Inference with Arbitrary Clustering

Fabrizio Colella[*], Rafael Lalive[*], Seyhun Orcan Sakalli[†], and Mathias Thoenig[*]

March 2, 2020

Abstract

In applied empirical work, statistical inference with spatial or network data is challenging since unobserved heterogeneity can be correlated across neighboring observational units. We develop a novel estimator for the variance-covariance matrix (VCV) in OLS and 2SLS settings that can accommodate, in a flexible way, dependence of the errors across arbitrary clustering structures (in space, in a network), and across time periods. In Monte Carlo simulations that use real data on U.S. metropolitan areas, or on co-authors in Economics, we find that our arbitrary clustering estimator of the VCV yields inference at the correct significance level in moderately sized samples, and it always dominates other commonly used approaches to inference. We provide guidance to the applied practitioners on (i) when the arbitrary clustering correction is necessary; (ii) whether to include potentially correlated control variables; and (iii) how to set the adequate correction bandwidth for the estimator in absence of prior knowledge about the Data Generating Process. Our companion statistical package (`acreg`) enables users to adjust the OLS and 2SLS coefficient's standard errors to account for arbitrary clustering dependence.

JEL: C13, C23, C26,

Keywords: Inference, Arbitrary Clustering, Geospatial Data, Network Data

[*]Department of Economics, HEC University of Lausanne, 1015 Lausanne, Switzerland

[†]King's Business School, King's College London, WC2R-2LS London, United Kingdom

[‡]Our companion statistical package (`acreg`) can be downloaded at the following address https://acregstata.weebly.com. For helpful comments and valuable feedback on early versions of our command, we thank Samuel Bazzi, Nicolas Berman, Richard Bluhm, Johannes Buggle, Mathieu Couttenier, David Drukker, Ruben Durante, Ruben Enikopolov, Elena Esposito, Matthew Jackson, Melanie Krause, and Eleonora Patacchini, as well as participants at the Swiss Meeting of Stata Users (Zurich, 2018) and the Workshop on Geodata and Economics (Braunchweig, 2018).

1 INTRODUCTION

Recent years have witnessed a tremendous surge of empirical studies with data endowed with a topology, such as spatial data or network data. In these data, unobserved shocks can be correlated across neighboring observational units, where the neighborhood refers to the physical space or to the network structure. In both settings, inference is challenging because the sampling structure of the data and of the VCV matrix exhibits overlapping clusters — a feature that is vastly ignored by applied econometricians.[1] Indeed, a common practice with spatial data consists of considering non-overlapping clusters (typically administrative units) defined at a level of aggregation that encompasses the scale of the resolution of the data by several orders of magnitude — e.g., standard errors are clustered at the region level, while observational units typically correspond to $0.5^o \times 0.5^o$ grid cells.[2] In addition to the loss of efficiency when it turns to estimation, such a practice is subject to caution for observational units that are located close to the frontier between two clusters and are likely to be correlated. In the case of network data, the practice is even more rudimentary, as many studies simply do not correct for the potential correlation of unobserved shocks across neighbors.

We develop a novel and flexible approach to obtain reliable inference in spatial and network settings with any type of arbitrary topological and temporal dependence between observational units in large samples. Arbitrary here refers to the way units are correlated with each other in space/network and time. We impose no restrictions so that our approach can be used with a wide range of data. Our estimator for the variance-covariance (VCV) matrix of the estimated parameters builds on the seminal insight by White (1980) who showed that a sandwich-type VCV can be estimated by constructing a consistent estimator of the VCV of the parameters. Specifically, the estimator uses estimated regression errors and knowledge of the clustering structure to estimates the unknown elements of the sandwich formula. In a network setting, the clustering structure is derived from the network structure itself (i.e. links between observational units); in space, our approach follows Conley (1999): A circle around each unit specifies how distance dependence is likely to reach, allowing for decay or not. This type of clustering structure is well known in spatial data, and statistical packages are available online only for ordinary least squares (OLS) estimations (Conley, 1999; Hsiang, 2010).

We foresee three main domains of application of this flexible inference method. The first

[1] Multiway clustering is somewhat more flexible than one-way clustering, allowing errors to correlate, for instance, within units over time and across time periods (Cameron *et al.*, 2011). However, multiway clustering assumes regularity in the clustering structure that may not hold in real-life settings with spatial and network data.

[2] For example, see the GAEZ v3.0 Global Agro-ecological Zones dataset of FAO: http://www.gaez.iiasa.ac.at/

one relates to a clustering structure allowing for spatial and temporal decays with geocoded data. Indeed, empirical work has been fueled by the growing availability of geocoded data and the integration of geographic information systems (GIS) in the toolkit of applied economists. From development and urban economics to economic history, big data at a high level of spatial resolution enable researchers to move the analysis within countries and to craft compelling empirical designs (e.g., RDD, DiD), for the purpose of causal analysis, as various endogeneity concerns are alleviated by exploiting fine-grained variations and discontinuities in the variables of interest.[3] We extend the Conley approach to two-stages least squares (2SLS) estimations. The second type of application relates to all clustering structures that are based on a metric that is not spatial distance (i.e., Euclidean or geodesic) such as a measure of contiguity (e.g. neighboring countries) or a travel distance (flight, road, or walking). More specifically, consider a scholar interested in studying economic outcomes at the county level in the U.S. In such a scenario, it is likely that contiguous counties are affected by common shocks and this should be reflected in the clustering structure. The issue here is that counties have different sizes (much larger in the West; see the map in Figure 3), preventing the researcher from imposing the same spatial kernel (as in Conley (1999)) across the entire sample. The third application relates to network topology. Consider a scholar interested in violence between rebel groups in Africa (König *et al.*, 2017, e.g.). These groups are affected by common shocks not only in the physical space through their location but also in the cultural/social space through their ethnic affiliations. Groups that are ethnically close tend to be affected by similar shocks. Hence, it is important that the clustering structure accounts for ethnic (or genetic or linguistic) relatedness.

In the paper we provide results from extensive Monte Carlo simulations based on real-life data to document our arbitrary clustering approach. A first set of simulations relates to spatial clustering in real data on U.S. metropolitan areas. We construct environments where OLS or 2SLS regressions with robust standard errors clustered at the administrative level reject the null hypothesis of no effect in approximately 10% of all cases when the significance level of the test is set at 5%. Inference using conventional methods does not improve as the sample size increases, suggesting that these methods produce inconsistent estimates of the variance-covariance matrix. By adopting the arbitrary clustering estimator, we find that the null-rejection rate is approximately 8% for small samples (about 150 counties) and approaches the true significance level of 5% in larger samples (from 300 counties on). This pattern suggests that the arbitrary clustering correction produces consistent estimates of the VCV, which dom-

[3] For a survey, see Michalopoulos and Papaioannou (2017).

inate those of other approaches, thereby enabling applied econometricians to conduct robust inference in the presence of spatial correlation. Our second Monte Carlo study deals with network clustering in data of co-authors in Economics from IDEAS RePEc. Here, we again find that applied econometricians adopting conventional inference using robust standard errors that neglect the network correlation in both regressors and outcomes would severely overstate the precision of their estimates. By contrast, inference that allows for arbitrary clustering yields rejection rates close to the correct 5% threshold.

We design several Monte Carlo simulations to provide guidance to the applied researcher on (i) when the arbitrary clustering correction is necessary; (ii) whether to include potentially correlated control variables; and (iii) how to set the adequate correction bandwidth for the estimator in absence of prior knowledge about the Data Generating Process. All our analyses can be readily implemented with our companion statistical package. Our `acreg` command enables Stata users to estimate panel OLS/2SLS models with arbitrary correlation structures, e.g. in space or across a network. We provide a user-friendly introduction to the command and hope this can ensure access to all empirical researchers who are interested in using the new estimator.[4]

This paper is related to several strands in the literature. First, our approach to conducting inference is inspired by White (1980)'s seminal work on consistent estimation of the VCV. Subsequent work by White (1984) and Arellano (1987) developed an estimator, the cluster-corrected estimator (CCE), that allows for robust inference when data are clustered, e.g., in random samples of units observed over multiple time periods. Bertrand *et al.* (2004) discuss how to implement the CCE in a difference-in-differences design. Cameron *et al.* (2011) extended this CCE approach to clustering in multiple dimensions. Recent contributions discuss the performance of the CCE estimators in situations where there are few heterogeneous clusters and propose robust test statistics (Wooldridge, 2003; Ibragimov and Müller, 2016).

Second, a large body of literature on spatial econometrics discusses inference approaches. Conley (1999) develops robust inference in settings where shocks to spatial units are spatially dependent, also allowing for decays. Conley (1999)'s approach extends White (1980)'s main idea to use estimated residuals along with a hypothesis of the correlation structure to construct an estimator of the VCV.[5] Hsiang (2010) further extends Conley (1999)'s approach to panel data

[4] Please see a guideline providing a set of instructions and examples using the following link: https://acregstata.weebly.com/uploads/2/9/1/6/29167217/faq_acreg_01_20.pdf

[5] The different strand in the spatial econometrics literature takes a somewhat different approach. Kelejian and Prucha (1998, 1999) develop estimators in a spatial setting with spatial dependence in both the dependent variable and the regressors.

with HAC decay in temporal dimension and provided a code to the research community. Kelly (2019) criticizes the applied empirical literature for failing to address the complex nature of spatial dependence, perhaps because no guidelines regarding how to implement corrections for spatial correlation was formerly available.[6]

Third, an extensive body of literature examines behavior of individuals as it is shaped by friends or peers in their social networks. Fafchamps and Lund (2003) study whether villagers in the Philippines can self-insure using their social network. Calvó-Armengol *et al.* (2009) assess peer effects in education in the U.S. In both settings, the unexplained parts of behavior, captured by residuals, might be connected along the network. Both studies address network dependence by introducing network-level clusters. Allowing for more flexible correlation patterns, Fafchamps and Gubert (2007) consider clusters that are correlated across all dyads, and Lalive *et al.* (2018) address clustering along adjacent regional passenger rail lines.[7]

We complement this literature by allowing for arbitrary forms of clustering. The recent surge in availability of data with complex spatial or network dependence structures creates unprecedented demand for flexible modeling in order to ensure unbiased inference in complex settings. Our approach can deal with spatial distance, travel distance, travel costs, contiguity, as well as any concept of distance in a network. We allow users to implement instrumental variables (IV) or two-stage least squares (2SLS), procedures that specify outside instruments, a requirement that is very important for applied papers but that seems overlooked or not discussed in the more theory-driven spatial econometrics literature. Finally, our simulations results are based on real data. And we show that the main results from Kelly (2019), who studies inference problems in spatial studies using artificial data, are actually specific to the nature and the structure of his data.

We discuss the econometric background that allows for arbitrary clustering in the next section. Section 3 presents Monte Carlo evidence on the effect of within-cluster correlation on inference for a spatial setting (U.S. counties) and a network setting (coauthors in economics) Section 4 concludes.

[6] A similar literature discusses inference in shift-share designs. A shift-share design combines information on an aggregate shift with local information on shares to build an instrument. This design induces correlation across space trough a highly correlated spatial regressor, the shift-share or Bartik instrument. Adão *et al.* (2019) discuss how to address inference in the shift-share setting.

[7] Bramoullé and Fortin (2010) discuss the econometrics of social networks.

2 A MODEL WITH CROSS-SECTION AND TIME DEPENDENCE

The purpose of this section is to present the estimator of the variance-covariance (VCV) matrix of the parameters that allows for arbitrary clustering. Specifically, we expose the structure of the estimator of the VCV both in situations without endogeneity and with endogeneity. We also discuss that the arbitrary clustering estimator is a straightforward extension of the one-way or multi-way clustering (Cameron *et al.*, 2011). Our core objective is to assess the quality of the inference — in particular, the likelihood of Type 1 error — produced with the estimates of the VCV through Monte Carlo simulations, which we present in Section 3.

Our key focus is on inference with arbitrary dependence of error terms across observations and over time. We have in mind a setting where each observation's error term may depend on other observations' error terms, and this dependence may change with time, or distance in arbitrary ways. Information on the pattern and strength of cross-observation correlations in errors is encoded in pattern matrix that we call S. In the spatial context, S is normally built from information on the geographic distance between spatial units, e.g., regions, cities, and countries. In a social network context, S reflects the direct links of each person, information that is capture in the so-called adjacency matrix. Entries of the S matrix range from 0 to 1, allowing for weights reflecting the strength of each link over time t. We also always include unit self-links in S, so its main diagonal contains ones.

Consider n observations at each t instant of time T from the following linear model:

$$y = X\beta + \epsilon$$

where we observe each unit i several times in different periods t. y is a dependent variable, and X is a matrix of k linearly independent components that could include a long list of dummies for each unit, in case we are interested in the within estimates. We can write the OLS estimator as:

$$b_{OLS} = (X'X)^{-1}X'y$$

and the theoretical VCV of the b_{OLS} is:

$$VCV(b_{OLS}) = (X'X)^{-1}X'\Omega X(X'X)^{-1}$$

where $\Omega \equiv E(\epsilon\epsilon'|X)$ is the unknown VCV of ϵ.

Building on the seminal insight from White (1980) and following the multiway cluster-

robust estimator structure designed by Cameron *et al.* (2011), we propose the following sandwich estimator for the VCV based on the estimated residuals $e \equiv y - Xb_{OLS}$:

$$\widehat{VCV}(b_{OLS}) = (X'X)^{-1}X'(S \times (ee'))X(X'X)^{-1}$$

where S is the pattern matrix, capturing how each observation's error term depends on other observations' error terms, and $\times$ is element-by-element matrix multiplication. The key element of this estimator is the "meat" in the sandwich:

$$X'(S \times (ee'))X = \sum_{i=1}^{n}\sum_{t=1}^{T}\sum_{j=1}^{n}\sum_{s=1}^{T} x_{it}e_{it}e_{js}x'_{js}s_{itjs}$$

where x_{it} is the (column) vector of regressors, and x'_{it} is the row it in matrix X. This estimator of the VCV departs from an estimator that assumes independence across observations in time if both regressors x_{it} *and* residuals e_{it} are correlated across units, or time, or both (Moulton, 1990).

This framework can also be used in situations with endogeneity. We consider the linear two-stage least squares with a number of instruments greater or equal than the number of endogenous regressors: once the endogeneity is taken into account and the causal effect of the explanatory variable on the dependent variable is uncovered through instruments, the procedure to estimate the VCV is qualitatively equivalent to the OLS case.

We consider the same linear model as before, where we add that m of the k components of X are endogenous and a set of $o \geq m$ excluded instruments for a total of $p \leq k$ exogenous variables that form the matrix Z. We can write the first stage as:

$$\hat{X} = (Z'Z)^{-1}Z'XZ$$

Then, the 2SLS estimator is:

$$b_{2SLS} = (\hat{X}'\hat{X})^{-1}\hat{X}'y$$

Under standard regularity conditions b_{2SLS} is asymptotically normal with the following theoretical estimated variance matrix:

$$VCV(b_{2SLS}) = (\hat{X}'\hat{X})^{-1}\hat{X}'\Omega\hat{X}(\hat{X}'\hat{X})^{-1}$$

Moreover, the core part $\hat{X}'\Omega\hat{X}$ can be estimated as before by the following:

$$\hat{X}'(S \times (uu'))\hat{X} = \sum_{i=1}^{n}\sum_{t=1}^{T}\sum_{j=1}^{n}\sum_{s=1}^{T}\hat{x}_{it}u_{it}u_{js}\hat{x}'_{js}s_{itjs} \tag{1}$$

where the estimated residuals now refer to the 2SLS estimator: $u \equiv y - Xb_{2SLS}$.

Arbitrary clustering extends one-way clustering and multi-way clustering by allowing for a flexible structure of the pattern matrix S. Consider a situation where a researcher studies earnings in counties, nested in states. In this context, there is reason to suspect that counties located in the same state may be affected by common shocks. One-way clustering specifies a pattern matrix S that is block diagonal with element in row it and column js equal to one if i and j are located in the same state, and equal to zero otherwise.

The multi-way clustering approach provides for more flexibility (Cameron *et al.*, 2011). The peculiarity of the multi-way clustering environment is the presence of several dimensions of clustering with non-overlapping clusters in each dimension, and each observation belonging to cluster in each dimension. Specifically, observations may share a cluster called state, and a second cluster called time, thereby allowing for correlation within state and year. With multi-way clustering, the row it and column js element of the S matrix is equal to one if observation it and observation js refer to counties located in the same state, or observed in the same year, and equal to zero otherwise. Multiway clustering is more flexible than one-way clustering but it has some drawbacks. Multi-way clustering imposes that if unit i is assumed to be correlated with unit j and l, then unit j and l must also be correlated. In addition, if the unit i is assumed to depend on unit j at time t, then they must also be dependent at time s. In many real-life settings, like the ones we present in the next section, this particular clustering structure does not hold.

Arbitrary clustering further relaxes assumptions on the shape of the pattern matrix. Units can be correlated among them in any possible way, without any kind of imposed structure. Simply, the row it and column js-th component of the matrix S can be zero, one or any other number in between, depending on the imposed strength of the dependence between errors of observations i and j. The flexibility of our structure allows for both cross-sectional and time dependence. In addition, it allows for changes in the strength of the correlation generated by alterations in the link structure over time or any kind of decay between two points in time. The flexibility offered by arbitrary clustering is an asset in many empirical applications. We next discuss two classes of applications in which only flexible arbitrary clustering offers reliable statistical inference.

3 SIMULATION STUDY

In this section we conduct various Monte Carlo experiments to illustrate how spatial or network correlation between neighboring observational units affects the quality of statistical inference. We focus on the likelihood of Type 1 errors and measure the quality of inference as the extent to which the rejection rate of the null hypothesis of no effect (of a random shock) approaches the nominal rejection rate of the test. We compare several procedures of correcting standard errors across different data environments. We show that the arbitrary clustering estimator provides better inference than estimators that correct for heteroscedasticity or clusters, two standard procedures commonly used in the applied literature. We now briefly describe the overarching structure of our simulation study and get into the details in the next sub-sections.

Our approach is pretty standard and draws on Bertrand *et al.* (2004) and Cameron *et al.* (2011). In all experiments, the Data Generating Process (DGP) starts by retrieving an outcome variable and covariates from real-life data, i.e., geocoded data on U.S. counties for the spatial setting and co-authorships from IDEAS RePEc for the network setting. Then, in each Monte Carlo draw, we generate fake policy/productivity shocks, $\texttt{shock}_i$, that are randomly assigned to some of the observational units i.[8] Depending on the experiment, the DGP is engineered such as to generate an environment where (i) the shocks correlate (or not) between neighboring units; (ii) where the shocks are endogenous, namely they spuriously correlate with the outcome variable under consideration.

Equipped with the generated data, we then regress with OLS or 2SLS the outcome variable on the fake policy/productivity shocks using alternative options for correcting standard errors (heteroskedasticity robust, non-overlapping clusters, arbitrary clusters):

$$Y_i = \alpha + \beta \texttt{shock}_i + X_i'\gamma + \epsilon_i \tag{2}$$

Then, pooling together all Monte Carlo draws, we compare across the estimators the average rejection rate of the null hypothesis of no effect ($\beta = 0$) at the 5% significance level. With an appropriate estimator and for a sufficiently large sample size, the rejection rate is expected to converge to 5% as the number of Monte Carlo draws increases.

[8] We refer to this variable as `policy` in the spatial environment and `productivity` in the network environment.

3.1 Spatial Setting

We now present in more detail the Monte Carlo experiments in a spatial setting. We first explain how the DGP is designed. Then, we report and comment the results. And finally, we conclude with a list of recommendations for the applied practitioner.

3.1.1 The Data Generating Process in the Spatial Setting

We extract tabular information on median earnings, education level, age, race, and gender aggregated at the county level for 2000 from the National Historical Geographic Information System (NHGIS) database (Manson *et al.*, 2017). The NHGIS is a part of the Integrated Public Use Microdata Series (IPUMS) project of the University of Minnesota and provides tabular U.S. Census data and GIS boundary files. Our observational units i consist in the 3,141 counties that are covered in our sample.

When considering equation (2), the outcome variable Y_i is the natural log of median earnings in county i in 2000 and X_i is a vector of county-level controls, which comprises the share of population with tertiary education, share of females, share of blacks, median age and its square, and natural log of total population in 2000. In this setting the random shock $\texttt{shock}_i$ is equal to a binary variable $\texttt{policy}_i$ indicating whether county i receives a policy shock. We now explain the various ways of generating this policy shock.

Scenario 1: No Spatial Correlation. In the most simple experiment, in each Monte Carlo iteration, we draw an iid random variable, $\tilde{u}_i$, from a standardized normal distribution (with cdf Φ), for each county. Then, we select the counties that are in the top quarter of the distribution of this random variable as counties that receive a fake policy shock: $\texttt{IIDpolicy}_i = 1$ iff $\Phi(\tilde{u}_i) > 0.75$ (and 0 otherwise). Panel (a) of Figure 1 visualizes an example of the distribution of the `IIDpolicy` variable drawn at random.

Scenario 2: Spatial Correlation. Next, we impose spatial correlation between neighboring counties when generating the policy shock. Specifically, we compute bilateral distances between counties' centroids. Then, we define a distance cutoff and construct an indicator variable, h_{ij}, that codes for the spatial cluster of each county: $h_{ij} = 1$ if the distance between county i and county j is lower than the cutoff and $h_{ij} = 0$ otherwise. We set the cutoff at 56 kms in our baseline analysis in order to have on average five counties in spatial clusters.[9] For each

[9] Note that we adopt a uniform spatial decay kernel in our simulations. We have explored Bartlett-type kernels as well and find that results are fairly comparable to those we present here.

county i, we then compute the share of neighboring counties within its spatial cluster that are affected by the policy shocks, $\texttt{SCshare}_i \equiv \sum_j h_{ij}\texttt{IIDpolicy}_j / \sum_j h_{ij}$. We define the spatially correlated policy shock as the sum of the idiosyncratic policy shock and the share of the neighboring counties that are also affected by the shock: $\texttt{SCpolicy}_i = \texttt{IIDpolicy}_i + \texttt{SCshare}_i$.

Panel (b) of Figure 1 visualizes the distribution of SCpolicy built on the underlying policy shocks IIDpolicy displayed in panel (a). While we observe no spatial pattern in panel (a), the panel (b) is marked with spatial correlation across counties that are in close proximity. As will be discussed in-depth in section 3.1.4, spatial correlation is susceptible to deteriorate the quality of inference only when both the explanatory variable (SCpolicy) and the outcome variable (log median earning) are spatially correlated. This is the case in this setting as shown by the spatial distribution of log median earnings in 2000 at the county level in Figure 2: We clearly see a pattern of spatial correlation between close counties.

ENDOGENEITY. We design an environment with endogeneity by generating fake policy shocks that are correlated with the outcome variable: $\texttt{ENDpolicy} \in \{0,1\}$. We do so by forcing the counties that receive a policy shock to be randomly drawn only among the subsample of wealthy counties, all poor counties getting no shock. Importantly, we rely on the same random variable $\tilde{u}_i$ we use to define the iid policy shock $\texttt{IIDpolicy}_i$.[10] Thus, $\texttt{IIDpolicy}_i$ is a valid instrument of $\texttt{ENDpolicy}_i$ when estimating a 2SLS version of the econometric model (2). Note that the exclusion restriction is also met by construction because the instrument IIDpolicy is uncorrelated with the outcome variable. Panel (a) of Figure 3 depicts an example of the spatial distribution of a randomly drawn ENDpolicy. Visual inspection confirms that it correlates with the county-level distribution of log median earnings in 2000 as depicted in Figure 2. We then factor in spatial correlation also into this endogenous environment by generating a spatially correlated endogenous policy shock, SCENDpolicy, in the same way we generate spatially correlated exogenous policy shocks in scenario 2.[11]

3.1.2 RESULTS

We run Monte Carlo simulations with 10,000 iterations. In each Monte Carlo draw, the DGP produces the random policy variables according to the two aforementioned scenarios. Us-

[10] More precisely, we select as $\texttt{ENDpolicy}_i = 1$ the counties that are in the top half of the distribution of $\tilde{u}_i$ *conditional* on being a wealthy county (i.e. above the median log earnings). In this way, both IIDpolicy and ENDpolicy take the value of 1 exactly for one quarter of the counties. Formally we set: $\texttt{ENDpolicy}_i = 1$ iff $\Phi^W(\tilde{u}_i) > 0.5$ where Φ^W is the cdf of the standardized normal distribution conditional on being wealthy.

[11] Formally, we define $\texttt{SCENDpolicy}_i \equiv \texttt{ENDpolicy}_i + \texttt{SCENDshare}_i$ where $\texttt{SCENDshare}_i \equiv \sum_j h_{ij}\texttt{ENDpolicy}_j / \sum_j h_{ij}$.

ing this set of fake data, we estimate the model (2) and test for the null hypothesis $\beta = 0$ for three types of standard error corrections (heteroskedasticity robust, state-level clustering, and arbitrary clustering).

Panel A of Table 1 displays the results in a setting with no endogeneity. The model is estimated with OLS. Each row corresponds to a particular combination of scenario-correction while each column refers to different estimation samples of counties; and each row-column cell displays the average rejection rate across Monte Carlo iterations of the null hypothesis of no effect at the 5% significance level. We start, in column 1, with the estimation results based on the full sample of counties (N=3,141). As a benchmark, we consider in the first row the scenario 1, in which the policy shock is iid across observational units. We estimate the model computing heteroscedasticity-robust standard errors. As expected, the null-rejection rate is close to 5%. We then consider the scenario 2, in which the DGP allows for spatial correlation in the regressor `SCpolicy`. We first estimate the model with the same robust correction as in scenario 1, row (2). The null-rejection rate jumps to 9.1%. When standard errors are clustered at the state-level in row (3), the null-rejection rate decreases to 6.8%. This improvement stems from the fact that the DGP is designed such that many of the counties that are in the same spatial cluster are also in the same state; therefore, clustering at the state level approximates the existing spatial correlation structure to a certain extent. Finally, we correct for the presence of spatial correlation across counties using our `acreg` estimator. We obtain a null-rejection rate of 5.5% in row (4).[12]

To highlight the role of clusters separated by state-lines, we replicate the analysis after splitting the sample into two mutually exclusive subsamples of within-state and cross-state spatial clusters; they are made of counties which are treated by the DGP as belonging to spatial clusters never separated by a state border (column 2) and clusters separated by a state border (column 3), respectively. The null-rejection rates are lower in the sample of clusters that are located within a state compared to the sample with clusters that are separated by a state: 8.2% vs 10.4% with robust standard errors and 6.9% vs 9.2% with state-level clustering. State-level clustering correction approximates the underlying "true" spatial correlation structure of the DGP better for clusters that are located within a state than clusters separated by state lines. More importantly, our `acreg` estimator always performs substantially better than the state-level clustering correction — and this is the case in both subsamples — producing

[12]Note that our estimator requires as input a distance cutoff for setting the radius of the spatial clusters (below which all observations are considered as being located in the same cluster). In this baseline analysis, we select a cutoff equal to the "true" radius, i.e., the one used in the DGP. We discuss below, in section 3.1.4, situations where the practitioner has to set a cutoff without prior knowledge of the DGP.

null-rejection rates of approximately 5.5% to 5.8% for the spatially correlated random shocks. As seen in Figure 2, the outcome variable is not uniformly correlated across all counties within the same state. The correlation is greater across counties that are closer to one another within the same state. Therefore, taking into account the physical distance between counties (spatial units) performs much better than applying the same statistical treatment to all units within the same state (the greater administrative unit).

ENDOGENEITY. We now redo the whole analysis in a setting where the main regressor is endogenous. Hence, we switch to $\texttt{ENDpolicy}_i$ for the (endogenous policy) shock and estimate the model (2) through 2SLS, using $\texttt{IIDpolicy}_i$ as an instrumental variable. Panel B of Table 1 displays the results. The conclusions are unchanged with respect to the OLS environment. Firstly, the average null-rejection rate in the scenario with no spatial correlation (row 5) is close to 5% for all three estimation samples. Secondly, the scenario with spatial correlation is considered in rows (6) to (8) for various standard error corrections. We see that the null-rejection rate follows a pattern identical to the one in panel A. In particular, the arbitrary clustering correction provides rejection rates that are closer to the nominal rejection rates of the test than the other approaches.

SAMPLE SIZE. We next assess how the performance of our arbitrary clustering estimator is affected by sample size. The manipulation of the sample size is engineered in a simple way. In each Monte Carlo draw we retain the n largest counties in each state (excluding Washington D.C.), with $n \in \{3,4,...,20\}$, to obtain a filtered map of the US. Then, the DGP scenarios are simulated on this map. Figure 4 displays the null-rejection rates by sample size for the OLS and 2SLS settings (panels a and b respectively).[13] Each connected curve corresponds to a different scenario-estimation pair: no spatial correlation and robust correction (black); spatial correlation and robust correction (red); spatial correlation and state-clustering correction (green); spatial correction and arbitrary clustering (blue). Visual inspection confirms the better performance of the arbitrary clustering estimator at any level of the sample size, i.e., the corresponding null rejection rates being always the closest to the theoretical benchmark of 5%. It also It also shows that a larger sample size does not alleviate the inference problem when correcting for heteroskedasticity robust or state-level clustering.

[13] To ensure enough predictive power of the instrument in the first stage, even in the case of small sample sizes, we run 10,000 Monte Carlo iterations but report the average null-rejection rates only for the top half of the Monte Carlo draws in terms of F-statistics of the first stage.

3.1.3 Alternative Approach - Artificial data.

Kelly (2019) recently argued that many important empirical works using spatial data suffer from serious inference flaws. His main argument – Section 2 in Kelly (2019) – is based on Monte Carlo simulations using *exclusively* artificially generated data. By contrast, our approach follows Bertrand *et al.* (2004) and Cameron *et al.* (2011) in using real data for outcome and covariates; only the policy variable is artificially generated. We now show that the choice of simulating with real vs fake data affects results. To this purpose, we replicate our Monte Carlo analysis of Figure 4 with a set of artificial data as done in Kelly (2019). Figure 5 reports the average rejection rates of the null hypothesis $\beta = 0$ in equation 2 when regressing a randomly generated outcome on randomly generated policy shocks and covariates.[14] We see that artificial data lead to much larger null-rejection rates, up to a fourfold increase, compared to the ones obtained with real data in Figure 4. We conjecture that this explosion is mechanically driven by a severe spatial correlation between artificial variables: Indeed, they are all generated by the same DGP, with the same spatial kernel. Hence, conclusions drawn on the exclusive use of artificial data tend to exaggerate the extent of the inference problem encountered with real data, and should hence be interpreted with caution.

In spite of this important caveat, it is reassuring to see that our arbitrary clustering estimator delivers higher quality inference, i.e., rejection rates that are closer to the nominal ones, as soon as there is spatial correlation in the sample. Quite reassuringly, the quality of inference increases with sample size and the null-rejection rate converges toward the theoretical benchmark 5%. For a sample size of 3,141 counties, our proposed estimator reduces the null-rejection rates in the presence of spatial correlation from 27.4% (heteroskedasticity-robust standard errors) to 6.5% in the OLS setting and from 27.1% to 5.4% in the 2SLS setting. These results point toward the robustness of our estimator when dealing with alternative dataset (even entirely artificial).

[14] We generate random variables, Y and X, that are independent and identically distributed (iid): $Y, X = \sim N(0,1)$. To introduce spatial correlation to these variables, we impose a Bartlett kernel decay across observations within the same cluster. In other words, we spread the random variables across observations within the cluster as an inverse function of the distance between them. Then, we sum them up. Formally: $Y_{i,sc} = \sum_{j \neq i}^{N} [1 - (dist_{ij}/distcut)] \times Y_j$ and $X_{i,sc} = \sum_{j \neq i}^{N} [1 - (dist_{ij}/distcut)] \times X_j$, where N is the number of observations in the cluster of observation i, $dist_{ij}$ is the distance between observations i and j, and $distcut$ is the distance cuttoff. To introduce endogeneity to the model, we define an endogenous variable, End, as a function of Y, X, and IV. IV is a random variable and iid to Y and X: $IV = \sim N(0,1)$. Then, we instrument End with IV.

3.1.4 Understanding Spatial Correlation: A Practitioner's Guide

We designed further Monte Carlo simulations to shed light on several ways of improving the quality of inference in presence of spatial correlation in the model. We document first the role played by spatial correlation in the outcome variable. Then, we investigate whether or not potentially correlated regressors should be included. Last, we come up with practical recommendations on how to set the radius of the arbitrary spatial clusters when correcting standard errors without any prior knowledge of the true DGP.

Spatial correlation in both outcome variable and regressor. The results presented previously show that in the absence of spatial correlation in the treatment variable, `policy`, the null-rejection rates are close to the theoretical 5% despite the presence of spatial correlation in the outcome variable. This finding suggests that spatial correlation increases the likelihood of Type 1 error if unaccounted for, only when both the outcome variable and the variable of interest exhibit spatial correlation. We highlight this point with a variant of the DGP that suppresses spatial correlation in the outcome variable by randomly reshuffling log median income across counties.

Table 2 presents the average null-rejection rates obtained from 10,000 Monte Carlo simulations with the data generating process as described in section 3.1.1. Column 1 presents the baseline results obtained using observed log median income as the outcome variable, whereas column 2 presents those obtained using the randomized outcome variable, i.e., reshuffled across counties. In the absence of spatial correlation in the outcome variable (column 2), neither spatial correlation in the policy variable nor correction for it substantially affects the null-rejection rates; they all remain in the vicinity of the theoretical 5% benchmark. Column 3 considers an intermediate case where we re-inject spatial correlation affecting the outcome variable. We start from the randomized outcome variable and spread it across observations within a given spatial cluster as an inverse function of the distance between them (using a bartlett kernel decay). The result shows that re-injecting spatial correlation into the outcome variable leads to an increase in the null-rejection rates only when the policy variable also exhibits spatial correlation.

Our findings confirm that the quality of inference is deteriorated only when spatial correlation affects both the outcome variable and the regressor of interest. An important implication is that we should refrain from testing for the presence of spatial autocorrelation in *residuals* as a diagnostic against the possibility that a model suffers from inflated t-statistics, as suggested by Kelly (2019). The simple reason for being cautious here is that spatial correlation of

residuals is compatible with only one of the variables (outcome or regressor) being spatially correlated. We recommend instead to test for the presence of spatial correlation separately in the outcome variable and in the regressor of interest. If the test is inconclusive for at least one of the two, the practitioner could safely conclude that inference is unlikely to suffer from inflation of t-statistics because of spatial correlation.

INCLUDING CONTROL VARIABLES. In Table 3, we investigate how the control variables in the model 2 affect null-rejection rates. In column 1, we report the baseline null-rejection rates. Column 2 shows that, in presence of spatial correlation, the average null-rejection rates increases when the controls are not included. Column 3 shows that controlling for state fixed effects in addition to the controls yields a quality of inference that is as good as the one in the baseline case. Our interpretation is that the quality of inference tends to improve with the inclusion of additional control variables that exhibit a spatial kernel comparable to the one of the outcome variable and/or that of the regressor of interest.

OPTIMAL CORRECTION RADIUS - A SIMPLE RECOMMENDATION. We now investigate how the practitioner should set the radius of the arbitrary spatial clusters when correcting standard errors without any prior knowledge of the true DGP. Actually, to our best knowledge, no clear-cut procedure currently exists to define the potential optimal correction radius using observational data.

We start with setting the *true distance radius* in the DGP at 168 kilometers, such that there are on average 50 counties by spatial clusters. Then, we generate the data over 10,000 Monte Carlo iterations. In each iteration, using the arbitrary clustering estimator, we correct for spatial correlation in the data using different *correction radiuses*, namely: 56 kms (one-third of the true threshold, 5 counties on average), 82 kms (~half of the threshold, 12 counties on average), 117 kms (~two-thirds of the threshold, 25 counties on average), 168 kms (the true threshold), 242 kms (~1.5 times the threshold, 100 counties on average), 327 kms (~twice the threshold, 175 counties on average), and 478 kms (~three times the threshold, 350 counties on average).

Figure 6 reports the average null-rejection rate corresponding to each correction radius. Panel (a) considers the baseline case (i.e. a single policy treatment as in equation 2). For the sake of benchmarking, we note that when spatial correlation is present in the policy variable the (non-reported) null-rejection rate amounts to 11.9% with heteroskedasticity-robust standard errors; it drops to 7.5% with state-level clustering. Correcting for spatial correlation using small radiuses such as 56 kms and 82 kms or a very large one such as 478 kms, already outper-

forms robust standard errors (with rates between 9.1% and 10.5%); however, performance is worse than with state-level clustering. Using a correction radius closer to the real one, from 117 kms to 327 kms, yields null-rejection rates between 5.9% and 7.4% — all being below the average null-rejection rate obtained with state-level clustering. Crucially, we observe a U-shaped pattern in the null-rejection rates when spanning across correction radiuses, with a minimun (the best performance, close to the theoretical benchmark of 5%) that is reached when the correction radius is set exactly at the value of the true DGP radius. This evidence suggests a rule of thumb for setting the correction radius: In absence of prior knowledge of the true DGP, the practitioner should (i) estimate standard errors for a large range of potential correction radiuses; (ii) check for the presence of a non-linear pattern; and (iii) retain the correction radius that yields the most conservative standard errors.

How to generalize the previous guideline in a context where the researcher is interested in estimating and inferring more than one parameter of the model? We shed some light on this question by looking for the optimal correction radius in a variant of equation 2 with two policy variables. To this purpose, we consider a DGP augmented with a second random policy variable that is also spatially correlated, but with a true distance radius of 242 kilometers (keeping 168 kms for the first variable). In other words, the spatial kernels of the two policy variables differ, as it likely to be the case with real-life applications. Panel (b) of Figure 6 presents the average null-rejection rates for each policy variable across the range of correction radiuses spanning from 56kms to 478kms. Note that our arbitrary clustering estimator imposes the same correction radius to the two variables. Here again, we find, for each variable, a U-shaped pattern in null-rejection rates with the lowest rate (6.2% and 6.5% respectively) reached when the correction radius is set at its true level in the DGP –a level that is specific to each variable. This graphical evidence makes clear that there is no *universal* correction radius that limits the inflation of t-statistics for all regressors in a given econometric specification. The practical consequence is that the researcher may want to apply the previous rule of thumb *separately* for each of the regressors of interest. This procedure would yield a set of correction radiuses, each one corresponding to the optimal radius for a given regressor. And then the researcher could report for each parameter of interest all the set of standard errors sequentially estimated with the set of radiuses. A more compact alternative could also be to report the most conservative standard error obtained for each variable of interest.

OPTIMAL CORRECTION RADIUS WITH MIXTURE OF SPATIAL KERNELS. So far, the DGP was designed to impose the same true distance radius for all observational units when generating

spatially correlated policy variables. However, with real-life observational data, it may happen that the size of spatial clusters differs across space and observations. For example, surface area of the counties in the Northeast, Midwest, and South regions of the US are substantially smaller than those in the West (see Figure 2). Similarly, clusters of low median income and high median income counties also tend to be smaller in size in these regions than they are in the West. We now investigate the consequence of this mixture of spatial kernels for setting the correction radius in practice.

We consider a variant of the DGP that allows the size of spatial clusters to vary across observations. To this purpose, we rank counties in terms of surface areas. We then set the true distance radius of the smallest county at 168 kilometers (corresponding to 50 counties per cluster on average) and that of the largest county at twice of it — 336 kilometers. We set the true distance radius for each county in between with a proportionality rule according to its rank and end up with a uniform distribution of cluster radiuses. Panel (a) of Figure 7 presents the histogram of the true cluster radiuses defined as such. Then, the model is estimated using our arbitrary clustering procedure with different correction radius, namely: 56 kms (5 counties on average), 82 kms (12 counties on average), 117 kms (25 counties on average), 168 kms (50 counties), 208 kms (75 counties on average), 242 kms (100 counties on average), 287 (138 counties on average), 327 kms (175 counties on average), 398 (250 counties on average) and 478 kms (350 counties) on average). Panel (b) of Figure 7 presents the average null-rejection rates obtained from using each of these correction radiuses over 10,000 Monte Carlo draws. Remarkably, we keep on observing a U-shaped profile of null-rejection rates, that is admittedly flatter than the one obtained with a unique spatial kernel in Figure 6. Correction radiuses between 168 kms and 336 kms yield very similar null-rejection rates (around 7.4%) that are all below the rate obtained with heteroskedasticity-robust standard errors (11.7%) or state-level clustering (8.5%). According to this finding, our simple rule of thumb for setting the correction radius can also be applied in presence of a mixture of spatial kernels in the underlying DGP. In that case, the procedure is likely to select the correction radius that overlaps with the true distance radius of a greater number of observations.

Finally, we explore a DGP that allows for continuous distribution of spatial bandwidths. With respect to the previous DGP, this variant allows us to generate cluster radius sizes that are correlated with the surface area of counties — a likely feature in observational data. Specifically, we set the true distance radius of the median county in terms of surface area at 168 kilometers. For all other counties, their true radius is defined as 168 times the square root of the ratio of their surface area over that of the median county. Panel (a) of Figure 8 presents

the resulting distribution of true cluster radius sizes.[15] The cluster radiuses are concentrated around the radius assigned to the median county — 168 kms. Then, we estimate the model with our arbitrary clustering procedure using the same set of correction radiuses than in the previous approach. Average null-rejection rates are reported in Panel (b) of Figure 8 and the evidence is qualitatively and quantitatively very comparable. An additional insight relates to the fact that the null-rejection rates follow a distribution similar to that of the cluster radiuses: We obtain lower null-rejection rates with correction radiuses around which the true cluster radius of a greater number of counties fall. Importantly, our simple rule of thumb will select the correction radius that matches the radius of the median county in terms of surface area (i.e. 168 kms), around which the cluster radiuses are concentrated.

IMPLICATIONS. As shown by our simulations, the presence of spatial correlation, if unaccounted for, can lead to inflated t-statistics only if both the outcome variable and the regressor of interest are spatially correlated. Controlling for covariates (that follow a similar spatial kernel) and clustering standard errors at a greater administrative unit can help with addressing the inflation in t-statistics. However, a more compelling approach is to explicitly model the spatial correlation structure with our arbitrary clustering estimator.

When deciding on how to set the correction radius in absence of prior knowledge about the underlying DGP, as Cameron and Miller (2015) put it: "You need to think carefully about the potential for correlations in your model errors, and how that interacts with correlations in your covariates." Our Monte Carlo simulations in a controlled environment suggest a simple rule of thumb for setting the optimal correction radius for each parameter of interest. Practitioners should correct standard errors with varying correction radiuses (and potentially using different distance metrics) and select as the baseline the radius that provides the largest standard errors for a given model. In the presence of multiple outcomes of interest, we advise selecting a correction radius that provides the largest standard errors for most of the variables of interest as the baseline. Overall, we recommend that researchers, as a healthy practice, be transparent about their choice of baseline correction radius and report the robustness of their findings to correcting the standard errors in their models using a wide range of correction radiuses.

In practice, however, it is possible that the correlation structure in the data cannot be approximated by spatial clusters defined as circles with a given radius. For example, topographic features such as mountain ranges could generate variations in the distribution of the outcome

[15] The histogram is scaled with counties radiuses between 56 kilometers and 478 kilometers, corresponding to 96% of the full sample.

variable and covariates across spatial units that are in close proximity in terms of Euclidean distance. To help address this issue, our proposed estimator's companion statistical package (`acreg`) allows users to provide a bilateral-distance matrix of any metric between observations. Then, the distance radius used for error correction can be defined as *effective distance* between observations in terms of time or cost of travel (flight, road, or walking) distance.

3.2 Network Setting

We now present the Monte Carlo experiments conducted in a network setting. The analysis follows closely the one conducted in the spatial part. As before, we first explain how the DGP is designed, we then report and comment the results. We conclude with a list of recommendations for the applied practitioner.

3.2.1 The Data Generating Process in the Network Setting

We extract information on author characteristics from IDEAS RePEc and complement it with information on coauthorship links between researchers coming from RePEc genealogy.[16] We start from the "Top 5% Authors, Number of Citations, as of October 2019" list of IDEAS RePEC.[17] We collect information on the primary affiliation of the listed authors — name of the institution they are affiliated to, the country and city where it's located — the year in which they obtained their PhD degree and from which school, in addition to the number of citations their work has received. We only consider researchers who are currently alive, affiliated with an institution, and whose coauthor network is observed in the RePEc genealogy. Our observational units i is one of the 1,637 researchers in economics that are covered in our sample. For each researcher, we only consider the part of her network of coauthors who are also in this sample.

When considering equation (2), the outcome variable Y_i is the log number of citations author i received. the vector of author-level controls X_i is empty in our baseline analysis but comprises fixed effects for affiliation and degree school and year of PhD graduation in additional analysis. In this setting the random shock is a variable $\texttt{productivity}_i$ indicating whether author i receives a fake productivity shock. We now explain how we construct this shock in the two different scenarios: with and without network correlation.

[16] https://genealogy.repec.org/

[17] This list is updated monthly. The most up-to-date version can be accessed using the following link: https://ideas.repec.org/top/top.person.nbcites.html

SCENARIO 1: NO NETWORK CORRELATION. Similar to the approach used in the spatial setting, we draw in each iteration an iid random variable, $\tilde{u}_i$, from a standardized normal distribution (with cdf Φ), for each author in the sample. Then, we select the authors who are in the top quarter of the distribution of this random variable as those who receive a fake productivity shock: $\texttt{IIDproductivity}_i = 1$ iff $\Phi(\tilde{u}_i) > 0.75$ (and 0 otherwise). This productivity shock is unrelated to the author's actual number of citations (outcome variable) and is uncorrelated with the productivity shock of her co-authors. As an example, panel (a) of Figure 9 visualizes the distribution of the `IIDproductivity` variable drawn at random in one of the Monte Carlo draws.

SCENARIO 2: NETWORK CORRELATION. In this scenario, we impose network correlation while generating the productivity shocks. Specifically, for each author, we construct an indicator that identifies a co-authorship connection between two researchers i and j in the sample: $g_{ij} = 1$ if authors i and j coauthored at least one paper and $g_{ij} = 0$ otherwise. For each author i, we then compute the share of their first degree coauthors who are affected by the productivity shocks, i.e., $\texttt{NCshare}_i = \sum_b g_{ij}\texttt{IIDproductivity}_j / \sum_j g_{ij}$.[18] Then, we define the correlated productivity shocks as the sum of the idiosyncratic productivity shock and the share of the author's coauthors hit by a productivity shock, i.e., $\texttt{NCproductivity}_i = \texttt{NCproductivity}_i + \texttt{NCshare}_i$. As an example, panel (b) of Figure 9 displays the distribution of the correlated productivity shock, `NCproductivity`, in the same Monte Carlo draw used for panel (a). Given the high number of observations, the figures in panels (a) and (b) do not provide a useful representation of the mechanism that builds the correlated productivity shock. Panels (c) and (d) of Figure 9 shows the distribution of the variables `IIDproductivity` and `NCproductivity`, respectively, for a subsample of the 250 most cited authors in the sample. While this subsample is not representative of the full sample, it is effective in showing how the random productivity shock dissipates across observations in the network. Among the nodes with a null random shocks (blue dots in panel (c)), the ones that are not connected to any affected node receive also a null correlated productivity shock (blue dots in panel (d)), while the ones that are connected to affected nodes get a value of shocks between zero and one. The highest value of the correlated shock (red dots in panel (d)) are reported by nodes that are affected by the random shock (red dots in panel (c)) and are connected to other affected nodes.

[18] We adopt a setting where shocks are correlated in coauthor neighborhoods of degree 1. Larger neighborhoods and decay in shocks can be accommodated in our estimator as well.

ENDOGENEITY. We introduce endogeneity into the model by generating a random productivity shock that is correlated with the outcome variable, i.e., log number of citations. Authors who receive an endogenous productivity shock, `ENDproductivity`, are randomly selected among highly cited authors, i.e., those who are in the top half of distribution of the log number of citations. We rely on the same random variable $\check{u}_i$ used to define the exogenous productivity shock $\texttt{IIDproductivity}_i$.[19] Consequently, `ENDproductivity` is correlated with the exogenous productivity shock, `IIDproductivity`, making the latter a valid instrument for the former when estimating a 2SLS version of the econometric model (2). Note that the exclusion restriction is also met by construction because the instrument `IIDproductivity` is uncorrelated with the outcome variable. We also factor in network correlation in this setting by creating an endogenous productivity shock that is correlated within the coauthorship network, `NCENDproductivity`, in the same way we generate an exogenous productivity that is correlated within coauthorship networks.[20]

3.2.2 RESULTS

We run Monte Carlo simulations with 10,000 iterations. In each Monte Carlo draw, the DGP produces random productivity variables according to the aforementioned scenarios. We use these generated datasets to assess the performance of our arbitrary clustering estimator when testing for the null hypothesis $\beta = 0$ in equation 2. Our main objective is to compare various types of standard error corrections across different scenarios.

Panel A of Table 4 presents the results in a setting with no endogeneity where the log number of citations is regressed on the exogenous random productivity shocks. Each row corresponds to a particular combination of DGP scenario and correction; each column refers to different specifications with respect to the control and outcome variables used. Each row-column cell displays the average rejection rate across Monte Carlo iterations of the null hypothesis of no effect at the 5% significance level. In column 1, we consider the full sample of authors (N=1,367) and a univariate model with no controls. In row (1), we start, as a benchmark, with iid productivity shocks from scenario 1 and heteroscedasticity-robust standard errors. As expected, the average null-rejection rate is close to 5%. Then, we move to scenario

[19] We select as $\texttt{ENDproductivity}_i = 1$ the authors who are in the top half of the distribution of $\check{u}_i$, *conditional* on being a more-productive author (i.e. above median in terms of log number of citations). In this way, both `IIDproductivity` and `ENDproductivity` take the value of 1 exactly for a quarter of the authors. Formally we set: $\texttt{ENDproductivity}_i = 1$ iff $\Phi^{MP}(\check{u}_i) > 0.5$ where Φ^{MP} is the cdf of the standardized normal distribution conditional on being a more productive author.

[20] Formally, we define $\texttt{NCENDproductivity}_i \equiv \texttt{ENDproductivity}_i + \texttt{NCENDshare}_i$ where $\texttt{NCENDshare}_i \equiv \sum_j g_{ij}\texttt{ENDproductivity}_j / \sum_j g_{ij}$.

2 that imposes correlation in the productivity shocks across first-degree connections in coauthorship networks. Heteroscedasticity-robust standard errors, row (2), produce average null-rejection rate of 9.8%. Next, we attempt to address network clustering using clusters at a level where coauthorship-network formation and productivity of authors correlate. Natural candidates for the clustering level correspond to institutions authors are affiliated with (row 3, 611 clusters), the city of location of these institutions (row 4, 259 clusters), the schools they graduated from (row 5, 202 clusters) and the city of location of these schools (row 6, 135 clusters). None of these non-overlapping clustering structures reduces substantially the null-rejection rates. If anything, we obtain null-rejection rates that are inflated compared to those derived under heteroskedasticity-robust standard errors. Finally, we correct standard errors using our arbitrary clustering estimator in row (7). To account for correlation across authors linked in coauthorship networks, we use information on all links between co-authors to form the pattern matrix required for arbitrary clustering.[21] The average null-rejection rate falls down substantially, to 5.6%, close to the theoretical prediction of 5%. This finding clearly indicates that, in presence of network correlation, arbitrary clustering provides null-rejection rates that are closer to nominal levels than standard correction procedures based on non-overlapping clusters.

ENDOGENEITY. In panel B, rows (8)–(14) of Table 4, we turn to a setting with endogenous productivity shocks and estimate the econometric equation 2 through 2SLS, using the exogenous productivity shock as an instrumental variable. The sequence of specifications follows the same logic as the one of panel A, rows (1)–(7). We see that the conclusion does not change with respect to the OLS case, with null-rejection rates following a comparable pattern. The average null-rejection rate in the iid scenario reported in row (8) is again 4.8% and it goes up to 9.8% in row (9) with network correlation and heteroskedasticity-robust standard errors. Here also, various methods based on non-overlapping clusters do not help in reducing the bias, augmenting rather than decreasing the null-rejection rates as shown in rows (10–13). Ultimately, the arbitrary clustering correction keeps on being the best performer producing null-rejection rates of 5.7% in row (14).

[21] Specifically, we collect information on links g_{ij} between author i and j into a matrix, which yields the adjacency matrix G, whose ij element is g_{ij}. The adjacency matrix does not contain self-links, while the pattern matrix S always does. So the pattern matrix that we use in the arbitrary clustering calculations is $S = I + G$ where I is the identity matrix. We use this matrix in equation 1.

SAMPLE SIZE. We next assess how the performance of the arbitrary clustering estimator is affected by sample size. To engineer changes in size, we retain only the top n authors with the highest number of coauthors within the sample, with $n \in \{150, 200, ..., 1000\}$. The other aspects of the Monte Carlo simulation are unchanged. Figure 10 reports average null-rejection rates for different sample sizes, both in a OLS setting (panel a) and a 2SLS setting (panel b). Each connected line corresponds to a specific scenario-estimation pair: no network correlation and heteroskedastisity-robust standard errors (black); network correlation and heteroskedastisity-robust standard errors (red); network correlation and arbitrary clustering (blue).[22] Visual inspection reveals that the performance of the arbitrary clustering estimator improves as the sample size increases with a null-rejection rate approaching the theoretical benchmark 5%. Quite importantly, this graphical evidence confirms that, in presence of network correlation, a larger sample size does not alleviate inference problem when correcting for heteroskedasticity-robust standard errors.

3.2.3 NETWORK CORRELATION: SOME GUIDELINES

In the following section, we aim at providing guidelines to the practitioner interested in estimating the standard errors in presence of network correlation in the model. Here again, the analysis closely follows the one conducted with spatial data. Thus, we tend to skip the technical details and go to the essentials.

The results presented in panel A, row 1 of Table 4, show that in the absence of network correlation in the regressor of interest (productivity shock), the average null-rejection rate is close to the theoretical 5% despite the presence of network correlation in the outcome variable, i.e., log number of citations. Hereafter we assess whether the presence of network correlation in the variable of interest leads to an increase in the likelihood of making a Type 1 error (if unaccounted for) when the outcome variable is not correlated across authors within the same network. In each Monte Carlo draw, we first suppress network correlation in the outcome variable through a random reshuffling of the log number of citations across authors; then we test for the null hypothesis of no effect at the 5% significance level. Column 2 of Table 4 presents the average null-rejection rates obtained from 10,000 Monte Carlo simulations using the reshuffled outcome variable. In all rows, whatever the correction method and the estimator (OLS in panel A, 2SLS in panel B), the null-rejection rate remains in the vicinity of the theoretical 5%.

[22] To ensure that the first stage has enough predictive power even in the case of small sample sizes, we run 10,000 Monte Carlo simulations in each iteration but report the average null-rejection rates for the top half of the Monte Carlo draws in terms of F-statistics of the first stage.

This finding indicates that the likelihood of wrongly rejecting the null hypothesis is inflated only if network correlation affects both the outcome variable and the regressor of interest. This confirms an important insight that emerged also in the analysis conducted with spatial data.

We also investigate how the inclusion of control variables in the model affects the quality of inference. We proceed by assessing how different types of controls, taken separately, impact null-rejection rates. We consider author characteristics that are likely to be correlated with coauthorship networks and productivity of authors. In columns 3, 4, and 5 of Table 4, we control for affililation-country fixed effects, degree-school fixed effects, and linearly by the year in which authors received their PhD degree, respectively. Controlling for affiliation-country fixed effects and degree-school fixed effects decreases the average null-rejection rate obtained when network correlation is not corrected for from 9.8% to 9.3% and 8.9% in the OLS setting (and from 9.6% to 9.2% and 8.8% in the 2SLS setting). In contrast, controlling for the year in which authors received their PhD degree increases the average null-rejection rates obtained when network correlation is not accounted for to 10.3% in both the OLS and 2SLS settings. Our arbitrary clustering estimator performs equally well in all of these specifications. This result suggests that the magnitude of t-statistics inflation due to network correlation depends on the degree of network correlation in the residual variation left in the outcome variable and variable of interest, conditional on the set of covariates.[23]

RECOMMENDATIONS. Our analysis shows that researchers should correct for network correlation when estimating the standard errors for parameters of interest. In this endeavor, our arbitrary clustering estimator tends to outperform existing alternative correction procedures (robust, non-overlapping clusters) in making inference. When implementing arbitrary clustering in a network context, the VCV matrix should be corrected using the adjacency matrix. Our findings also show that the quality of inference tends to improve with the inclusion of additional control variables that exhibit a level of network correlation comparable to the one of the outcome variable and/or of the regressor of interest.

4 CONCLUSION

We implement a novel approach to obtain an asymptotically valid inference in settings with spatial or network topology, allowing for any type of dependence between observation units.

[23] Adão *et al.* (2019) present inference in shift-share designs in terms of residual variation left in the regressors.

Our proposed variance-covariance matrix (VCV) estimator, accompanied by a companion statistical package `acreg` for Stata, allows researchers to obtain cluster-robust inference in OLS and 2SLS settings with arbitrary dependence across observations and over time. Arbitrary here refers to the way units could be correlated with each other in space and time. Our approach allows units to be correlated with each other in any possible way: The estimator can account for indirect links in the cross-sectional dependence, time dependence and alteration in the correlation structure over time. This allows our estimator to be suitable for many applications.

Our empirical validation approach is to compare the quality of inference based on arbitrary clustering to conventional estimators, e.g., one-way clustering and the heteroscedastic-consistent estimator. Our Monte Carlo simulations use real-life data on U.S. counties and co-authorship in Economics. We find three key results. First, arbitrary clustering inference dominates inference based on conventional estimators, i.e., the rejection rate of the null hypothesis of no effect is always closer to its nominal size (of 5% in our simulations) for the arbitrary clustering estimator compared to conventional estimators. Second, arbitrary clustering inference improves as the sample size gets larger, while inference based on conventional estimators does not improve. Third, we show that the main source of biased inference is simultaneous presence of (spatial or network) correlation in both the outcome residuals and the regressors. We obtain this pattern of results both in spatial data on U.S. counties and in data on co-author networks in Economics.

These results suggest that accounting for arbitrary clustering appears to be important in spatial or network data. Conventional estimators of standard errors are not flexible enough to address correlations across state borders or across co-authors from different PhD programs or affiliated with different institutions. Granted, a key requirement for the arbitrary clustering procedure is the bandwidth choice, which reflects the maximum distance in which units are thought to be correlated. We designed simulations to guide applied work on choosing the right bandwidth. We find that estimated standard errors are largest, and rejection rates closest to nominal size, when the bandwidth chosen by the researcher is close to the true (DGP) bandwidth. These simulations offer practical guidance for implementation in one of the challenging problems of spatial research.

REFERENCES

Adão, R., Kolesár, M., and Morales, E. (2019). Shift-Share Designs: Theory and Inference*. *The Quarterly Journal of Economics*, **134**(4), 1949–2010.

Arellano, M. (1987). Computing robust standard errors for within-groups estimators. *Oxford Bulletin of Economics and Statistics*, **49**(4), 431–34.

Bertrand, M., Duflo, E., and Mullainathan, S. (2004). How much should we trust differences-in-differences estimates?*. *The Quarterly Journal of Economics*, **119**(1), 249–275.

Bramoullé, Y. and Fortin, B. (2010). *social networks: econometrics*, volume 4 of *The New Palgrave Dictionary of Economics*. Palgrave Macmillan.

Calvó-Armengol, A., Patacchini, E., and Zenou, Y. (2009). Peer Effects and Social Networks in Education. *The Review of Economic Studies*, **76**(4), 1239–1267.

Cameron, A., Gelbach, J., and Miller, D. (2011). Robust inference with multiway clustering. *Journal of Business and Economic Statistics*, **29**(2), 238–249.

Cameron, C. A. and Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. *Journal of Human Resources*, **50**(2), 317–372.

Conley, T. (1999). Gmm estimation with cross sectional dependence. *Journal of Econometrics*, **92**(1), 1–45.

Fafchamps, M. and Gubert, F. (2007). The formation of risk sharing networks. *Journal of Development Economics*, **83**(2), 326–350.

Fafchamps, M. and Lund, S. (2003). Risk-sharing networks in rural Philippines. *Journal of Development Economics*, **71**(2), 261–287.

Hsiang, S. M. (2010). Temperatures and cyclones strongly associated with economic production in the caribbean and central america. *Proceedings of the National Academy of Sciences*, **107**(35), 15367–15372.

Ibragimov, R. and Müller, U. K. (2016). Inference with few heterogeneous clusters. *The Review of Economics and Statistics*, **98**(1), 83–96.

Kelejian, H. H. and Prucha, I. (1998). A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances. *The Journal of Real Estate Finance and Economics*, **17**(1), 99–121.

Kelejian, H. H. and Prucha, I. (1999). A generalized moments estimator for the autoregressive parameter in a spatial model. *International Economic Review*, **40**(2), 509–33.

Kelly, M. (2019). The standard errors of persistence. *CEPR Discussion Paper Series 13783*.

König, M. D., Rohner, D., Thoenig, M., and Zilibotti, F. (2017). Networks in conflict: Theory and evidence from the great war of africa. *Econometrica*, **85**(4), 1093–1132.

Lalive, R., Luechinger, S., and Schmutzler, A. (2018). Does expanding regional train service reduce air pollution? *Journal of Environmental Economics and Management*, **92**, 744 – 764.

Manson, S., Schroeder, J., Van Riper, D., and Ruggles, S. (2017). Ipums national historical geographic information system: Version 12.0 [database]. Minneapolis: University of Minnesota. http://doi.org/10.18128/D050.V12.0.

Michalopoulos, S. and Papaioannou, E. (2017). Spatial patterns of development: A meso approach. NBER Working Papers 24088, National Bureau of Economic Research, Inc.

Moulton, B. R. (1990). An Illustration of a Pitfall in Estimating the Effects of Aggregate Variables on Micro Unit. *The Review of Economics and Statistics*, **72**(2), 334–338.

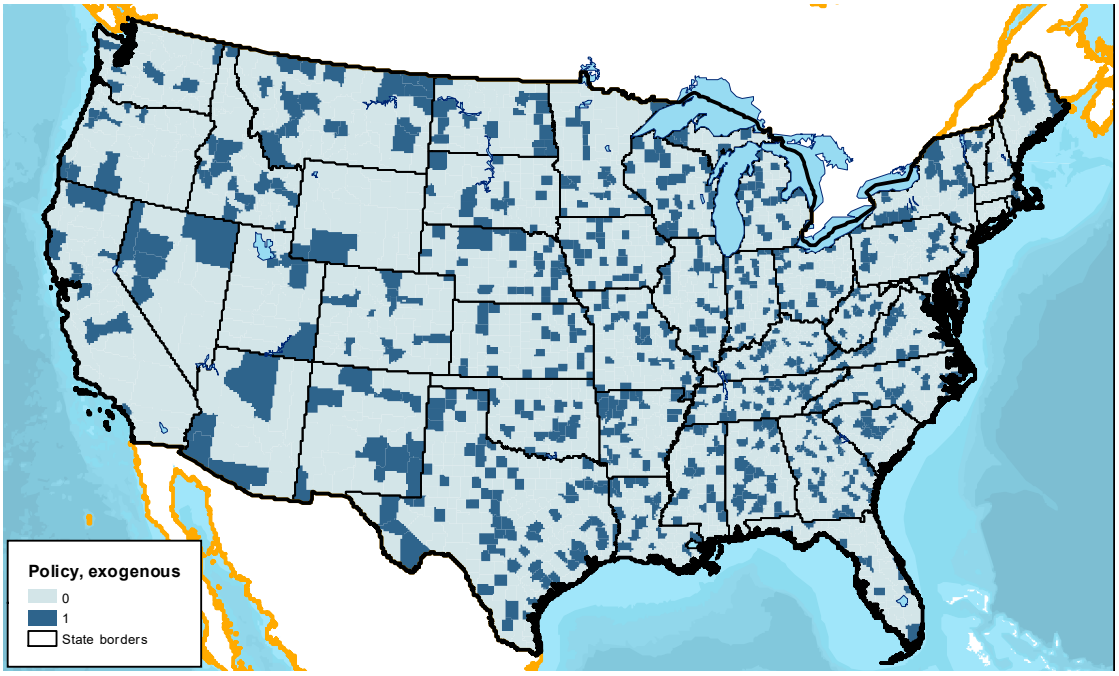
White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, **48**(4), 817–38.

White, H. (1984). *Asymptotic Theory for Econometricians*.

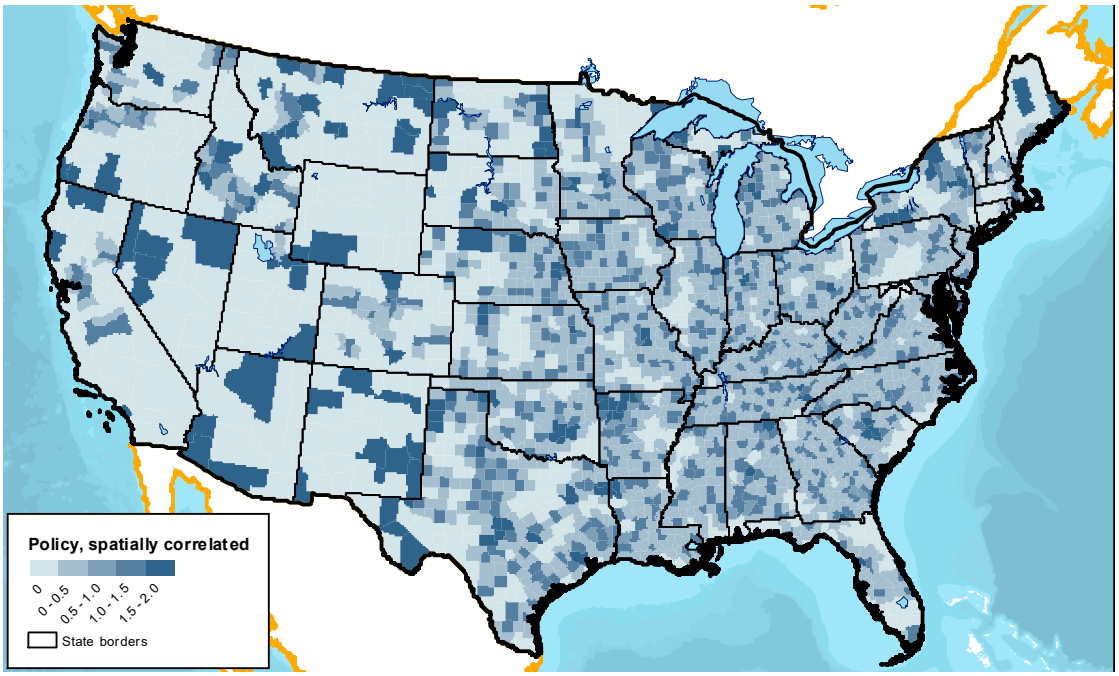
Wooldridge, J. M. (2003). Cluster-Sample Methods in Applied Econometrics. *American Economic Review*, **93**(2), 133–138.

Figure 1: Illustration of data generation process: exogenous shocks in U.S. counties

(a) Random policy shock

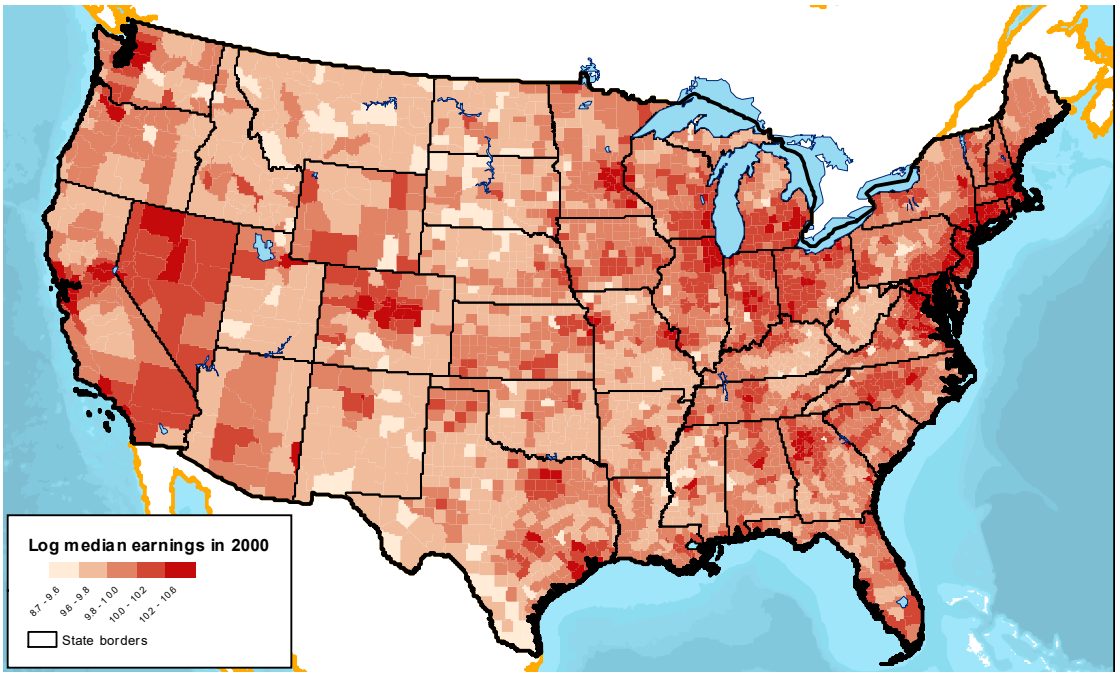


(b) Spatially correlated policy shock



Notes: Data source for the county boundaries: NHGIS (Manson *et al.*, 2017). The values of the exogenous policy shocks represented are randomly generated with the algorithm described in section 3.1.1.

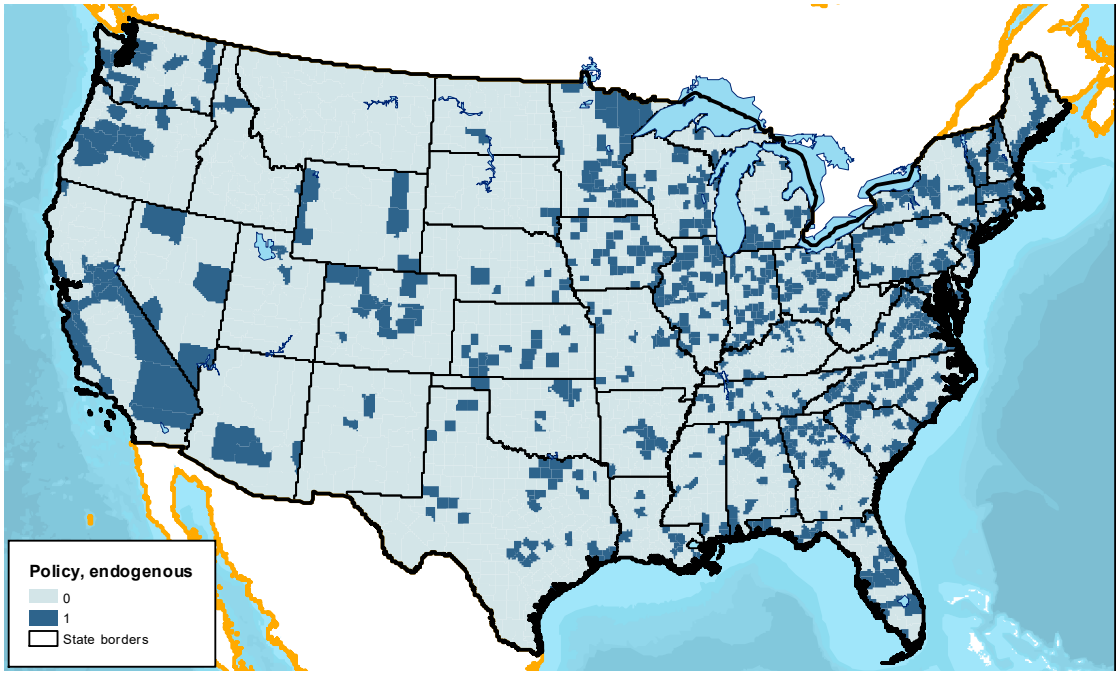
Figure 2: Log median earnings across US counties in 2000



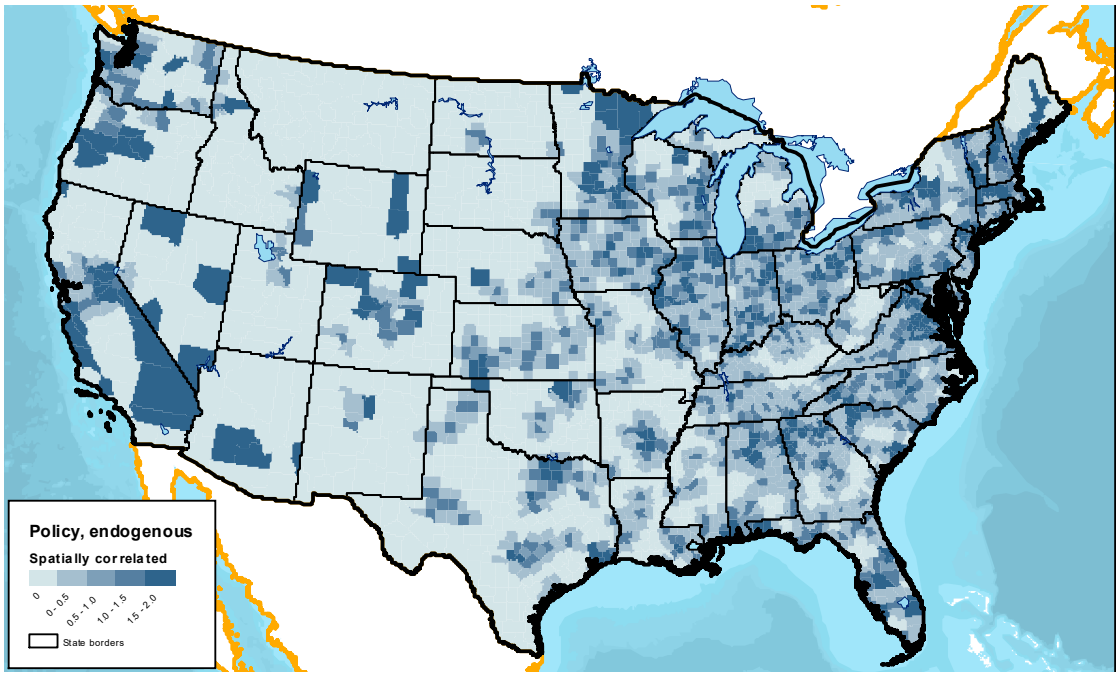
Notes: Data source for the county boundaries and log median earnings in 2000: NHGIS (Manson *et al.*, 2017).

Figure 3: Illustration of data generation process: endogenous shocks in U.S. counties

(a) Random endogenous policy shock



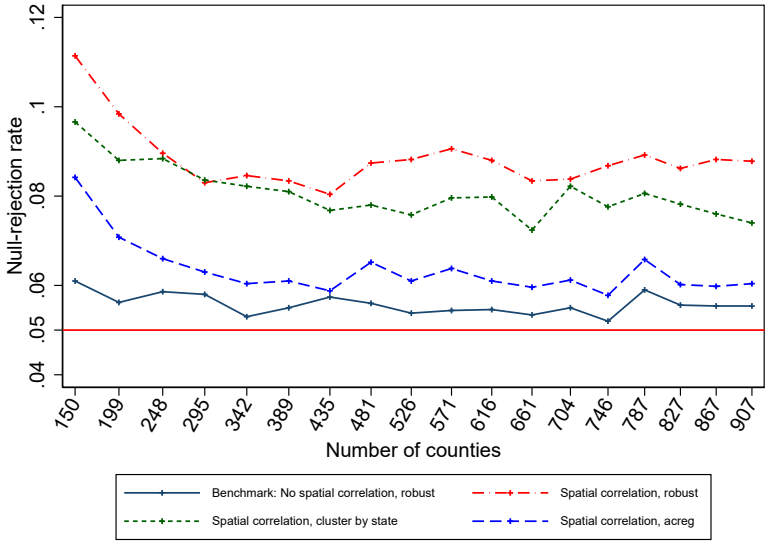
(b) Spatially correlated endogenous policy shock



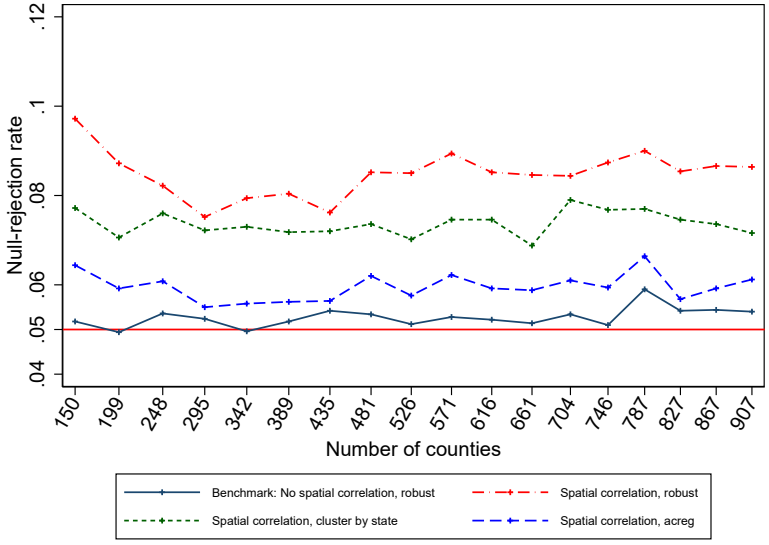
Notes: Data source for the county boundaries: NHGIS (Manson *et al.*, 2017). The values of the endogenous policy shocks represented are randomly generated with the algorithm described in section 3.1.1.

Figure 4: Null-rejection rate in the presence of spatial correlation: U.S. counties

(a) OLS



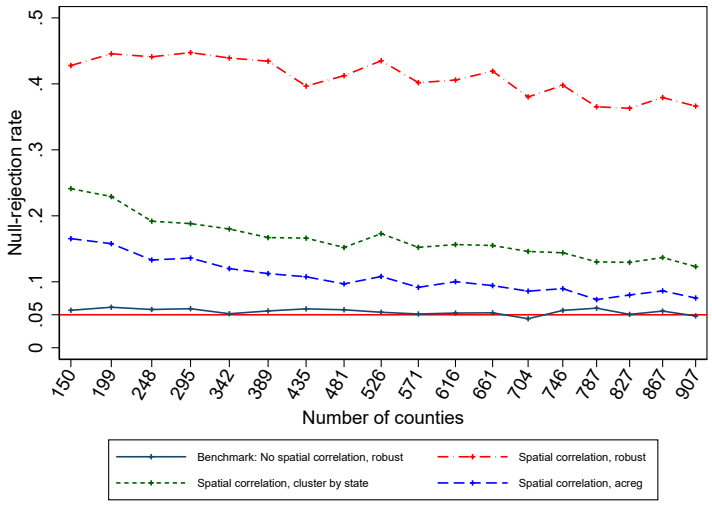
(b) 2SLS



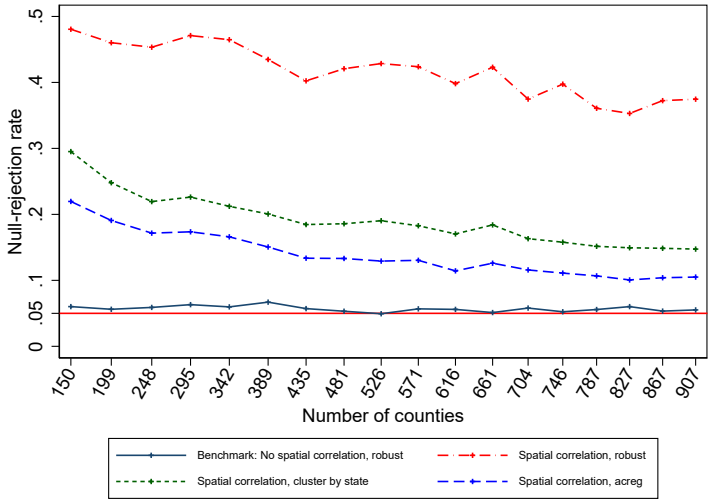
Notes: The red horizontal line represents the benchmark null-rejection rate of 5%. The vertical axis represents the rejection rate of the average null-rejection over 10,000 Monte Carlo simulations. Each point in the figure represents a different Monte Carlo simulation × estimation pair. The horizontal axis represents the sample size.

Figure 5: Null-rejection rate in the presence of spatial correlation: U.S. counties with fake data

(a) OLS



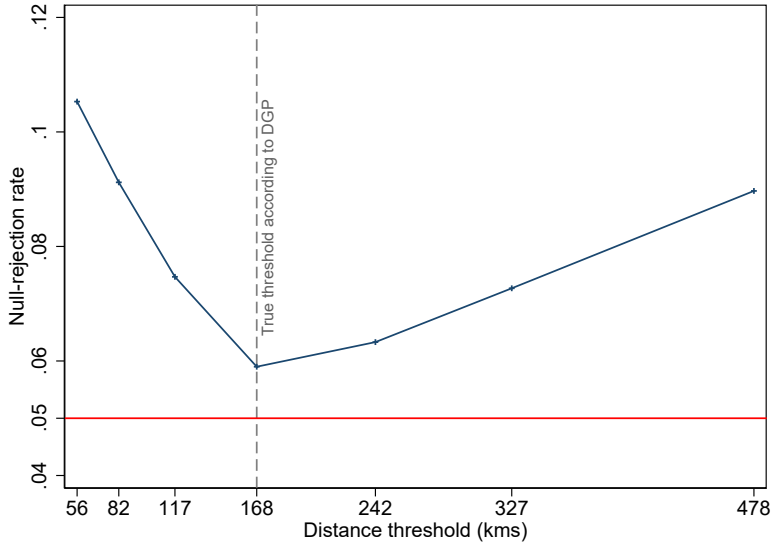
(b) 2SLS



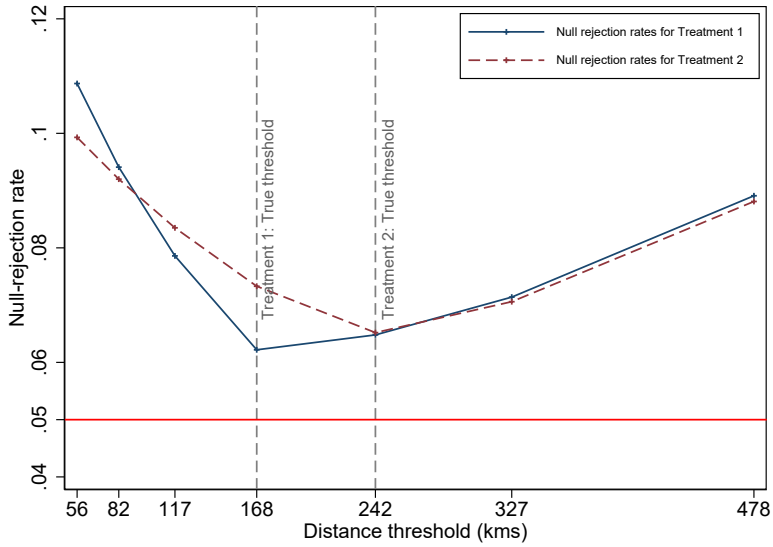
Notes: The red vertical line represents the benchmark null-rejection rate of 5%. The vertical axis represents the rejection rate of the average null-rejection over 10,000 Monte Carlo simulations. Each point in the figure represents a different Monte Carlo simulation × estimation pair. The horizontal axis represents the sample size.

Figure 6: Spatial setting: Optimal distance threshold and null-rejection rates

(a) Single policy treatment



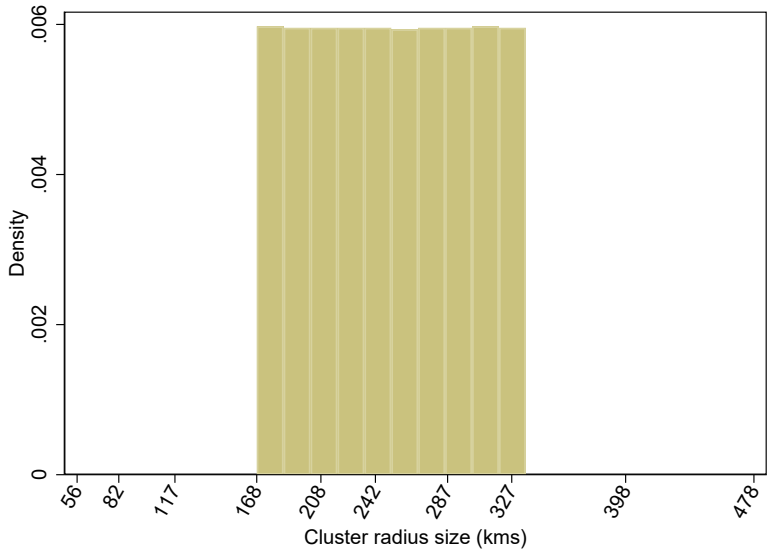
(b) Multiple policy treatments



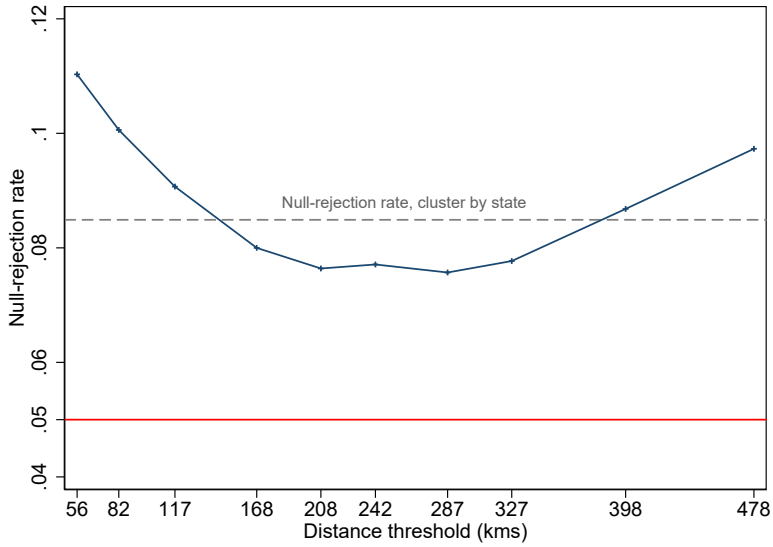
Notes: The red vertical line represents the benchmark null-rejection rate of %5. The vertical axis represents the rejection rate of the average null-rejection over 10,000 Monte Carlo simulations on a sample of 3,141 counties. Each point in the figure represents a different Monte Carlo simulation uses a different distance threshold used for error correction.

Figure 7: Spatial setting: Varying cluster radiuses following a uniform distribution

(a) Histogram of the cluster radiuses sizes



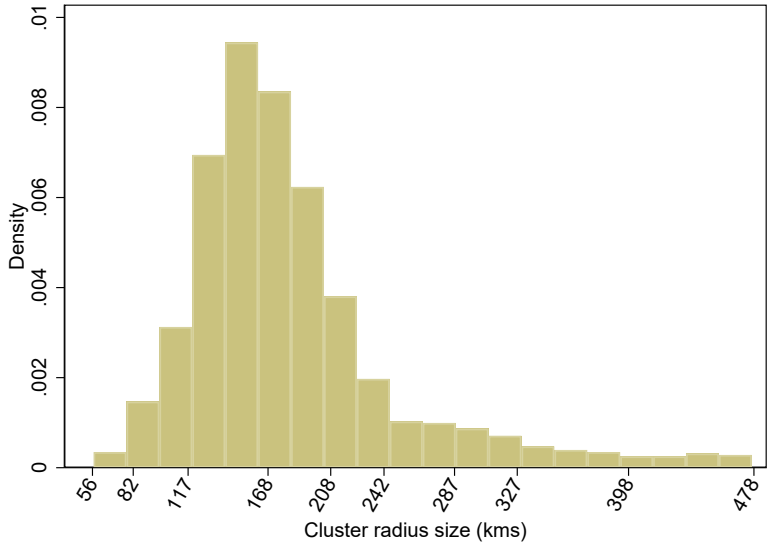
(b) Null-rejection rates by different Conley correction thresholds



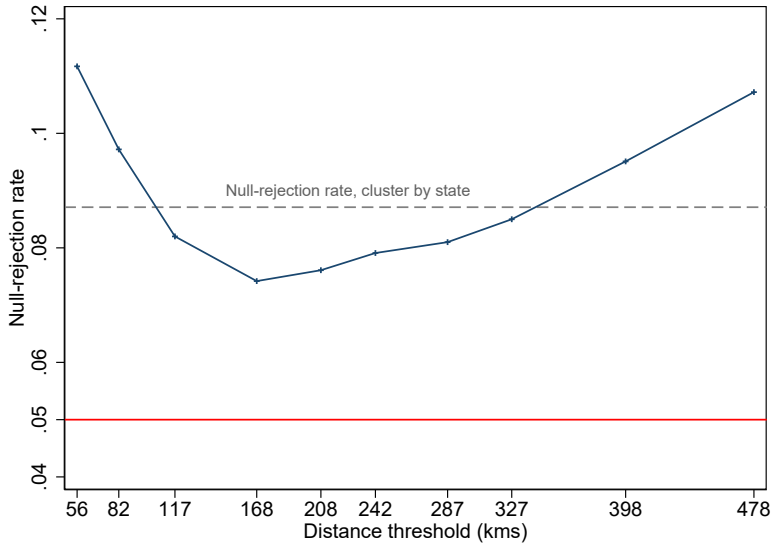
Notes: The red vertical line represents the benchmark null-rejection rate of %5. The vertical axis represents the rejection rate of the average null-rejection over 10,000 Monte Carlo simulations on a sample of 3,141 counties. Each point in the figure represents a different Monte Carlo simulation uses a different distance threshold used for error correction.

Figure 8: Spatial setting: Varying cluster radiuses following a nonuniform distribution

(a) Histogram of the cluster radiuses sizes



(b) Null-rejection rates by different Conley correction thresholds

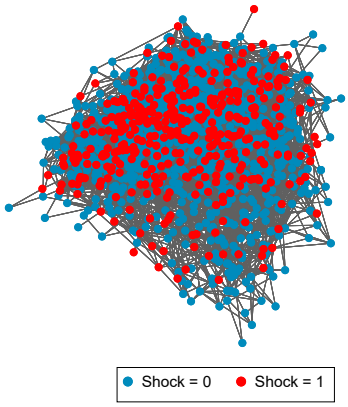


Notes: The red vertical line represents the benchmark null-rejection rate of %5. The vertical axis represents the rejection rate of the average null-rejection over 10,000 Monte Carlo simulations on a sample of 3,141 counties. Each point in the figure represents a different Monte Carlo simulation uses a different distance threshold used for error correction.

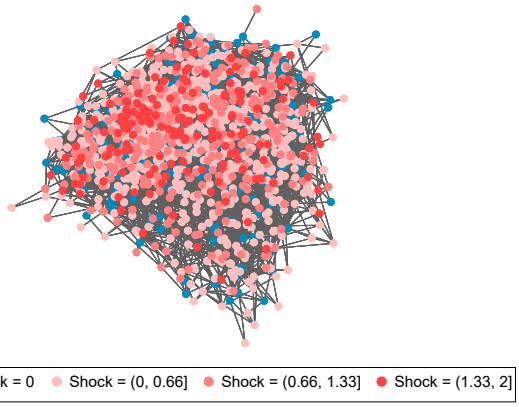
Figure 9: Illustration of data generation process: exogenous shocks in coauthorship networks

Full Sample

(a) Productivity shocks

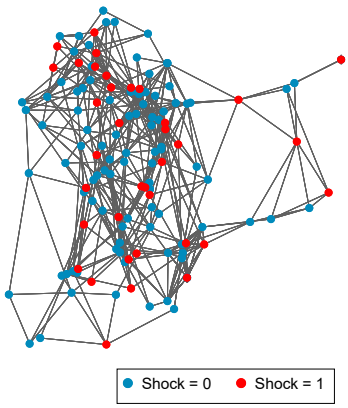


(b) Productivity shocks with network correlation

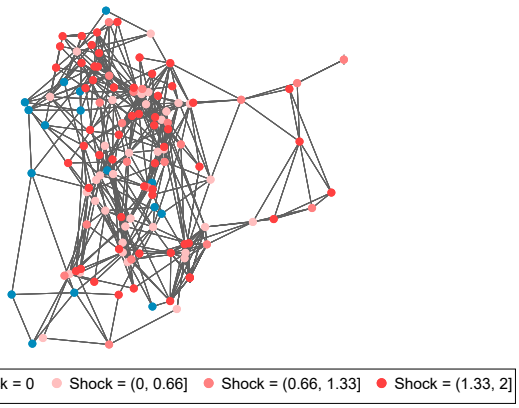


Subsample of the 250 most cited authors

(c) Productivity shocks



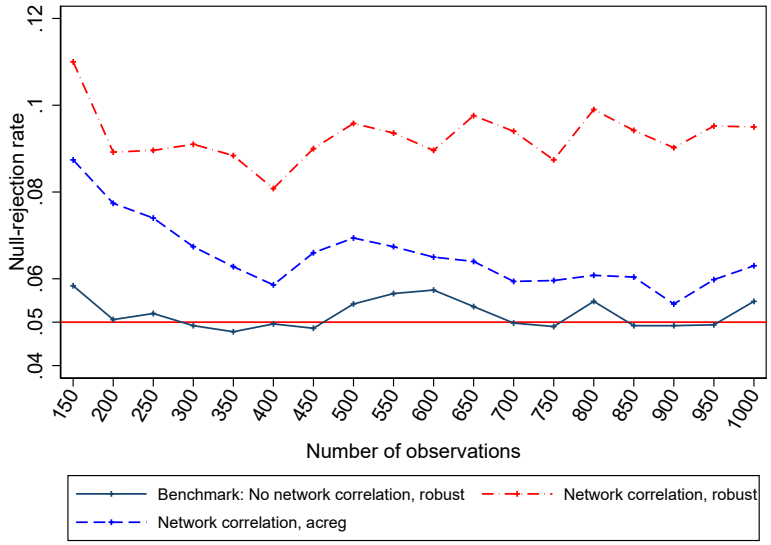
(d) Prod. shocks with network correlation



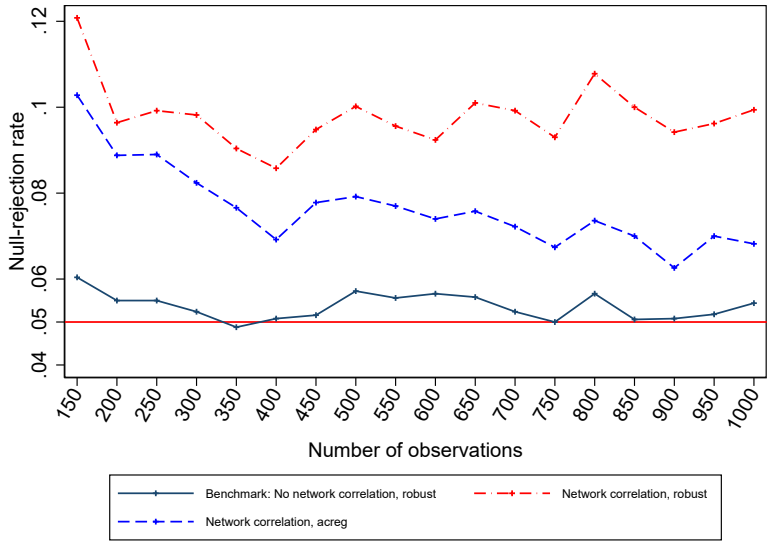
Notes: The figure maps the coauthorship links between authors. The sample consists of the authors indexed in the "Top 5% Authors, Number of Citations, as of October 2019", list of IDEAS RePEc (N = 1,637). The values of the exogenous productivity shocks represented are randomly generated with the algorithm described in section 3.2.1. Panels (a) and (b) refer to the full sample, panels (c) and (d) refer to a subsample of the top 250 authors in terms of number of citations.

Figure 10: Null-rejection rate in the presence of network correlation: Top-cited authors

(a) OLS



(b) 2SLS



Notes: The red vertical line represents the benchmark null-rejection rate of 5%. The vertical axis represents the rejection rate of the average null-rejection over 10,000 Monte Carlo simulations. Each point in the figure represents a different Monte Carlo simulation × estimation pair. The horizontal axis represents the sample size.

Table 1: Null-rejection rates in the spatial setting

Unit:					U.S. counties		
Sample:					All	Within-state	Cross-state
Sample size:					N=3,141	N=2,126	N=1,015
Data Generating Process			**Estimation**				
	Endogeneity	**Spatial corr.**	**Estimator**	**Correction**	**Null-rejection rate**		
					(1)	(2)	(3)
Panel A: Baseline Model							
(1)			OLS	robust	5.2%	5.1%	5.1%
(2)		✓	OLS	robust	9.1%	8.2%	10.4%
(3)		✓	OLS	cluster	6.8%	6.9%	9.2%
(4)		✓	OLS	acreg	5.5%	5.8%	5.6%
Panel B: Endogeneity							
(5)	✓		2SLS	robust	5.1%	5.1%	5.0%
(6)	✓	✓	2SLS	robust	9.0%	8.2%	10.1%
(7)	✓	✓	2SLS	cluster	6.6%	6.9%	8.8%
(8)	✓	✓	2SLS	acreg	5.3%	5.5%	5.6%

Note: This table reports the average null-rejection at the 5% level for Monte Carlo simulation experiments for different environments and sample sizes. The number of replications is 10,000 for each simulation. Panel A refers to a model with no endogeneity in which the β coefficients are estimated with OLS; Panel B refers to a model with endogeneity in which the β coefficients are estimated with 2SLS. Each column-row pair represents a different environment (data generating process and error correction) and sample pair. The outcome variable is log median earnings. In column 1 the whole sample is used; in column 2 only counties that are not at the state-border are considered; in column 3, we use only counties at the border. The data generating process simulates two different models: a baseline model without any spatial correlation in the policy treatment variable across units and another model imposing a spatial correlation in the policy treatment variable among the units within an arbitrary cluster. Each row indicates the model and the way we estimate it. Unit of observation is U.S. counties.

Table 2: Null-rejection rates in the spatial setting: Spatial correlation in the outcome

Unit:					U.S. counties, N=3,141		
Spatial correlation in the outcome:					Observed	Random	Fake
	Data Generating Process		**Estimation**				
	Endogeneity	**Spatial corr.**	**Estimator**	**Correction**	**Null-rejection rate**		
					(1)	(2)	(3)
Panel A: Baseline Model							
(1)			OLS	robust	5.2%	5.5%	4.8%
(2)		✓	OLS	robust	9.1%	5.1%	8.8%
(3)		✓	OLS	cluster	6.8%	6.1%	6.2%
(4)		✓	OLS	acreg	5.5%	5.2%	5.1%
Panel B: Endogeneity							
(5)	✓		2SLS	robust	5.1%	5.5%	4.7%
(6)	✓	✓	2SLS	robust	9.0%	5.0%	8.7%
(7)	✓	✓	2SLS	cluster	6.6%	5.7%	5.4%
(8)	✓	✓	2SLS	acreg	5.3%	5.0%	4.9%

Note: This table reports the average null-rejection at the 5% level for Monte Carlo simulation experiments for different environments and sample sizes. The number of replications is 10,000 for each simulation. Panel A refers to a model with no endogeneity in which the β coefficients are estimated with OLS; Panel B refers to a model with endogeneity in which the β coefficients are estimated with 2SLS. Each column-row pair represents a different environment (data generating process and error correction) and different outcome. The outcome variable in column 1 is the observed log median earnings; in column 2, the outcome variable is the observed log median earnings randomly reshuffled across counties; in column 3, we impose spatial correlation to the randomly shuffled log median earnings used in column 2. The data generating process simulates two different models: a baseline model without any spatial correlation in the policy treatment variable across units and another model imposing a spatial correlation in the policy treatment variable among the units within an arbitrary cluster. Each row indicates the model and the way we estimate it. Unit of observation is U.S. counties.

Table 3: Null-rejection rates in the spatial setting: Controls

Unit:					U.S. counties, N=3,141		
Controls:					Baseline	No controls	State FEs
Data Generating Process			**Estimation**				
	Endogeneity	**Spatial corr.**	**Estimator**	**Correction**	**Null-rejection rate**		
					(1)	(2)	(3)
Panel A: Baseline Model							
(1)			OLS	robust	5.2%	4.9%	5.1%
(2)		✓	OLS	robust	9.1%	12.8%	8.4%
(3)		✓	OLS	cluster	6.8%	7.2%	6.2%
(4)		✓	OLS	acreg	5.5%	5.7%	5.6%
Panel B: Endogeneity							
(5)	✓		2SLS	robust	5.1%	4.9%	5.0%
(6)	✓	✓	2SLS	robust	9.0%	12.7%	8.3%
(7)	✓	✓	2SLS	cluster	6.6%	6.8%	5.8%
(8)	✓	✓	2SLS	acreg	5.3%	5.6%	5.5%

Note: This table reports the average null-rejection at the 5% level for Monte Carlo simulation experiments for different environments and sample sizes. The number of replications is 10,000 for each simulation. Panel A refers to a model with no endogeneity in which the β coefficients are estimated with OLS; Panel B refers to a model with endogeneity in which the β coefficients are estimated with 2SLS. Each column-row pair represents a different environment (data generating process and error correction) and models. The outcome variable is log median earnings. The model in column 1 is the baseline one; in column 2, controls are omitted; in column 3, state Fixed effects are added to the regression. The data generating process simulates two different models: a baseline model without any spatial correlation in the policy treatment variable across units and another model imposing a spatial correlation in the policy treatment variable among the units within an arbitrary cluster. Each row indicates the model and the way we estimate it. Unit of observation is U.S. counties.

Table 4: Null-rejection rates in the network setting

Unit:					Authors, N = 1,637				
Network correlation in the outcome:					Observed	Random	Observed	Observed	Observed
Data Generating Process			**Estimation**						
	End.	**Network corr.**	**Estimator**	**Correction**	**Null-rejection rate**				
					(1)	(2)	(3)	(4)	(5)
Panel A: Baseline Model									
(1)			OLS	robust	4.8%	5.1%	4.8%	5.5%	4.7%
(2)		✓	OLS	robust	9.8%	5.2%	9.3%	8.9%	10.3%
(3)		✓	OLS	cluster, affiliation	9.6%	5.2%	9.4%	9.0%	9.9%
(4)		✓	OLS	cluster, affiliation city	10.4%	5.6%	10.1%	9.8%	10.7%
(5)		✓	OLS	cluster, degree school	11.0%	6.7%	10.9%	10.2%	11.3%
(6)		✓	OLS	cluster, degree city	12.0%	7.2%	12.1%	11.3%	12.6%
(7)		✓	OLS	acreg	5.6%	5.3%	5.6%	6.2%	5.4%
Panel B: Endogeneity									
(8)	✓		2SLS	robust	4.8%	5.0%	4.7%	5.5%	4.6%
(9)	✓	✓	2SLS	robust	9.6%	5.0%	9.2%	8.8%	10.3%
(10)	✓	✓	2SLS	cluster, affiliation	9.6%	5.0%	9.3%	8.9%	10.0%
(11)	✓	✓	2SLS	cluster, affiliation city	10.6%	5.4%	10.4%	10.2%	11.1%
(12)	✓	✓	2SLS	cluster, degree school	11.2%	6.5%	11.1%	10.2%	11.6%
(13)	✓	✓	2SLS	cluster, degree city	12.2%	7.0%	12.3%	11.3%	12.8%
(14)	✓	✓	2SLS	acreg	5.7%	5.1%	5.7%	6.2%	5.5%
Affiliation country					No	No	Yes	No	No
Degree school					No	No	No	Yes	No
PhD obtention year					No	No	No	No	Yes

Note: This table reports the average null-rejection at the 5% level for Monte Carlo simulation experiments for different environments and sample sizes. The number of replications is 10,000 for each simulation. Panel A refers to a model with no endogeneity in which the β coefficients are estimated with OLS; Panel B refers to a model with endogeneity in which the β coefficients are estimated with 2SLS. Each column-row pair represents a different environment (data generating process and error correction) and models. The outcome variable in columns 1,3,4,5 is the observed log citation score; in column 2, the outcome variable is the observed log citation score randomly reshuffled across authors. In columns 1,2, no additional covariates are added to the model; in columns 3,4,5 we controls for two distinct sets of covariates separately. The data generating process simulates two different models: a baseline model without any network correlation in the productivity treatment variable across units and another model imposing a network correlation in the productivity treatment variable among the units within an arbitrary cluster. Each row indicates the model and the way we estimate it. Unit of observation is authors indexed in the "Top 5% Authors, Number of Citations, as of October 2019", list of IDEAS RePEc.