; et al. 2020. The effect of human mobility and control measures on the COVID-19 epidemic in China. *Science*, 368(6490): 493–497.

Liu, L.-P.; and Blei, D. 2017. Zero-Inflated Exponential Family Embeddings. In *Proceedings of the 34th International Conference on Machine Learning*.

Newsom, G. 2020. California Health Officials Announce a Regional Stay at Home Order Triggered by ICU Capacity. Available at https://www.gov.ca.gov/2020/12/03/california-health-officials-announce-a-regional-stay-at-home-order-triggered-by-icu-capacity/.

Nguyen, T. D.; Gupta, S.; Andersen, M.; Bento, A.; Simon, K. I.; and Wing, C. 2020. Impacts of State Reopening Policy on Human Mobility. *NBER Working Paper*, 27235.

Nouvellet, P.; Bhatia, S.; Cori, A.; Ainslie, K. E. C.; Baguelin, M.; Bhatt, S.; Boonyasiri, A.; Brazeau, N. F.; Cattarino, L.; Cooper, L. V.; et al. 2021. Reduction in mobility and COVID-19 transmission. *Nature Communications*, 12(1090).

Papay, J. P.; Willett, J. B.; and Murnane, R. J. 2011. Extending the regression-discontinuity approach to multiple assignment variables. *Journal of Econometrics*, 161(2): 203–207.

Ribeiro, M. H.; Jhaver, S.; Zannettou, S.; Blackburn, J.; Stringhini, G.; De Cristofaro, E.; and West, R. 2021. Do Platform Migrations Compromise Content Moderation? Evidence from r/The_Donald and r/Incels. In *Proceedings of the ACM on Human-Computer Interaction*, volume 5, CSCW2.

Sigman, H. 2005. Transboundary spillovers and decentralization of environmental policies. *Journal of Environmental Economics and Management*, 50(1): 82–101.

Wong, V. C.; Steiner, P. M.; and Cook, T. D. 2013. Analyzing Regression-Discontinuity Designs With Multiple Assignment Variables: A Comparative Study of Four Estimation Methods. *Journal of Educational and Behavioral Statistics*, 38(2): 107–141.

Zhao, M.; Holtz, D.; and Aral, S. 2021. Interdependent program evaluation: Geographic and social spillovers in COVID-19 closures and reopenings in the United States. *Science Advances*, 7(31).

A1 Details on Data and Model Fitting

Data and code availability. The code to run our experiments and regenerate figures is available online.[6] We also make our constructed Z variables available, to facilitate future research that uses them in regression discontinuity designs to estimate the effects of California Blueprint tiers on spillovers and other outcomes of interest.

Documentation about the California Blueprint for a Safer Economy is provided by the California Department of Public Health (CDPH), such as how tiers were assigned (CDPH 2021a,b) and what the tier restrictions were for different sectors (Table A1). CDPH has also archived historical tier assignments and COVID-19 metrics per county over the course of the Blueprint.[7] Our mobility data comes from SafeGraph Weekly Patterns[8], which is available to researchers through Dewey.[9] SafeGraph also provides each POI's "top" category (first 4 digits of NAICS code) and "sub" category (first 6 digits), which we use to learn heterogeneous effects for different POI groups (Table A2). Finally, we use data from the US Census Bureau's 5-year American Community Survey about census block groups, which is available online.[10]

Bandwidth selection. As we describe in Section 4, we filter the data based on a number of criteria, including that Z_{iw} for CBG c_i's county and Z_{jw} for POI p_j's county both lie within a bandwidth h of 0 (since $Z = 0$ is the threshold between assignment to the purple tier and red tier). Bandwidth selection introduces a bias-variance trade-off, with larger bandwidths corresponding to greater bias but reduced variance. We err on the side of larger bandwidths, so that we retain enough representation from different county pairs for each of the treatment conditions. When we use $h = 5$ in our experiments and apply all our other filtering criteria, we are left with 1,408,317,656 data points overall (Table A3).

684,604 of those data points represent non-zero visits, and among those, 289,199 belong to the PP treatment condition, 48,608 to PR, 37,243 to RP, and 309,554 to RR. In Table A4, we list, for each treatment condition, the number

[6] https://github.com/snap-stanford/covid-spillovers

[7] https://www.cdph.ca.gov/Programs/CID/DCDC/Pages/COVID-19/CaliforniaBlueprintDataCharts.aspx

[8] https://docs.safegraph.com/docs/weekly-patterns

[9] https://www.deweydata.io/

[10] https://www.census.gov/programs-surveys/acs/data.html

Sector	Purple Tier	Red Tier
Critical Infrastructure (e.g., hospitals, emergency services, grocery stores, gas stations)	Open with modifications	Open with modifications
Limited Services (provides services with limited contact, e.g, laundry services, auto repair shops, pet grooming)	Open with modifications	Open with modifications
Outdoor playgrounds & recreational facilities	Open with modifications	Open with modifications
Hotels and Lodging	Open with modifications	Open with modifications
Hair salons & barbershops	Open indoors with modifications	Open indoors with modifications
Personal Care Services	Open indoors with modifications	Open indoors with modifications
All Retail	Open indoors with modifications, max 25% capacity	Open indoors with modifications, max 50% capacity
Shopping Centers (Malls, Destination Centers, Swap Meets)	Open indoors with modifications, max 25% capacity	Open indoors with modifications, max 50% capacity
Museums, Zoos, and Aquariums	Outdoor only with modifications	Open indoors with modifications, max 25% capacity
Places of Worship	Outdoor encouraged; indoors allowed with modifications, max 25% capacity	Open indoors with modifications, max 25% capacity
Movie Theaters	Outdoor only with modifications	Open indoors with modifications, max 25% capacity or 100 people (whichever is fewer)
Restaurants	Outdoor only with modifications	Open indoors with modifications, max 25% capacity or 100 people (whichever is fewer)
Gyms and Fitness Centers	Outdoor only with modifications	Open indoors with modifications, max 10% capacity

Table A1: The California Blueprint for a Safer Economy's section-specific restrictions for the purple and red tiers. For the full list of tiers and sectors, see CDPH (2021a), "Risk Criteria".

NAICS code	Full Name	Description	# POIs
4411	Automobile Dealers	New car and old car dealers	1429
4413	Automotive Parts, Accessories, and Tire Stores	Retailers for automotive parts and repair	2034
4441	Building Material and Supplies Dealers	Retailers for home improvement goods, paint and wallpaper, tools and builders' hardware	1146
4451	Grocery Stores	Supermarkets, convenience retailers, vending machine operators	6449
4551	Department Stores	Department stores for apparel, jewelry, home furnishings, toys, etc.	1138
4552	General Merchandise Stores, including Warehouse Clubs and Supercenters	Warehouse clubs, supercenters, dollar stores, home and auto supply stores, variety stores	2280
4561	Health and Personal Care Stores	Retailers for drugs (i.e., pharmacies), beauty supplies, optical goods, food supplements	4227
4571	Gasoline Stations	Gasoline stations, sometimes with convenience stores	7514
4581	Clothing Stores	Sells clothing, clothing accessories (e.g., hats, gloves, wigs)	1656
4591	Sporting Goods, Hobby, and Musical Instrument Stores	Retailers for sporting goods, hobbies, toys, games, sewing and needlework supplies, musical instruments and supplies	3362
4594	Office Supplies, Stationery, and Gift Stores	Retailers for office supplies, stationery, office equipment, greeting cards, decorations	1212
4595	Used Merchandise Stores	Sells used goods, antiques, auctions	1292
4599	Other Miscellaneous Store Retailers	Retailers for pet supplies, art dealers, mobile home dealers, smoking supplies, other miscellaneous things	3697
5311	Lessors of Real Estate	Lessors of real-estate for residential, non-residential (e.g., malls), and storage purposes	3144
7121	Museums, Historical Sites, and Similar Institutions	Museums, historical sites, zoos, gardens, and nature parks	7511
713940	Fitness and Recreational Sports Centers	Sports facilities, including exercise centers, ice or roller skating rinks, tennis club facilities, and swimming pools	4730
7139	Other Amusement and Recreation Industries	Golf courses, skiing facilities, marinas, bowling centers	1834
7211	Traveler Accommodation	Hotels, motels, casino hotels, bed-and-breakfasts	2252
7224	Drinking Places (Alcoholic Beverages)	Bars, taverns, nightclubs	1791
722511	Full-Service Restaurants	Provides food services to patrons who order and are served while seated, then pay after	21972
722513	Limited-Service Restaurants	Provides food services where patrons order and pay before eating; food may be consumed on premises, taken out, or delivered	15074
722515	Snack and Nonalcoholic Beverage Bars	Prepares specialty snacks (e.g., ice cream) or non-alcoholic beverages (e.g., coffee)	8528
8131	Religious Organizations	Churches, religious temples, synagogues, mosques, monasteries	1065
–	Other	All other POIs that were not in one of these groups	20181

Table A2: POI groups for which we learn heterogeneous treatment effects. We keep all POIs in California with at least 30 non-zero visits to CBGs, which leaves 128,655 POIs. Then, as POI groups, we keep all 'top"-categories (first 4 digits of the NAICS code) with at least 1000 POIs (besides "Elementary and Secondary Schools", which we drop due to poor coverage from cell phone data) and 4 "sub"-categories (first 6 digits of NAICS code) of interest. The category descriptions are based on https://www.naics.com/six-digit-naics/.

Week	# Counties in Purple	# Counties in Red	# Data Points
2021-02-01	9	0	2,659,022
2021-02-08	14	1	26,031,894
2021-02-15	20	2	58,585,151
2021-02-22	20	9	64,334,107
2021-03-01	28	16	497,292,295
2021-03-15	10	37	506,971,688
2021-03-22	7	30	204,963,737
2021-03-29	3	31	47,479,762
Total	**111**	**126**	**1,408,317,656**

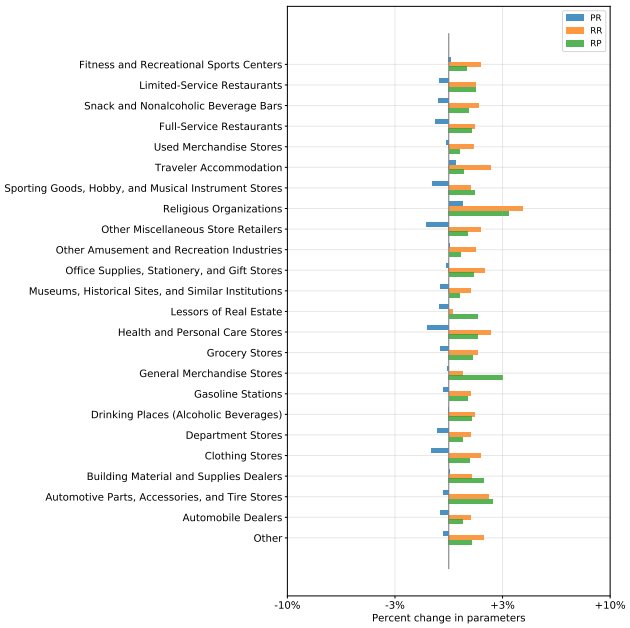
Table A3: Size of the data we keep for fitting our model. We keep counties that are in the purple or red tier, comply with the expected assignment based on Z (its tier is red if and only if $Z < 0$), and its Z variable lies within a bandwidth $h = 5$ of 0. Then, we keep all CBG-POI data points between adjacent kept counties. We also drop the week of March 8, 2021, since the purple-red threshold for adjusted case rate was changed in the middle of the week, due to the statewide vaccine goal being met.

of unique adjacent county pairs that appear for this condition, the top 5 most-represented pairs, and the proportion of all non-zero data points that each pair accounts for within this treatment condition. We see that all treatment conditions, including the less common PR and RP, still retain substantial diversity across counties, with over 70 unique pairs for each condition and no single county or county pair seriously dominating the data. The county pairs that appear more often are, as expected, the ones with a large number of CBGs in the source county and a large number of POIs in the target county.

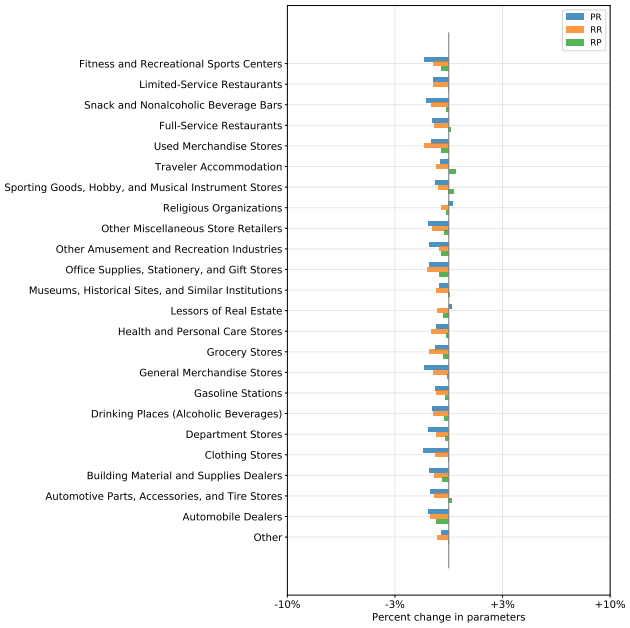
We also conduct sensitivity analyses with $h = 4$ and $h = 6$. Compared to the estimated parameters when $h = 5$, the estimated parameters when $h = 4$ typically only change by 2-3% and at most 5% (Figure A1a). The change is even smaller when we compare $h = 6$ to $h = 5$; the change is at most 2% and mostly smaller than 1% (Figure A1b).

Loss-corrected negative sampling. We perform negative sampling such that we sample each zero data point (i, j, w) with probability s_{ijw}. Then, we fit our model on sample S, which contains all of the non-zero data points from the original data set and our sampled zero data points. However, negative sampling biases our model parameters by greatly reducing the number of zeros in the training data. To correct this bias, we weight each sampled data point by $\frac{1}{s_{ijw}}$ when computing the overall loss (negative log likelihood). These corrections ensure that our stochastic gradient $\nabla_\theta \mathcal{L}_S(\theta)$, computed over sample S, forms an unbiased estimate of the true gradient $\nabla_\theta \mathcal{L}(\theta)$, computed over the full data.

The proof of this is very straightforward, utilizing the fact that the negative log likelihood is a sum over negative log likelihoods per data point, and then applying linearity of expectation. Let indicator variable $b_{ijw} \sim \text{Bern}(s_{ijw})$ repre-



(a) $h = 4$ vs. $h = 5$.



(b) $h = 6$ vs. $h = 5$.

Figure A1: Percent change in estimated spillover parameters for different choices of bandwidth h.

	PP	PR	RP	RR
N	126	71	74	177
1	6067 → 6061, 7.6%	6077 → 6067, 8.9%	6067 → 6077, 10.5%	6067 → 6061, 7.6%
2	6037 → 6059, 5.8%	6067 → 6113, 8.3%	6113 → 6067, 8.3%	6065 → 6071, 7.5%
3	6013 → 6001, 5.1%	6075 → 6081, 6.4%	6017 → 6067, 8.1%	6037 → 6059, 6.0%
4	6059 → 6037, 4.0%	6029 → 6037, 5.9%	6107 → 6019, 7.5%	6071 → 6065, 5.6%
5	6061 → 6067, 3.7%	6115 → 6101, 5.9%	6081 → 6075, 5.3%	6065 → 6059, 4.9%

Table A4: Distribution of top 5 most-represented adjacent county pairs per treatment condition. N represents the number of unique county pairs seen per treatment condition.

sent whether data point (i, j, w) is in our sample S:

$$\nabla_\theta \mathcal{L}_S(\theta) = -\sum_{i,j,w \in C} b_{ijw} \frac{1}{s_{ijw}} \nabla_\theta \ln(\Pr(Y_{ijw}|\theta)) \tag{12}$$

$$\mathbb{E}[\nabla_\theta \mathcal{L}_S(\theta)] = -\sum_{i,j,w \in C} \mathbb{E}[b_{ijw}] \frac{1}{s_{ijw}} \nabla_\theta \ln(\Pr(Y_{ijw}|\theta)) \tag{13}$$

$$= -\sum_{i,j,w \in C} \nabla_\theta \ln(\Pr(Y_{ijw}|\theta)) = \nabla_\theta \mathcal{L}(\theta). \tag{14}$$

Thus, incorporating a correction $\frac{1}{s_{ijw}}$ into the loss per data point ensures that the stochastic gradient forms an unbiased estimate of the true gradient, *regardless of the negative sampling scheme used*.

However, different negative sampling schemes, i.e., different choices of s_{ijw}, may be preferable in order to reduce the variance of the stochastic gradient. First, note that the variance of the stochastic gradient is the sum of the variances per data point, since each indicator variable b_{ijw} is sampled independently. Second, observe that the non-zero data points contribute no variance, since they are sampled with probability 1. So, the variance of the stochastic gradient is a sum of variances over the zero data points, which we refer to as C_0:

$$\text{Var}[\nabla_\theta \mathcal{L}_S(\theta)] = \sum_{i,j,w \in C_0} \text{Var}[b_{ijw} \frac{1}{s_{ijw}} \nabla_\theta \ln(\Pr(Y_{ijw} = 0|\theta))] \tag{15}$$

$$= \sum_{i,j,w \in C_0} s_{ijw}(1 - s_{ijw})(\frac{1}{s_{ijw}} \nabla_\theta \ln(\Pr(Y_{ijw} = 0|\theta)))^2 \tag{16}$$

$$= \sum_{i,j,w \in C_0} (\frac{1}{s_{ijw}} - 1)(\nabla_\theta \ln(\Pr(Y_{ijw} = 0|\theta)))^2. \tag{17}$$

Since s_{ijw} is a probability, then $\frac{1}{s_{ijw}} \geq 1$. We can see, firstly, that larger sampling probabilities reduce the variance, and when $s_{ijw} = 1$ for all i, j, w (meaning we sample all zero data points with probability 1), the variance is 0. Barring an increase in sample size, observe that we would want to prioritize sampling "harder" data points, i.e., those with larger gradients, where the model is learning more from $Y_{ijw} = 0$. While we cannot know before sampling which data points would have larger gradients, we can use proxies, such as distance between the CBG and POI, to estimate harder samples. Thus, we set $s_{ijw} \propto \frac{1}{1+d_{ij}}$, where d_{ij} is the distance between CBG c_i and POI p_j. Then, we scale the s_{ijw} terms such that in expectation, we sample a certain percentage, such as 2%, of the zero data points.

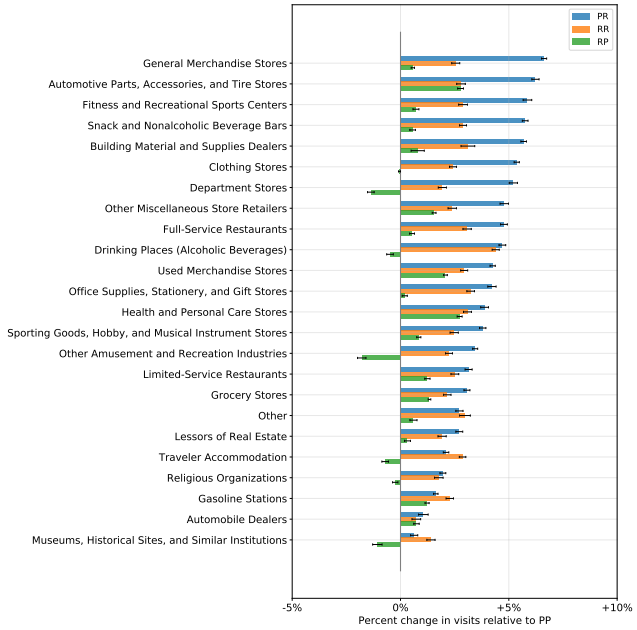
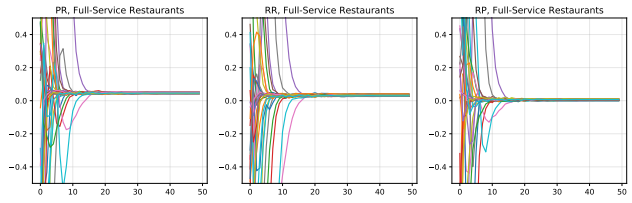
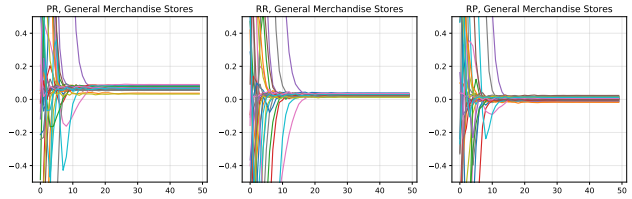


Figure A2: Estimated spillover effects across POI groups, with 95% confidence intervals only capturing uncertainty over negative samples and not over the data.

Recall that in our bootstrapping procedure, we sample the non-zero data points with replacement and draw a fresh negative sample in every trial. This procedure allows us to capture uncertainty in the data as well as uncertainty from negative sampling. We conduct an additional experiment where we only capture uncertainty from negative sampling, by conducting 10 more trials with fresh negative samples but without sampling the non-zero data points. We use the same negative sampling scheme as we do in our main experiments, sampling 2% of zero data points with distance-weighted sampling probabilities. Compared to our main results (Figure 3), we show in Figure A2 that negative sampling only accounts for a very small proportion of the overall uncertainty,



(a) Full-Service Restaurants.



(b) General Merchandise Stores.

Figure A3: Model parameters converge over epochs. We plot one line for each of the 30 trials. The example model parameters are the spillover weights β_{PR}, β_{RR}, and β_{RP} for two POI groups.

confirming that our negative sampling scheme is sufficient.

Model fitting. As described in Section 4, we quantify uncertainty with 30 trials. In each trial, we re-sample the data and fit the model using loss-corrected gradient descent, running for 50 epochs. Empirically, the model parameters reliably converge within this number of epochs and, across trials, parameters also converge to similar values (Figure A3).

A2 Details For Computing Results

Local vs. global restrictions. In this section, we discuss in more detail how we conduct our analysis of county-level vs. hypothetical statewide restrictions. First, let $\mathbf{T} \in \mathbb{R}^{58}$ represent a treatment vector for all 58 counties. Then, let $\mathbf{T}_R$ represent the all-red treatment condition, $\mathbf{T}_P$ represent the all-purple treatment condition, and $\mathbf{T}_A$ represent the condition where all counties are in red except county A, which is in purple. Recall that our goal is to compare the reduction in county A's expected out-degree, $\mathbb{E}[out(A)]$, from $\mathbf{T}_R$ to $\mathbf{T}_A$, compared to $\mathbf{T}_P$:

$$r(A) = \frac{\mathbb{E}[out(A)|\mathbf{T}_R] - \mathbb{E}[out(A)|\mathbf{T}_A]}{\mathbb{E}[out(A)|\mathbf{T}_R] - \mathbb{E}[out(A)|\mathbf{T}_P]}. \quad (18)$$

We calculate $\mathbb{E}[out(A)|\mathbf{T}]$ as the sum over within-county visits and out-of-county visits:

$$\mathbb{E}[out(A)|\mathbf{T}] = \mathbb{E}[Y_{AA}|T_A] + \sum_{B \in N(A)} \mathbb{E}[Y_{AB}|T_A, T_B], \quad (19)$$

where Y_{AB} represents the total number of visits from any CBG in A to any POI in B. When we use our fitted model to compute the conditional expectation of Y_{AB} given tiers, we assume $Z = 0$ for all CBGs and POIs, since our RD-based framework estimated tier effects at the joint cutoff. We also marginalize over the remaining dynamic covariate, the CBG's weekly device count, by taking each CBG's average device count over the 9-week period that we study. This produces a static vector $\mathbf{X}_{ij}$, representing the CBG and POI covariates.

Using our zero-inflated Poisson regression model, we compute $\mathbb{E}[Y_{AB}]$ as the sum over visits from each CBG in A to each POI in B, where we use POI group-specific weights β_{g_j,T_A,T_B} to capture heterogeneous tier effects:

$$\mathbb{E}[Y_{AB}|T_A, T_B] = \sum_{i \in A} \sum_{j \in B} \mathbb{E}[Y_{ij}|T_A, T_B] \quad (20)$$

$$= \sum_{i \in A} \sum_{j \in B} \frac{1}{1 + \alpha_1 d_{ij}^{\alpha_2}} \exp(\beta_0 + \beta_3^T \mathbf{X}_{ij} + \beta_{g_j,T_A,T_B}) \quad (21)$$

$$= \sum_g \exp(\beta_{g,T_A,T_B}) \underbrace{\sum_i \sum_{j;g_j=g} \frac{1}{1 + \alpha_1 d_{ij}^{\alpha_2}} \exp(\beta_0 + \beta_3^T \mathbf{X}_{ij})}_{\phi(g,A,B)}. \quad (22)$$

We can simplify computation by pre-computing weights $\phi(g, A, B)$ for each adjacent county pair and POI group g. Then, the numerator of $r(A)$ becomes

$$\mathbb{E}[out(A)|\mathbf{T}_R] - \mathbb{E}[out(A)|\mathbf{T}_A] \quad (23)$$

$$= \mathbb{E}[Y_{AA}|R] - \mathbb{E}[Y_{AA}|P] + \sum_{B \in N(A)} \mathbb{E}[Y_{AB}|R, R] - \mathbb{E}[Y_{AB}|P, R] \quad (24)$$

$$= \mathbb{E}[Y_{AA}|R] - \mathbb{E}[Y_{AA}|P] + \sum_{g \in G} \exp(\beta_{g,R,R}) - \exp(\beta_{g,P,R}) \sum_{B \in N(A)} \phi(g, A, B). \quad (25)$$

Similarly, the denominator is

$$\mathbb{E}[out(A)|\mathbf{T}_R] - \mathbb{E}[out(A)|\mathbf{T}_P] \quad (26)$$

$$= \mathbb{E}[Y_{AA}|R] - \mathbb{E}[Y_{AA}|P] + \sum_{g \in G} \exp(\beta_{g,R,R}) - \exp(\beta_{g,P,P}) \sum_{B \in N(A)} \phi(g, A, B). \quad (27)$$

We fit a separate model on only within-county visits to calculate $\mathbb{E}[Y_{AA}]$ and use it to estimate the change in within-county visits from the red to purple tier. With this analysis, we estimate that counties applying local restrictions can only achieve, on average, 54.0% (46.4%–61.7%) of the reduction in mobility that they would experience under a statewide shutdown. We also compute the averages over only the 23 small counties (population under 106,000), which yields 41.7% (31.0%–52.4%), and over the 35 large counties, which yields 62.1% (56.3%–67.9%).

We also analyze a less extreme setting, where instead of $\mathbf{T}_A$, where all counties are in red except county A, we consider the actual tier configuration from the California Blueprint. First, we take the real assignments from a week w in our study period, such as March 15, 2021, when 11 counties were in the purple tier, 42 were in the red tier, 4 were in the orange tier, and 1 was in the yellow tier (Figure 1a). We construct a new treatment vector $\mathbf{T}_w'$, where a county is

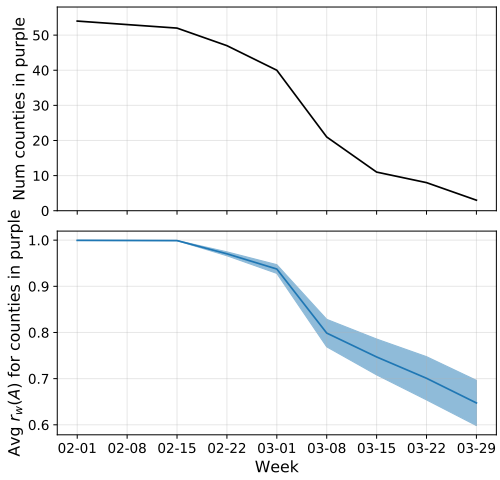


Figure A4: Evaluating the efficacy of realistic tier configurations. Over the course of our 9-week study period, we visualize the number of counties kept in the purple tier based on the real Blueprint tier assignments from that week (top) and the average percentage of mobility reduction kept for those counties in purple, with 95% confidence intervals (bottom).

assigned to the purple tier if it was in the purple tier in week w and assigned to the red tier, otherwise. Then, we compute a similar ratio $r_w(A)$:

$$r_w(A) = \frac{\mathbb{E}[out(A)|\mathbf{T}_R] - \mathbb{E}[out(A)|\mathbf{T}'_w]}{\mathbb{E}[out(A)|\mathbf{T}_R] - \mathbb{E}[out(A)|\mathbf{T}_P]}. \tag{28}$$

$r_w(A)$ represents the proportion of mobility reduction kept under this more realistic scenario, compared (as before) to a statewide shutdown. Then, we take the average $r_w(A)$ over the counties that were assigned to purple in week w.

Over the course of our study period, the number of counties in purple fell from 54 to 3. As expected, as the number of counties in purple fell, the amount of mobility reduction retained for the counties still in purple fell as well (Figure A4). For example, by the week of March 15, 2021, when there were only 11 counties left in the purple tier, those counties only kept 74.7% (70.8%–78.6%) of the statewide mobility reduction. Two weeks later, the remaining 3 counties in the purple tier could only retain 64.7% (59.8%-69.6%) of their statewide reductions. While these percentages are higher than the worst-case (54%, if each county goes to purple alone), they are still far below the full efficacy of the statewide restrictions, demonstrating the cost of spillovers on local policy regimes even under more realistic realizations of policies.

Balancing efficacy and flexibility. To analyze macro-county restrictions, we introduce $\mathbf{T}_{M(A)}$, which represents the treatment condition where all counties are in red except county A's *macro-county*, which is in purple. Then, we define $r_M(A)$ for a county partition M as a a simple extension of $r(A)$:

$$r_M(A) = \frac{\mathbb{E}[out(A)|\mathbf{T}_R] - \mathbb{E}[out(A)|\mathbf{T}_{M(A)}]}{\mathbb{E}[out(A)|\mathbf{T}_R] - \mathbb{E}[out(A)|\mathbf{T}_P]} \tag{29}$$

We can also compute this quantity efficiently using the precomputed weights per county pair and POI group.

To find optimal county partitions, we define an undirected graph G between counties, where the edge weight w_{AB} between two adjacent counties A and B is

$$w_{AB} = \frac{\mathbb{E}[Y_{AB}|P,R] - \mathbb{E}[Y_{AB}|P,P]}{\mathbb{E}[out(A)|\mathbf{T}_R] - \mathbb{E}[out(A)|\mathbf{T}_P]} + \frac{\mathbb{E}[Y_{BA}|P,R] - \mathbb{E}[Y_{BA}|P,P]}{\mathbb{E}[out(B)|\mathbf{T}_R] - \mathbb{E}[out(B)|\mathbf{T}_P]}, \tag{30}$$

and the edge weight between non-adjacent counties is 0. Now, we show that finding the county partition that maximizes the average $r_M(A)$, for a fixed number of macro-counties k, is equivalent to solving a minimum k-cut problem on G. First, observe which parts of $r_M(A)$ actually vary with our choice of M. The denominator is constant and the numerator we can expand out to

$$\mathbb{E}[Y_{AA}|R] - \mathbb{E}[Y_{AA}|P] + \sum_{\substack{B \in N(A); \\ M(A)=M(B)}} \mathbb{E}[Y_{AB}|R,R] - \mathbb{E}[Y_{AB}|P,P] + \sum_{\substack{B \in N(A); \\ M(A)\neq M(B)}} \mathbb{E}[Y_{AB}|R,R] - \mathbb{E}[Y_{AB}|P,R]. \tag{31}$$

The terms for within-county visits, Y_{AA}, are also constant, and the remaining terms are the summations over neighbors of A. At best, all of A's neighbors are in its macro-county, so $M(A) = M(B)$ applies to all neighbors. If we consider moving one neighbor B outside of $M(A)$, this will add c_{BA} to $r_M(A)$:

$$c_{BA} = \frac{\mathbb{E}[Y_{AB}|P,P]) - \mathbb{E}[Y_{AB}|P,R])}{\mathbb{E}[out(A)|\mathbf{T}_R] - \mathbb{E}[out(A)|\mathbf{T}_P]}. \tag{32}$$

This quantity tends to be negative since we showed that $\mathbb{E}[Y_{AB}|P,R]$ is typically larger than $\mathbb{E}[Y_{AB}|P,P]$, due to spillovers. Furthermore, the more spillover there is from A to B, the more negative this quantity will be. Thus, to maximize $r_M(A)$, we want to choose a partition M that maximizes c_{BA} and c_{AB} (i.e., minimizes spillovers) over all pairs of adjacent counties that are *not* in the same macro-county. This is equivalent to finding the minimum k-cut in G, since we defined the edge weight w_{AB} as the sum of $-c_{BA}$ and $-c_{AB}$. In practice, we use the mean model parameters over our 30 trials to construct G. Then, to approximate a solution to the minimum k-cut problem, we run the METIS algorithm[11], with the load imbalance tolerance set to 1.05 to encourage macro-counties of similar sizes.

[11] https://metis.readthedocs.io/en/latest/