MANAGEMENT SCIENCE

Network Effects in Contagion Processes: Identification and Control

Journal:	*Management Science*
Manuscript ID	MS-17-02880
Manuscript Type:	Operations Management
Keywords:	Networks-Graphs, Economics : Econometrics, Information systems: IT Policy and Management, Network economics
Abstract:	In this paper, we study the problem of identifying network effects in contagion processes and present an application to the propagation of influenza in the United States. In particular, using data on the evolution of infections over time, the travel intensity between states as well as environmental conditions we first provide a framework to identify the true network effect of traveling between states. Any identification strategy in this context needs to handle the following challenges: the reflection problem and the time correlation problem. The reflection problem arises from the observation that when sampling from the contagion process is frequent (in our case, weekly), the (potential) endogenous network effect cannot be discriminated from the correlation effect (such as that due to similar environmental conditions). The time-correlation effect stems from the observation that contagion processes are naturally characterized by correlation across different lags. We propose an instrumental variable approach, based on a spatiotemporally lagged versions of the observed data, and we show that our approach effectively tackles the aforementioned issues both theoretically and through a series of robustness checks. Finally, we use our estimates to propose and evaluate the performance of intervention and control policies, illustrating the benefits of network-based interventions.

Network Effects in Contagion Processes: Identification and Control

In this paper, we study the problem of identifying network effects in contagion processes and present an application to the propagation of influenza in the United States. In particular, using data on the evolution of infections over time, the travel intensity between states as well as environmental conditions we first provide a framework to identify the true network effect of traveling between states. Any identification strategy in this context needs to handle the following challenges: the reflection problem and the time correlation problem. The reflection problem arises from the observation that when sampling from the contagion process is frequent (in our case, weekly), the (potential) endogenous network effect cannot be discriminated from the correlation effect (such as that due to similar environmental conditions). The time-correlation effect stems from the observation that contagion processes are naturally characterized by correlation across different lags. We propose an instrumental variable approach, based on a spatiotemporally lagged versions of the observed data, and we show that our approach effectively tackles the aforementioned issues both theoretically and through a series of robustness checks. Finally, we use our estimates to propose and evaluate the performance of intervention and control policies, illustrating the benefits of network-based interventions.

1. Introduction

Contagion phenomena frequently dominate news headlines and public discussion, including the propagation of information [1] or misinformation [2], viral marketing and product adoption through word of mouth [3], the spread of computer viruses [4], the diffusion of innovations [5], financial contagion [6, 7], supply chain disruptions [8, 9] and, more traditionally, epidemics[1] and pandemics[2] at the national or international level. Due to the severity of these phenomena, online platforms, government officials, public health and pharmaceutical industry professionals, policy makers, and infectious-disease researchers increasingly need to understand the transmission dynamics, make better predictions, and design effective intervention policies. The significance of such contagion phenomena has led to extensive work on modeling their evolution and on understanding the resulting dynamics [10, 11, 12]. The main characteristic of these models is the presence of an underlying

[1] https://www.cdc.gov/flu/about/disease/burden.htm

[2] https://www.cdc.gov/vhf/ebola/outbreaks/2014-west-africa/index.html

contact network. Depending on the context, the network may represent contacts between individuals [13], influence among them [14, 15], or influence among different blogs in the blogspace [1, 3, 16]. Building on these models, a rich line of work on designing intervention policies to control (typically minimize or maximize) the effect of contagion processes has been developed. Network-based interventions have been proposed in the literature, primarily involving the design of techniques to identify central nodes in a static [17, 18, 19, 20] or dynamic [21, 22] manner. Both of these research directions assume that the magnitude of the network effect is known. However, both the modeling predictions and the policy designs greatly depend on the magnitude of the network effect, making the accurate estimation of the latter of utmost importance.

In this paper, we propose a method for estimating network effects for contagion processes, and we evaluate intervention strategies for their effective containment using the estimation result. The application explored in this paper is that of influenza, in the context of which network effects arise due to traveling of infected individuals between different states in the United States. The methodology developed, however, can be applied to any contagion process, as long as the dynamics are driven by a combination of endogenous and exogenous factors.

Influenza, commonly known as "the flu", is an infectious disease caused by an influenza virus [23]. Symptoms can be mild to severe. The most common symptoms include high fever, runny nose, sore throat, muscle pains, headache, coughing, and fatigue. These symptoms typically begin two days after exposure to the virus and usually last less than a week. While the impact of flu varies, overall, it places a substantial burden on the health of people in the United States each year. In fact the Center for Disease Control (CDC) estimates that influenza has resulted in between 9.2 million and 35.6 million illnesses, between 140,000 and 710,000 hospitalizations, and between 12,000 and 56,000 deaths annually since 2010. The transmission dynamics of influenza are mostly determined by the following three factors:

(i) **Human contact:** People with the flu can spread it to others up to about 6 feet away. Most experts think that flu viruses are spread mainly by droplets emitted when people with flu cough, sneeze, or talk. These droplets can land in the mouths or noses of people who are nearby or can possibly be inhaled into the lungs. Less often, a person might also get the flu by touching a surface or object that has flu virus on it and then touching their own mouth or nose.[3]

(ii) **Vaccination:** Immunization with an annual seasonal flu vaccine is the best way to reduce the risk of getting sick with seasonal flu and spreading it to others. When more people get vaccinated against the flu, the infection spread through a community will be more limited.

[3] Center for Disease Control and Prevention: https://www.cdc.gov/flu/about/disease/spread.htm

The size of the vaccination coverage among the adult population during the 2015-2016 season was 41.7%, while among the child population was 59.3%, in the United States.

(iii) **Environmental conditions:** Recent advances in epidemiology literature (see [24] and references therein) show that the transmission dynamics of the influenza viruses greatly depend on absolute humidity, explaining the seasonal patterns of influenza.

These three mechanisms affect the transmission dynamics simultaneously, and therefore, identifying their separate effects on the propagation of the disease, although challenging, as we will shortly see, has significant implications both for the understanding of contagion processes and the designing of intervention policies to minimize their impact.

More rigorously, estimating causal network effects in contagion processes is challenging for two reasons. First of all, since the processes are dynamic, properly dealing with the serial correlation across observations in the analysis is important. Most common regression methods such as ordinary least squares (OLS) require that the observations are independent and identically distributed, which is not applicable in such a dynamic setting. Therefore, one needs to resort to panel data methods in econometrics, which provide techniques for properly accounting for the diffusion of the processes.

Second, estimating the network effects is challenging even in a static environment. This problem was first introduced by [25] in the setting of estimating peer effects among students in the same class. The general idea is as follows. If we are interested in estimating the network effect between two nodes i and j, the most straightforward way is to regress the outcome observed at i, e.g., number of infected population, on the outcome observed at j. However,there are two problems with this approach. First, there is the usual endogeneity issue: j affects i, but i also affects j. Regressing the outcome of one on the other is not recovering the causal network effect of interest. Second, and more importantly, we cannot separately identify the network effect of interest from the correlation between the determinants of the outcome variable. For example, if we observe that the infected populations in New York and New Jersey are highly correlated, we cannot separate the effect of people traveling between the two states from the effect of the two states having similar environmental conditions. Proper handling of these two challenges leads to accurate identification of network effects, which, on top of the methodological interest, has significant implications on designing intervention policies.

The two principal strategies for containing serious human outbreaks of influenza are therapeutic countermeasures (e.g., vaccines and antiviral medications), from now on referred to as *medical interventions*, and public health interventions (e.g., infection control, social separation, and quarantine), from now on referred to as *network interventions*. Vaccination and, to a lesser extent,

antiviral medication are perhaps the most important *medical interventions* for reducing the morbidity and mortality associated with influenza. In fact, these approaches are considered so effective that the United States devotes over 90% of pandemic influenza spending to medical interventions [26]. The country's policy and strategic plan against influenza to a great extent revolves around increasing the production capacity of vaccines in order to ensure timely satisfaction of the demand [27].

Medical intervention is clearly effective when a vaccine is abundantly available and vaccine coverage is high: in the extreme case where the whole population is vaccinated at the beginning of the season, the possibility of a pandemic would be minimal if not obsolete. Unfortunately, such a scenario is not realistic: the strain of the virus is different among different seasons, and thus new vaccines need to be produced after the population has been exposed to the virus leading to a mismatch between public health needs and private-sector production capabilities. This drawback of medical intervention necessitates the development of different approaches for the containment and control of contagion processes.

Since the mechanism of the transmission of contagion processes is based on interactions between humans, decreasing social mixing (or increasing social distance) is another intervention approach that has been effective on a smaller scale. For example, when societies are faced with pandemics, public places such as schools and mass transit hubs may be closed, or high-risk individuals, such as teachers, doctors and other health professionals, may be distanced from from social or professional interactions. On a larger scale, the approach of decreasing interaction may take the form of quarantining individuals [28], groups [29], or whole geographic regions [30]. Such approaches are based on the principle of reducing the network interactions among humans, groups, or regions, hence decreasing the effective infection rate of the disease. These approaches, although hard to implement, have the significant advantage of being effective immediately (in contrast to the production delays of medical interventions) and independently of the strain of the virus. Clearly, however, such approaches have several disadvantages related to practical and ethical issues; these are extensively discussed in the public health literature [31].

Understanding and comparing the efficacy of medical versus network interventions is crucial when policy makers are making budgeting and planning decisions. The trade-offs are closely related to the intensity of the network effects as we shall shortly see, underlining the practical relevance of accurately measuring the latter.

1.1. Main Results

Our contributions in this paper are threefold:

(i) *Methodological Contributions:* We develop identification methods to simultaneously tackle the two identification challenges mentioned above: time correlation and network endogeneity. In particular, using the network structure as well as past measurements of exogenous factors (absolute humidity in the influenza application), we construct appropriate instrumental variables and show how they can be used to identify network effects in a contagion process. We should emphasize that our methodology is not tied to the specific application of influenza or epidemics; rather it is applicable whenever observational panel data on a contagious phenomenon that is driven by a combination of endogenous and exogenous factors are available.

(ii) *Practical Insights:* Applying our methodology to our dataset, consisting of 6 years of weekly data on influenza-related infections, pairwise travel intensities, vaccine coverage, environmental factors, health indices, gas prices, and household incomes, we identify the network (traveling) effects on the propagation of influenza. We discover that traveling between states has a significant impact on the transmission of the virus in the country, leading to an aggregate effect of $10\% - 110\%$ of the effect of in-state infections.

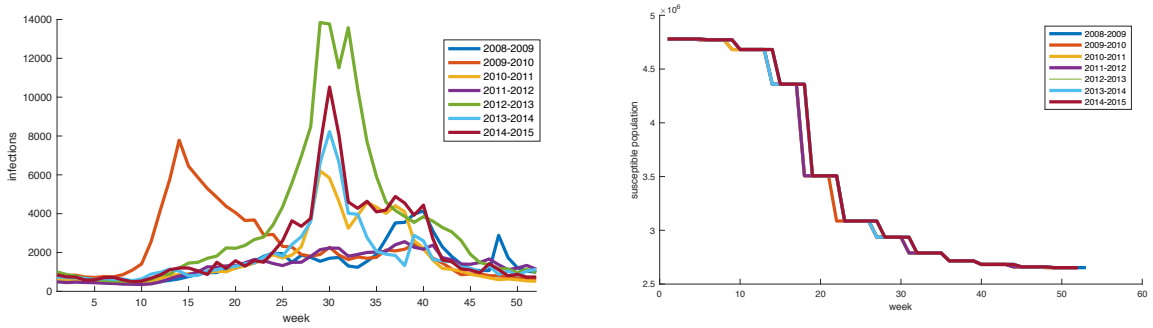
(iii) *Policy Implications:* Using our estimates, we are able quantify the efficacy of different interventions. In particular, we show that targeted medical interventions (increasing the vaccine availability in specific "central" states) can lead to substantial improvement in terms of the number of infections compared to the currently adopted uniform increase of vaccine availability. Furthermore, we identify the set of states as well as the set of state pairs where network intervention (travel monitoring and regulation) would lead to the largest decrease in infection levels.

1.2. Related Literature

The problem of identifying network effects was first introduced by [25] in the context of studying peer effects in classrooms. In general, depending on the context, several sources of bias have been identified including contextual and correlated effects [25, 32], simultaneity and other time-related factors [33, 34], and homophily [35]. The approaches adopted in the literature to correct these biases and accurately estimate network effects can be broadly categorized as follows:

(i) *Randomized experiments:* Appropriately designed experiments [36] have been proposed and used in the literature to identify network effects in different contexts: [37, 38] study peer effects on educational outcomes, [39] studies the effect of word of mouth on product virality, [35] studies the effect of social influence on knowledge propagation, and belief formation and [40] studies network effects on product demand.

(ii) *Inference from observations:* [32, 41, 42] and others have developed different methods to solve the identification problem in the setting of social network effects using observational data.



(a) Number of influenza-related infections in Alabama for all seasons. The single-peaked shape of the time series is typical in the dataset.

(b) Size of susceptible population in Alabama for all seasons.

Figure 1

Typically, these studies use instrumental variables to remove the endogeneity and obtain unbiased estimates of the network effects. Several applications have been explored, including studies of network effects between different products [43], social effects on behavior [44], and viral marketing [45].

In our context, where the effect of inter-state travel on the propagation of influenza is under study, deploying a randomized experiment would be impractical, if not infeasible. Instead, we extend the work of [46, 25, 32] and use observational data to obtain estimates, but extend existing literature by showing that the identification of network effects can be achieved in dynamic settings with panel data.

The rest of this paper is organized as follows. In Section 2 we present our dataset. In Section 3 we present our model an identification technique. In Section 4 we present our empirical findings while in Section 5 we present the policy implications. Finally, in Section 6 we present the conclusions of this paper an identify directions of future research.

2. Data

In this section, we describe the source and nature of the data that we use for our analysis. We present the different elements of data in order of appearance in our model.

Infection Data

In an attempt to provide faster detection and more detailed reporting, innovative surveillance systems have been created to monitor indirect signals of influenza activity, such as call volume to telephone triage advice lines and over-the-counter drug sales. About 90 million American adults are believed to search online for information about specific diseases or medical problems each

year, making web search queries a uniquely valuable source of information about health trends. Google created a system [47] that builds on this observation by utilizing an automated method of discovering influenza-related search queries. By processing hundreds of billions of individual searches from five years of Google web search logs, this system generates comprehensive models for use in influenza surveillance, providing regional and state-level estimates of influenza-like illness (ILI) activity in the United States. In the present work, we use the dataset provided by the Google Flu Trends tool, which consists of *weekly* estimates of influenza related infections for all 50 states over 8 seasons, namely, 2008-2015. Although an early version of this tool produced data with low accuracy on rare occasions [48], multiple studies have found that the overall accuracy level of the data to be high [49, 50, 51]. We use the most recent version of the tool introduced in 2014 which improves on the previous versions and has been shown to have high accuracy [52].

A typical snapshot of this dataset for a specific state during a specific season is shown in Figure 1(a). Note that we *exclude* from the dataset the season 2009-2010 due to the H1N1 influenza virus outbreak, which featured non-typical behavior (see Figure 1(a)), as well as the season 2008-2009 due to the a lack of enough samples. We denote by $\hat{I}_i(t)$ the number of infections, as provided by this tool.

Environmental Conditions

As Figure 1(a) illustrates, influenza spreads around the world in yearly outbreaks, resulting in about three to five million cases of severe illness and about 250,000 to 500,000 deaths per year. In the northern and southern parts of the world, outbreaks occur mainly in the winter, while in areas around the Equator, outbreaks may occur at any time of year. Previous studies indicate that environmental factors may affect such seasonality patterns, and have analyzed data from laboratory experiments to explore the effects of different parameters on influenza virus transmission and influenza virus survival. Recent studies (see [24] and references therein) find that absolute humidity constrains both transmission efficiency and, more significantly, influenza virus survival. In particular, in these studies, 50% of influenza virus transmission variability and 90% of influenza virus survival variability are explained by absolute humidity. In line with these experimental findings, in temperate regions, both outdoor and indoor absolute humidity possesses a strong seasonal cycle, with the lowest values occurring during the winter. This seasonal cycle is consistent with a wintertime increase in influenza virus transmission and influenza virus survival and can be used to explain the seasonality of influenza. Thus, differences in absolute humidity provide a single, coherent, physically sound explanation for the observed variability of influenza seasonality. We use meteorological data from the National Oceanic and and Atmospheric Administration[4] and use

[4] https://www.ncdc.noaa.gov/data-access/land-based-station-data/land-based-datasets

$\hat{H}_i(t)$ to denote the absolute humidity in state i at time t. We should note that the dataset, consists of daily samples from 1600 stations across the country and that $\hat{H}_i(t)$ is the average measurement over the corresponding week across all stations in state i.

Susceptible Population

Immunity may develop after an individual takes a flu vaccine, which causes antibodies to be produced in the body about two weeks after vaccination. These antibodies provide protection against infection with the viruses that are covered by the vaccine. Understanding the evolution over time of the vaccinated population is crucial to understanding the evolution of the susceptible population in each state. CDC provides weekly estimates of the fraction $\hat{V}_i(t)$ of the non-vaccinated population for all seasons of interest through the FluVaxView[5] tool. Using the latter, as well as estimates on the population of each state $\hat{N}_i$ from the 2010 Census data[6] we write

$$\hat{S}_i(t) = \hat{N}_i \hat{V}_i(t)$$

Note here that we exclude from our analysis the number of infected individuals who recovered from the flu and developed immunity. This simplifying assumption is further justified by observing that whereas the size of the susceptible population is in the order of millions, the number of recovered individuals is in the order of thousands and therefore can be safely omitted without affecting our results.

Travel Intensity

In order to obtain an estimate of the travel intensity between different pairs of states, we use data from the National Household Travel Survey (NHTS)[7]. This is the flagship survey of the U.S. Department of Transportation (DOT) and is conducted periodically to assess the mobility of the American public. The survey specifically gathers trip-related data, such as mode of transportation, duration, distance, and purpose, and then links the travel-related information to demographic, geographic, and economic data for analysis. The NHTS survey data that we used for our analysis were collected in 1990, 1995, 2001, 2009, and 2014. Each survey covers about 25,000 households representing all 50 States and the District of Columbia. During the survey period, each household was sent a travel diary and asked to report all travel by household members. For the purposes of our analysis, we identified for each (directed) pair of states the number of individuals traveling

[5] https://www.cdc.gov/flu/fluvaxview/interactive-general-population.htm

[6] https://www.census.gov/2010census/data/

[7] http://nhts.ornl.gov/download.shtml

from state to state, after normalizing by the population size. Specifically, our calculation for the travel intensity is equal to

$$A_{ij} = \frac{\hat{N}_{ji}}{\hat{D}_j}\hat{N}_j,$$

where $\hat{N}_{ij}$ is the total number of individuals (in the dataset) in the sample traveling from state i to state j, $\hat{D}_i$ is the sample size for state i, and $\hat{N}_i$ is the population of state i. The reader should note that the entry A_{ij} corresponds to the flow *from* state j into state i. We use this convention in order to make the notation in the following sections simpler.

3. Model and Estimation

In this section, we present the model that we will be using for the remainder of this paper. We first introduce the baseline model, which is the simplest one that allows us to obtain interpretable and robust results. Then, we discuss the identification and estimation of the model, and provide a series of improvements to the baseline model, to underline the robustness of our findings.

3.1. Baseline Model

We consider each of the 50 states of the United States as being represented by a node on a weighted directed graph G. We use A to denote the travel matrix of the graph, where each entry A_{ij} is proportional to the intensity of traveling from state $i \in V$ to state $j \in V$. We denote by

$$E = \{(i,j) : \ A_{ij} > 0\},$$

the set of edges of the underlying network on which there is non-zero travel intensity. Moreover, we denote by

$$Y_i(t) = \log(\hat{I}_i(t)),$$

the logarithm of the number of infected individuals in state i at time t, by

$$S_i(t) = \log(\hat{S}_i(t)),$$

the logarithm of the number of susceptible individuals in state i at time t, and by

$$H_i(t) = \log(\hat{H}_i(t)),$$

the logarithm of the absolute humidity in state i at time t.

Note that we take the log transformation for all quantities for the following reason: the growth of the infected population is likely to follow an exponential pattern and taking logarithms makes

the relationship closer to a linear one thus providing a stationary environment for the regression analysis, which is key for the analysis to be valid. Moreover, since all variables in the model are positive, taking the log transformation removes the restriction on the value range of those variables and makes the linear relationship a more reasonable approximation. Note that this is common practice in the econometrics literature when the dependent variable of interest is growing over time or when some variables in the analysis only take on positive values [53].

Using standard notation, whenever we omit the index i from any of the aforementioned objects, we denote the 50×1 vector, whose entries are the values for each corresponding state. With this notation in mind, our baseline model for the propagation of the flu can be written as

$$Y_i(t+1) = \beta_0 + \beta_1 H_i(t) + \beta_2 S_i(t) + \beta_3 Y_i(t) + \beta_4 A_i \cdot Y(t) + r_i + \nu_i(t), \tag{1}$$

where the operator $\cdot$ denotes the inner product and A_i denotes the i-th row of the travel matrix A. Equation (1) explicitly models the dependence of the infection process on absolute humidity, $H_i(t)$ (exogenous effect), the number of susceptible individuals $S_i(t)$ (state-level characteristic), the number of infected individuals within the state $Y_i(t)$, and the number of infected individuals who travel from all neighboring states j, with $(j,i) \in E$ (network effect). Note that we add an unobserved, state-specific effect r_i to control for the various unobserved demographic, environmental, and economic inhomogeneities across states, which would potentially affect the evolution of the flu. For the baseline model, we assume that r_i for all $i \in V$ is unknown. Later, in Section 4, we include additional control variables, such as the Health Index data as described in Section 2.

Throughout the paper, we uppercase all variables that are *observable* to us in the data, and we lowercase all variables that are unknown to us (either unobserved, such as the noise $\nu_i(t)$, or to be estimated, such as the parameters $\beta_1, \beta_2, \beta_3, \beta_4$).

3.2. Estimation

Correctly estimating the unknown parameters in (1) is challenging, mostly due to the difficulty of identifying the network effect, β_4. This difficulty is in turn due to the following two effects that act simultaneously:

(i) *time correlation:* The variables $Y_i(t)$ and $S_i(t)$ in (1), corresponding to the infected population and the susceptible population respectively are naturally correlated over time.

(ii) *spatial correlation:* If two states are connected through the network and have similar infection patterns, this can be explained by either the causal effect of traveling between them, or by the similarity in the patterns between the explanatory variables such as absolute humidity $H_i(t)$ or the state-level characteristics $S_i(t)$.

In order to clearly illustrate the two effects and explain the proposed estimation method, we start by decomposing the two problems: we first explain the identification and propose a solution to the *time correlation* issue without the network structure, and we then explain the *spatial correlation* issue in a static model without the *time correlation* issue. Finally, we discuss the identification and estimation of the network effect in the presence of both *time correlation* and *spatial correlation*.

3.2.1. Identification of dynamic model without network effect Suppose, for the purposes of this discussion, that the nodes on the network are isolated, i.e., there is no travel between states. Equation (1) becomes

$$Y_i(t+1) = \beta_0 + \beta_1 H_i(t) + \beta_2 S_i(t) + \beta_3 Y_i(t) + r_i + \nu_i(t). \tag{2}$$

This is the classic panel data regression setting. Following the common practice in the literature, we take the difference between $Y_i(t+1)$ and $Y_i(t)$ [46, 54]. This allows us to difference out any unobserved state-specific effect captured by r_i, as well as the time trend in $\nu_i(t)$. After taking the difference, Equation (2) becomes

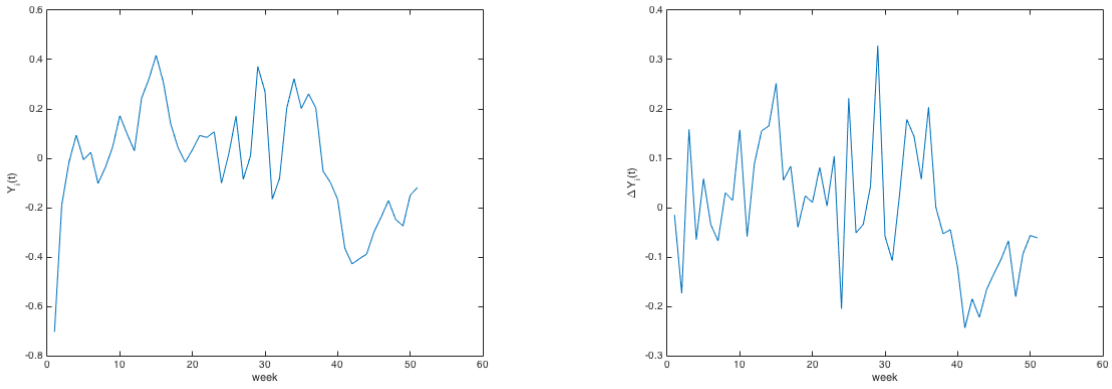
$$\Delta Y_i(t+1) = \beta_1 \Delta H_i(t) + \beta_2 \Delta S_i(t) + \beta_3 \Delta Y_i(t) + \varepsilon_i(t), \tag{3}$$

where $\Delta Y_i(t) = Y_i(t) - Y_i(t-1)$, $\Delta H_i(t) = H_i(t) - H_i(t-1)$, $\Delta S_i(t) = S_i(t) - S_i(t-1)$, and $\varepsilon_i(t) = \nu_i(t) - \nu_i(t-1)$.

An alternative approach to taking the difference is to include fixed effects and estimate r_i. However, this imposes stronger independence assumptions on the error term $\nu_i(t)$ if we would like to obtain consistent estimates of the parameters in the model. By taking the differences, we are able to relax this assumption. We plot $Y_i(t)$ and $\Delta Y_i(t)$ in Figure 2(a) and Figure 2(b), respectively, to support our modeling choice. These figures show that $Y_i(t)$ exhibits strong serial correlation, while $\Delta Y_i(t)$ more closely resembles an independent identically distributed sample. We include the one period lag $\Delta Y_i(t)$ as an explanatory variable to account for the residual serial correlation.

To recover unbiased estimates for the unknown parameters, simply regressing $\Delta Y_i(t+1)$ on the $\Delta H_i(t)$, $\Delta S_i(t)$, and $\Delta Y_i(t)$ is not sufficient. There are two endogenous variables in the equation, $\Delta S_i(t)$ and $\Delta Y_i(t)$, i.e., $E[\varepsilon_i(t)\Delta S_i(t)] \neq 0$, and $E[\varepsilon_i(t)\Delta Y_i(t)] \neq 0$. Intuitively, $\varepsilon_i(t)$ captures the determinant of newly infected individuals unobserved to the researcher in period t. Since both the infected population and the susceptible population in period $t-1$ affect the number of new infections in period t, $S_i(t-1)$ and $Y_i(t-1)$ are correlated with $\varepsilon_i(t)$. Therefore, in turn, $\Delta Y_i(t)$ and $\Delta S_i(t)$ are correlated with $\varepsilon_i(t)$.

In order to resolve this endogeneity issue, we observe that (3) can be solved as



(a) Serial correlation in $Y_i(t)$ (b) Independence of $\Delta Y_i(t)$

Figure 2

$$\Delta Y_i(t) = \sum_{j=1}^{t-2} \beta_3^{j-1}(\beta_1 \Delta H_i(t-j) + \beta_2 \Delta S_i(t-j) + \epsilon_i(t-j)) + \beta_3^{t-2} \Delta Y(2),$$

for all $t \geq 2$, hence establishing that the time-lagged values $\Delta H_i(t-1), \Delta H_i(t-2), \Delta H_i(t-3), \ldots$ of the humidity affect $\Delta Y_i(t)$. Note that *total population* $=$ *infected population* $+$ *susceptible population immune population*. Thus, the size of the infected population and susceptible population in the same period are correlated. This implies that the time-lagged humidity values also affect $\Delta S_i(t)$. Therefore, $\Delta H_i(t-1), \Delta H_i(t-2), \Delta H_i(t-3), \ldots$ satisfy the relevance condition to be valid instrumental variables for $\Delta Y_i(t)$ and $\Delta S_i(t)$.

Next, we show that the lagged humidity variables satisfy the exclusion restriction to be valid instrumental variables. First, humidity levels are exogenously determined and thus they are not affected by the spread of the flu. This implies that given the state specific effect r_i, any unobserved determinant of the infected population is mean independent of the history of the humidity levels. In other words, humidity is a "truly exogenous" variable in the standard panel data method introduced by [46]:

Assumption 1.

$$E[\nu_i(t)|H_i(-\infty), \cdots, H_i(\infty), r_i] = 0.$$

This is a strong assumption, but we believe that it is a reasonable one, given that the nature of the variation in humidity levels is random. Following [46], we take the differences between the equation in Assumption 1 and the same equation in period $t-1$: $E[\nu_i(t-1)|H_i(-\infty), \cdots, H_i(\infty), r_i] = 0$. We have

$$E[\varepsilon_i(t)|\Delta H_i(-\infty), \cdots, \Delta H_i(\infty)] = 0,$$

which implies that $\Delta H_i(-\infty), \cdots, \Delta H_i(\infty)$ satisfy the exclusion restriction.

To summarize, the time lagged humidity measures $\Delta H_i(t-k)$ where $k \geq 1$ satisfy

1. **relevance condition:** they affect $\Delta Y_i(t)$ and $\Delta S_i(t)$,
2. **exclusion restriction:** the humidity measures two or more periods before are not directly correlated with the size of infected population in the current period.

As a result, $\Delta H_i(t-k)$, where $k \geq 1$ are valid instrumental variables for $\Delta Y_i(t)$ and $\Delta S_i(t)$, and the unknown parameters in Equation (3) can be identified. As we will see in the discussion below, similar intuition for identification extends to the setting with the presence of network effects.

3.2.2. Identification of a static model with network effects In this section, we discuss the identification of a static model with network effects. To simplify the notation and for the purposes of this preliminary discussion, we ignore the dependence on time t in Equation (1). We also ignore the state specific effect r_i in this subsection since we can difference it out, as described in the previous section.

We start with a simple network with two nodes i and j. Equation (1) then becomes

$$Y_i = \beta_0 + \beta_1 H_i + \beta_2 S_i + \beta_4 Y_j + \nu_i. \tag{4}$$

To further simplify the discussion, we assume in this section that both H_i and S_i are exogenous determinants of Y_i. In other words, $E[\nu_i | H_i, S_i] = 0$. Taking the conditional expectation on both sides, we have

$$E[Y_i | H_i, S_i] = \beta_0 + \beta_1 H_i + \beta_2 S_i + \beta_4 E[Y_j | H_i, S_i],$$

or, in matrix form,

$$E[Y | H, S] = \beta_0 + \beta_1 H + \beta_2 S + \beta_4 E[Y | H, S].$$

When $\beta_4 \neq 1$, it is clear from the following equation that the network effect β_4 is not separately identified from the other parameters:

$$E[Y | H, S] = \beta_0/(1-\beta_4) + \beta_1/(1-\beta_4) H + \beta_2/(1-\beta_4) S.$$

This is a simple example of the "reflection problem" studied in [25]. The intuition of this result is as follows. First, if i and j are connected on the network, it is likely that H_i and H_j, or S_i and S_j are also correlated through the connection. As explained in detail in [25], any correlation between Y_j and Y_i could be explained by either the network effect we are interested in, namely, β_4, or the correlation between the determinants of Y_i and Y_j. Therefore, it is difficult to separately identify the network effect between two connected nodes on the network and the correlation between their respective determinants.

Second, there can be unobserved variables that affect Y_i and are also correlated with Y_j through the network. In other words, r_i and r_j in Equation (1) are correlated through the connection of i and j on the network. This correlation also leads to issues in the identification of the network effects that we are interested in estimating. Since this problem is easier to deal with in the panel data setting, we postpone the discussion on this issue to the next subsection, when dynamics are reintroduced into the model. Meanwhile, we focus on the first identification issue in this subsection.

In order to resolve the "reflection problem", we generalize the setting to networks with more than two nodes and show that the identification of the network effect can be achieved under certain conditions. We utilize the network structure and exogenous determinants of Y_i to identify the network effect. The main idea is similar to that in [32], where in a static setting of estimating peer effects on a social network, spatially "lagged" measurements of the exogenous determinants are valid instrumental variables for the network effect as long as the (binary) matrices G, G^2, and G^3 are not linearly dependent, where G denotes the (binary) travel matrix. If j is i's neighbor on the network, k is j's neighbor, but k is not i's neighbor, then the exogenous variable acting on agent k will not have a direct impact on Y_i but is correlated with it through Y_j. Therefore, since the exogenous variable acting on k is only affecting Y_i through Y_j, it is a valid instrument for the network effect of j on i.

Adapting this idea to the preliminary setting of this subsection, with more than two nodes on the network, Equation (4) becomes

$$Y_i = \beta_0 + \beta_1 H_i + \beta_2 S_i + \beta_4 A_i \cdot Y + \nu_i, \tag{5}$$

where again we assume $E[\nu_i | H_i, S_i] = 0$. We construct instrumental variables later to deal with the endogeneity problem of S_i. We first show that if certain conditions on the network structure A are satisfied, β_4 is identified. Rewriting Equation (5) in matrix format, we have

$$Y = \beta_0 + \beta_1 H + \beta_2 S + \beta_4 A \cdot Y + \nu. \tag{6}$$

If $(I - \beta_4 A)$ is invertible, we can write

$$Y = (I - \beta_4 A)^{-1}\beta_0 \mathbf{e} + (I - \beta_4 A)^{-1}\beta_1 H + (I - \beta_4 A)^{-1}\beta_2 S + (I - \beta_4 A)^{-1}\nu,$$

where $\mathbf{e}$ is an $N \times 1$ vector of ones. Since $(I - \beta_4 A)^{-1} = \sum_{k=0}^{\infty} \beta_4^k A^k$, we have

$$Y = (I - \beta_4 A)^{-1}\beta_0 \mathbf{e} + \beta_1 \sum_{k=0}^{\infty} \beta_4^k A^k H + \beta_2 \sum_{k=0}^{\infty} \beta_4^k A^k S + \sum_{k=0}^{\infty} \beta_4^k A^k \nu.$$

Then, we can write

$$E(A \cdot Y|S,H) = A(I-\beta_4 A)^{-1}\beta_0 \mathbf{e} + \beta_1 AH + \beta_2 AS + \beta_1 \sum_{k=1}^{\infty} \beta_4^k A^{k+1} H + \beta_2 \sum_{k=1}^{\infty} \beta_4^k A^{k+1} S, \quad (7)$$

if $E[\nu|H,S] = 0$. Therefore, the spatially lagged values $\{A^{k+1}H\}_{k\geq 1}$ of the humidity satisfy the following conditions

1. **relevance:** by Equation (7), they affect the network term $A \cdot Y$,
2. **exclusion restriction:** Since humidity condition is exogenous, $E[\nu|H] = 0$. If A, A^2, and A^3 are linearly independent, then the network effect is separately identified from the "exogenous effect". In [32], this condition is only satisfied for binary A with certain structures. In our setting, since A indicates travel intensity between states, this condition is generally satisfied. Therefore, A^2H and A^3H can be used as instrumental variables for AY. The intuition as follows: state i's neighbors' neighbors' exogenous determinant H affects i's infection level only through AY if the intensity of connection between two nodes on the network is generally different. The additional variation found in the neighbors' neighbor's humidity levels H is what we use to identify the network effect. Note that theoretically, the same argument applies to A^4, A^5, etc.[8]

The term AH Equation (7) captures what Manski refers to as "exogenous effect" in [25]. It can be interpreted as the effect of i's neighbors' infection level on Y_i due to the correlation between their humidity levels. To control for the exogenous effect, [32] includes the neighbors' covariates, such as AH in the current setting, in the regression Equation (6). In our setting, however, since the network A of interest is about travel patterns, it does not make sense to assume that humidity levels are correlated through travel. Instead, humidity levels are more likely to be correlated if two states are geographically adjacent to each other. Therefore, in the full specification of the model, we include the additional covariate BH, where the matrix B indicates the spatial adjacency of every pair of states. Similarly, we also include BS as a additional covariate.

Given that the size of the susceptible populations is correlated with the number of infected individuals and is correlated with the humidity condition, the same set of instruments A^2H, A^3H, etc. are also valid for S. We summarize this identification result in Proposition 1.

ASSUMPTION 2. $E[\nu|H] = 0$.

ASSUMPTION 3. *Matrices I, A, A^2, A^3 are linearly independent.*

PROPOSITION 1. Under Assumptions 2 and 3, A^2H and A^3H are valid instrumental variables for AY and S_i in Equation (6), and the unknown parameters in (6) are identified.

[8] In practice, we stop at A^3, as higher order terms prove to have a weaker correlation to AY.

With additional control variables BH and BS, Equation (6) becomes

$$Y = \beta_0 + \beta_1 H + \beta_2 S + \beta_4 A \cdot Y + \beta_5 BH + + \beta_6 BS + \nu, \tag{8}$$

and Equation (7) becomes

$$\begin{aligned} E(A \cdot Y | S, H) =& G(I - \beta_4 A)^{-1} \beta_0 \mathbf{e} + (\beta_1 A + \beta_5 AB) H + (\beta_2 G + \beta_6 AB) S \\ &+ (\beta_1 A + \beta_5 AB) \sum_{k=1}^{\infty} \beta_4^k A^k H + (\beta_2 A + \beta_6 AB) \sum_{k=1}^{\infty} \beta_4^k A^k S. \end{aligned} \tag{9}$$

By the same reasoning as above, all the cross terms, such as ABH, are also valid instrumental variables (under some linear independence assumptions), hence providing us with a richer set of instrumental variables and better estimation results. We summarize this result in Proposition 2.

Assumption 4. *Matrices I, A, B, AB, A^2B are linearly independent.*

Proposition 2. Under Assumptions 2 and 4, ABH and A^2BH are valid instrumental variables for AY and S_i in Equation (8), and the unknown parameters in (8) are identified.

Intuitively, the instrumental variables ABH and A^2BH exploit the differences between the two networks A and B to separately identify the network effect from the correlated effect that we discussed above. For example, if states i and j are adjacent, i and k are not adjacent, but j and k are connected through the travel network A. The instrument ABH measures the effect of H_k on Y_i through Y_j, i.e., the humidity level in state k affects the infection level in state j through their link on network A, and then affects Y_i through its link to j on network B. Since i and k are not directly linked on AB, the correlated effect does not contaminate the endogenous network effect.

This identification result utilizes the structure of both network A and network B. The additional information used in this identification method compared with the method in Proposition 1 makes the result more robust. Assumption 4 is also less demanding than Assumption 3 if network B contains different types of information than network A. We think that this identification result can be applicable in many other settings where understanding of the causal network effect is important.

3.2.3. Identification of dynamic models with network effects As we established in Sections 3.2.1 and 3.2.2, when estimating network effects from panel data, there are two different challenges that unbiased estimation faces: the endogeneity due to time correlation and the endogeneity due to network or spatial correlation. As discussed above, for both challenges, the solution requires the construction of instrumental variables using appropriately lagged measurements of the exogenous variable (humidity). Specifically, when dealing with the time endogeneity issue, the instruments are time-lagged measurements ($H(t-k-1)$) of the exogenous variable. Similarly, when

dealing with the network endogeneity, the instruments are network-lagged measurements ($A^{k+1}H$), hence uncovering a fundamental similarity between the two settings. In this section, we generalize these results to the dynamic setting with network effects by appropriately constructing instruments using time- and network-lagged measurements of the exogenous variable.

As explained in the previous section, in order to account for the potential correlations between the exogenous variables and susceptible populations in neighboring states, we add the covariates $BH(t)$ and $BS(t)$ in the main specification. In vector notation, we write

$$Y(t) = \beta_0 \mathbf{e} + \beta_1 H(t) + \beta_2 S(t) + \beta_3 Y(t) + \beta_4 A \cdot Y(t) + \beta_5 BH(t) + \beta_6 BS(t) + r + \nu(t), \tag{10}$$

where B denotes the geographic adjacency matrix. We first take the difference between $Y_i(t+1)$ and $Y_i(t)$ to obtain

$$\Delta Y(t+1) = (\beta_1 I + \beta_5 B)\Delta H(t) + (\beta_2 I + \beta_6 B)\Delta S(t) + (\beta_3 I + \beta_4 A)\Delta Y(t) + \varepsilon(t). \tag{11}$$

This allows us to difference out any unobserved state-specific effect captured by r_i as well as the time trend in $\nu_i(t)$. It also alleviates the identification concern coming from the potential correlation between r_i and r_j discussed in the previous subsection. Note that taking the difference does not change the interpretation of the parameters in Equation (1). Solving (11) for $\Delta Y(t)$, we obtain

$$\begin{aligned} A\Delta Y(t) = \sum_{k=0}^{t-3} A(\beta_3 I + \beta_4 A)^k \left((\beta_1 I + \beta_5 B)\Delta H(t-k-1) + (\beta_2 I + \beta_6 B)\Delta S(t-k-1) + \varepsilon(t-k-1)\right) \\ + (\beta_3 I + \beta_4 A)^{t-2}\Delta Y(2), \end{aligned} \tag{12}$$

for all $3 \le t \le 52$. We would like to note that for applications with a finite number of samples, using the inverse mapping imposes conditions on the largest eigenvalue of the corresponding matrices. Such conditions are not necessary for finite sums as in (12), as the summation on the right-hand side is well defined.

Equation (12) readily implies that the tempospatially lagged measurements of absolute humidity that take the form

$$A^l \Delta H(t-k)\,, \text{ and } A^l B \Delta H(t-k)$$

for $l \ge 1$ and $k \ge 1$ [9] are valid instrumental variables for $A\Delta Y(t)$ (and therefore for $\Delta S(t)$ and $\Delta Y(t)$) since they satisfy the following conditions:

[9] Note that we do not necessarily need $k \ge l$; this condition is only needed if we are only studying the very first few periods of the process. In addition, the regression analysis itself requires more periods of observations. For example, if $l = 1$ and $k = 1$, we need at least 3 periods of data in the regression analysis, which is more than enough for $A^2 H$ to be relevant.

1. **relevance:** by Equation (12), they affect $A \cdot \Delta Y$.
2. **exclusion restriction:** Since humidity condition is exogenous, $E[\nu|H(-\infty),\cdots,H(\infty)]=0$, and $E[\varepsilon|\Delta H(-\infty),\cdots,\Delta H(\infty)]=0$. The reason that the time- and spatially-lagged measurements of humidity separately identify the correlated effect from the network effect is the combination of the arguments in Sections 3.2.1 and 3.2.2. The time lag solves the correlation problem over time, and the spatial lag solves the "reflection problem" on the network.

Therefore, we have argued that the following identification results are true.

PROPOSITION 3. Under Assumptions 1 and 3, $A^2\Delta H(t-1)$, $A^2\Delta H(t-2)$, $A^3\Delta H(t-1)$, and $A^3\Delta H(t-2)$ are valid instrumental variables for $AY(t)$, $Y(t)$, and $S(t)$ in Equation (10), and the unknown parameters in (10) are identified.

PROPOSITION 4. Under Assumptions 1 and 4, $AB\Delta H(t-1)$ and $A^2B\Delta H(t-1)$ are valid instrumental variables for $AY(t)$, $Y(t)$, and $S(t)$ in Equation (10), and the unknown parameters in (10) are identified.

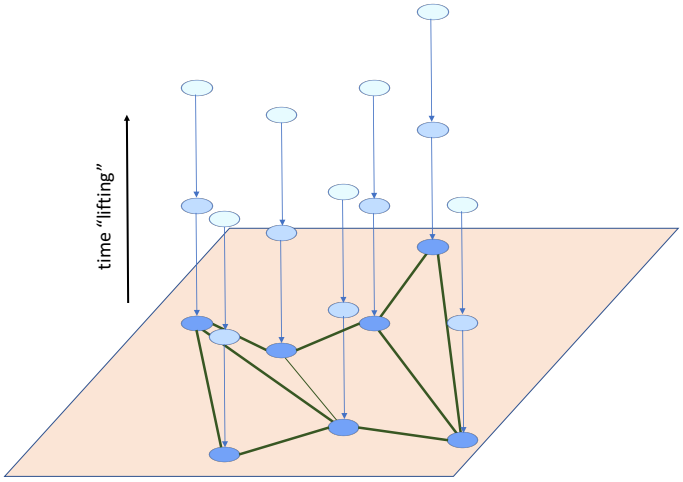


Figure 3 Lifting of a static network using time

One way to understand the intuition of the identification result is to think of the time dimension correlation as a "lifted" version of the network. Two nodes are horizontally linked on this network if they are observations in the same state one period apart from each other. With such a construction, we have a new "lifted" network, as depicted in Figure 3, instead of the original one in the static model. At a high level, we exploit the structure of this network the same way in which one would exploit the original network to construct instrumental variables (in the absence of the time dimension).

The identification method that we develop is not restricted to the application of studying contagious disease. In particular, the method only requires having an exogenous determinant and

variation on the weights of the links on the network. Therefore, the framework is general enough to be applicable to many other settings where estimating causal network effects in contagion processes is of interest.

4. Empirical results

In this section, we present, in increasing complexity and richness, the different models that we use to identify the different effects that determine the propagation of influenza in the United States. The main results of our estimation are given in Table 1, while the details of the estimated models are presented below.

Table 1 Effect of travel on the spread of influenza

Parameter estimates and standard errors

	Model I	Model II	Model III	Model IV	Model V
Humidity	-0.04^{***}	-0.07^{***}	-0.07^{***}	-0.06^{***}	-0.06^{***}
	(0.01)	(0.01)	(0.02)	(0.01)	(0.01)
Susceptible	-0.00	9.22^{***}	9.36^{***}	9.38^{***}	8.96^{***}
	(0.01)	(2.14)	(2.13)	(2.16)	(1.98)
$\Delta Y_i(t)$	-0.01	2.80^{***}	2.79^{***}	2.87^{***}	2.79^{***}
	(0.02)	(0.53)	(0.53)	(0.54)	(0.52)
Travel	1.34^{***}	4.04^{***}	4.23^{***}	4.02^{***}	3.82^{***}
	(0.08)	(1.49)	(1.48)	(1.48)	(1.40)
$(\Delta Y_i(t))^2$			-1.95^{***}	0.41	0.43
			(0.53)	(0.48)	(0.48)
IV		x	x	x	x
Add'l controls				x	x
Add'l IV					x
Year FE	x	x	x	x	x
Number of observations	15450	15450	15450	15450	15450
Weak IV test		1.11	1.11	1.06	1.37

Robust standard errors are in parentheses. $*$, $**$, and $***$ indicate statistical significance at 10%, 5%, and 1% levels, respectively.

4.1. Model I: Basic model without instrumental variables

The main specification of the empirical analysis is as follows:

$$\Delta Y_i(t+1) = \beta_1 \Delta H_i(t) + \beta_2 \Delta S_i(t) + \beta_3 \Delta Y_i(t) + \beta_4 A \cdot \Delta Y(t) + \gamma_1 V(t) + \varepsilon_i(t), \tag{13}$$

where $V(t)$ denotes the season (year) in which week t of the dataset belongs and γ_1 represents the season (year) fixed effect. The estimation results of the first column in Table 1 are obtained by ordinary least squares on (13), thus neglecting the endogeneity issues described in Sections 3.2.1-3.2.3.

In this case, only humidity and travel have statistically significant effects on the spread of flu. Neither the number of individuals in the susceptible population nor the number of individuals

in the infected population in the previous period has a significant effect. As the discussion in the previous section illustrates, the parameters are not identified without instrumental variables, and the results are invalid. These results are also very different from the estimation results using instrumental variables reported in columns 2 to 5, underlining the importance of correcting for the endogeneity issues in the estimation.

4.2. Model II: Basic Model with instrumental variables

The specification of this model is (13). As suggested by Section 3.2.3, we use

$$A^2\Delta H(t-1), \qquad A^2\Delta H(t-2), \qquad A^3\Delta H(t-1), \qquad A^3\Delta H(t-2)$$

as instrumental variables. We use two-stage least squares [53] to obtain the estimates as follows:
Step 1: We regress the endogenous variables $\Delta S_i(t)$, $\Delta Y_i(t)$, and $A\cdot\Delta Y(t)$ on the four instrumental variables as well as the other covariates $\Delta H_i(t)$ and $V(t)$ in the main specification (13).
Step 2: We run the linear regression specified in (13), replacing the endogenous variables $\Delta S_i(t)$, $\Delta Y_i(t)$, and $A\cdot\Delta Y(t)$ with their predicted values the first step. The estimated coefficients in this step are unbiased estimators of the unknown parameters.

4.3. Model III: Incorporating nonlinear terms

In this specification, we also include $(\Delta Y(t))^2$ in the regression to allow for the possible nonlinear relationship between $Y(t+1)$ and $Y(t)$. We observe that nonlinear terms, although significant, do not change the estimates of the different parameters. Furthermore, we find a negative coefficient which shows that although the growth rate of the infected population increases over time, the rate of increase slows down over time. However, this effect goes away when we include additional control variables, which suggests that the linear approximation is reasonable.

4.4. Model IV: Controlling for state correlations and state characteristics

In this specification, we add several additional control variables:

(i) **neighbors' covariates:** $B\cdot\Delta H(t)$, $B\Delta S_i(t)$: the reasoning behind adding these control variables is given in Section 3.2.2. The main idea is to control for the correlation in humidity levels and vaccine availability across nearby states.

(ii) **household income, health index, and gas prices**: Household income[10] and health index[11] are observed at the state level and do not change over time. These are important demographic

[10] https://www.census.gov/2010census/data/

[11] "America's Health Rankings Annual Report is the longest-running annual assessment of the nation's health on a state-by-state basis. For nearly 3 decades, America's Health Rankings Annual Report has analyzed a comprehensive set of behaviors, community and environmental conditions, policies, and clinical care data to provide a holistic view

variables we think might affect the spread of influenza. Gas prices (gas) are also included because they are correlated with in-state travel and thus the spread of flu within the state, conditional on the current size of infected population. As a result, we also control for $G \cdot gas$ and $B \cdot gas$. We do not use gas prices in constructing instrumental variables for the following reason: gas prices are not exogenously determined. There can be unobserved factors, such as the level of economic activities in a state, that are correlated with both gas prices and the spread of flu.

4.5. Model V: Additional instrumental variables

We include additional instrumental variables in this specification. The rest of the model is the same as model IV. The additional instruments are, as suggested in Section 3.2.3,

$$AB\Delta H(t-1), \qquad A^2B\Delta H(t-1).$$

In other words, we use six instrumental variables in total in this specification. As in the previous specifications, we obtain the parameter estimates using two-stage least squares regression.

4.6. Results and interpretation

Our results are summarized in Table 1. Across all specifications with instruments, we see that the coefficients on the main variables of interest are robust and as expected. For example, humidity has a significant negative effect on the spread of influenza. This result is consistent with the findings in [55], [56], and [57] within the medical literature. Moreover, both the size of the susceptible population and the size of the infected population have positive effects on the level of infection. More interestingly, we find that travel has a significant positive effect on the spread of influenza. As it is not straightforward to compare the effect of in-state infections against the effect of infections in the neighboring states, in Figure 4, we provide the histogram of the ratio of the aggregate network effect to the in-state effect,

$$\frac{\beta_3 Y_i(t)}{\beta_4 \sum_j G_{ij} Y_j(t)}.$$

of the health of the nation. The America's Health Rankings Annual Report combines individual measures of each of these determinants with the resultant health outcomes to produce a comprehensive view of the overall health of each state" (https://www.americashealthrankings.org/explore/2017-senior-report). Our variable R_i for each state $i \in V$ is readily obtained by the America's Health Rankings Annual Report.

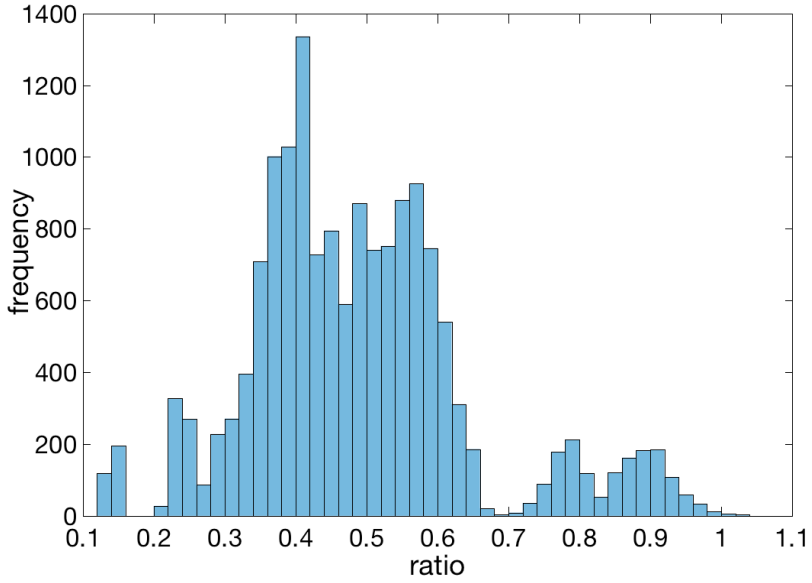


Figure 4 **Ratio of the aggregate network effect divided by the in-state effect.**

As shown in Figure 4, the aggregate network effect is in generally a non-negligible fraction of the in-state effect, underlining the importance of taking travel into account both when analyzing contagion phenomena, and more importantly, when proposing and designing policies, as we will discuss in the next section.

4.7. Strength of instrumental variables

To test the strength of the instrumental variables, we perform the Cragg and Donald test for weak instruments introduced by [58, 59]. This test is equivalent to the common first-stage F test for weak instruments when there is one endogenous variable in the analysis. Since we have three endogenous variables, we compute the matrix analog of the F statistic, compute the smallest eigenvalue, and follow the procedure in [59] to find the critical value. The test statistics are reported in Table 1. In all specifications, we reject the null hypothesis that the instruments are weak.

5. Prevention of contagion: medical vs. network interventions

In this section, we use the estimates from the previous section to evaluate the effect of different types of interventions. As discussed in Section 1, we consider two different classes of intervention: medical and network. Medical interventions decrease the size of the susceptible population by increasing the availability of vaccines. Network interventions include social distancing (decreasing the contact rate) or travel monitoring (control of travel flow or incentives to discourage inter-state travel). For both classes, we propose and analyze different policies in order to facilitate better data-driven decision making for public health officials.

Unfortunately, cost estimates for increasing vaccine availability at the national or state level are not available to us. Similarly, we cannot evaluate the ethical, practical, or financial costs associated with network interventions such as travel monitoring or social distancing. Instead, our approach is to provide quantitative insights into the benefit of different types of interventions in terms of the total number of infections, irrespective of cost considerations.

On the technical side, we emphasize that our estimates are based on observed data and that the goal of our analysis was not *prediction* accuracy, but instead *identification* of causal relationships between different factors (such as traveling behaviors and environmental conditions) and the propagation of an epidemic. Therefore, performing a counter-factual analysis using our model and estimates by simulating the whole trajectory of the epidemic process would be erroneous and inaccurate if attempted. Instead, in the analysis that follows, we evaluate different policies by considering "small perturbations" around the current specification, which allows us to perform comparisons without sacrificing the accuracy of the results. Concretely, we start with our estimated model

$$Y(t+1) = \beta_1 H(t) + \beta_2 S(t) + \beta_3 Y(t) + \beta_4 G \cdot Y(t) + r + \nu(t),$$

where r is the vector of state fixed effects. We define the policy maker's objective to be

$$U(z) = \mathbb{E}\left[\sum_{t=1}^{T-1}\sum_{i=1}^{N} Y_i(t) z^t\right],$$

where z corresponds to the weight that the policy maker assigns to future weeks within the season. For the remainder of this section, we denote by

$$X(z) = \sum_{t=1}^{T-1} x(t) z^t,$$

the Z-transform of a time series $x(t)$. The policy maker's objective coincides with the sum of the Z-transforms of the infection time series over all states. A popular property of the Z-transform that makes our analysis tractable is that the Z-transform of $x(t+1)$ can be written as

$$\sum_{t=1}^{T-1} x(t+1) z^t = \sum_{t=2}^{T} x(t) z^t z^{-1} = z^{-1}(X(z) - z x(1) + x(T) z^T). \tag{14}$$

Using Equation (14) and denoting by $F_i(z)$ the Z-transform of the time series $Y_i(t)$, we write

$$z^{-1}F(z) - Y(1) + Y(T)z^{T-1} = rZ(1) + \beta_1 H(z) + \beta_2 S(z) + \beta_3 F(z) + \beta_4 GF(z) + N(z),$$

where

$$Z(1) = \sum_{t=1}^{T-1} z^t = \frac{1 - z^{T-1}}{1 - z}.$$

Furthermore, using $\mathbb{E}[v_i(t)] = 0$, we get $\mathbb{E}[N(z)] = 0$, and hence, the performance metric of the policy maker can be written as

$$U(z) = \mathbf{e}^T F(z), \tag{15}$$

where

$$F(z) = ((z^{-1} - \beta_3)I - \beta_4 G)^{-1} \left(Y(1) - Y(T)z^{T-1} + r\mathbf{1}Z(1) + \beta_1 H(z) + \beta_2 S(z) \right), \tag{16}$$

which allows us to obtain a closed-form solution to the policy maker's objective as a function of the estimates, the available data, and the factor z. We should emphasize that z should be small enough so that $((z^{-1} - \beta_3)I - \beta_4 G)$ is an M-matrix, allowing us to obtain insights into the efficacy of different policies by perturbing different parameters of (15) around their nominal values, as we shall shortly see.

5.1. Medical interventions

In this section, we consider two different interventions. The first one involves increasing the availability of vaccines (and hence descreasing the size of the susceptible population) uniformly across the country. This approach, although simple conceptually and fair towards the population, treats all states similarly. On the other hand, as we established above, the network effects are strong, and therefore, each state should be treated differently depending on the position of the network. We capture this intricacy by considering the second intervention, which allows the policy maker to increase the availability of vaccines in a targeted manner by focusing on a specific state, thus exploiting the various network effects that are present in the process.

5.1.1. Increasing nationwide availability of vaccines We start by examining the most basic medical intervention strategy: increasing the availability of vaccines by a small amount s throughout the country. In particular, we first observe that increasing the availability of vaccines by an amount s in each week, for every state, leads to an equivalent decrease in the susceptible population by the same amount. Therefore, we can model the intervention as

$$\hat{S}_i(t) = S_i(t) - s,$$

for all $i \in V$ and $t \in [2, \dots T-2]$. In that case,

$$\hat{S}(z) = S(z) - sZ(1)\mathbf{e},$$

where $\mathbf{e}$ denotes an $n \times 1$ vector of all ones, and the resulting performance of the intervention is equal to

$$U(z) = \mathbf{e}^T \hat{F}(z),$$

where

$$\hat{F}(z) = ((z-\beta_3)I - \beta_4 G)^{-1} \left(Y(1) - Y(T) z^{T-1} + rZ(1) + \beta_1 H(z) + \beta_2 S(z) - \beta_2 s Z(1)\mathbf{e} \right).$$

As discussed above, we are interested in small perturbations around the nominal solution, which implies that the effect of a small increase in the availability of vaccines throughout the country is equal to

$$\frac{\partial U}{\partial s}\Big|_{s=0} = -\mathbf{e}^T((z^{-1}-\beta_3)I - \beta_4 G)^{-1}\beta_2 Z(1)\mathbf{e}.$$

5.1.2. Increasing number of vaccines in state i The second intervention that we consider assumes that the policy maker can increase the availability of vaccines for a particular state $i \in V$. In order to be able to make a comparison with the previous case of increasing nationwide availability, we assume that the number of vaccines in state $i \in V$ is increased by $n \cdot s$. The new susceptible vector we can be written as

$$S(t) - ns\mathbf{e}_i,$$

where e_i is an $n \times 1$ vector with an entry of one at the i-th row and zero entries otherwise. In that case,

$$\hat{S}(z) = S(z) - nsZ(1)\mathbf{e}_i,$$

and the resulting performance of the intervention is equal to

$$U(z) = \mathbf{e}^T \hat{F}(z),$$

where

$$\hat{F}(z) = ((z-\beta_3)I - \beta_4 G)^{-1} \left(Y(1) - Y(T) z^{T-1} + rZ(1) + \beta_1 H(z) + \beta_2 S(z) - n\beta_2 s Z(1)\mathbf{e}_i \right).$$

As discussed above, we are interested in small perturbations around the nominal solution, which implies that the effect of a small increase in the availability of vaccines throughout the country is equal to

$$\frac{\partial U}{\partial s}\Big|_{s=0} = -n\mathbf{e}^T((z-\beta_3)I - \beta_4 G)^{-1}\beta_2 Z(1)\mathbf{e}_i.$$

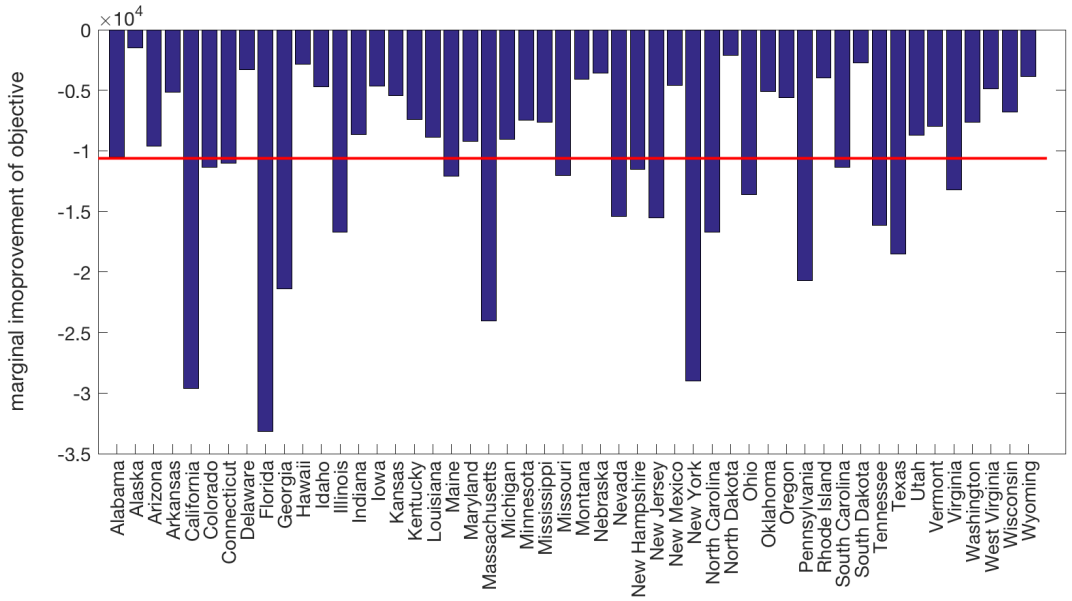


Figure 5 **Marginal improvement of the objective for state-focused medical intervention (blue bars) and nationwide medical intervention (red line).**

5.1.3. Comparison and insights In Figure 5, we show the results of the two different medical intervention approaches: increasing the nationwide availability of vaccines and increasing the availability of vaccines for a specific state $i \in V$ (by the same total amount after aggregating over all states).

Our results crystallize the strength and importance of the network effect. In particular, we identify states (for example, Florida, California, New York, etc.) where increasing the availability of vaccines (and decreasing the size of the susceptible population) leads to a substantially larger benefit for the policy maker. Taking Florida as an example, the benefit is three times as big as that of the nation-wide intervention. This is precisely due to the various network effects present in the propagation of influenza. The inflow and outflow of travelers from such well-connected states leads to a domino effect on the rest of the network, as captured by the term

$$C_i = -n\mathbf{e}_i^T((z-\beta_3)I - \beta_4 G)^{-1}\beta_2 Z(1)\mathbf{e}_i \tag{17}$$

in the marginal benefit of the objective. We would like to emphasize that this term resembles the well-known and widely used Bonacich centrality of state $i \in V$, strengthening the intuition that in the presence of network effects, intervening to more central nodes leads to higher benefits for the welfare of the system.

5.1.4. Relevance of state characteristics The calculation of (17) relies on the estimation of the epidemic parameters, which in turn depend on the different characteristics of each state:

position on the network, health index, population, and median household income. In order to understand the effect of each of these characteristics on the marginal benefit of targeting state $i \in V$ as calculated by (17), we use

$$C_i = \alpha_1 F_i + \alpha_2 R_i + \alpha_3 P_i + \alpha_4 W_i + \epsilon_i,$$

where F_i is the median household income of state i, R_i is the health index of state i, P_i is the population of state i, and W_i is the PageRank centrality of state i, a commonly used centrality measure. The results of the linear model above are presented in Table 2. As expected, the most relevant characteristics of the state are the population and the network centrality. Specifically, the larger the population of a state, the larger the marginal benefit from increasing vaccine availability since in the United States, a larger population typically implies a higher population density and hence higher effective infection rates. Similarly, the more central a state, the larger the marginal benefit from increasing vaccine availability, due to the strong network effects.

Table 2 Role of state characteristics in determining marginal benefit of intervention

	Estimate	Standard Error	pValue
intercept	20.37	8.6694	0.023235
income ($ thousands)	-0.00949	0.01038	0.36504
health	1.5261	1.7198	0.3796
population (millions)	-0.94587	0.0097	1.137710^{-12}
centrality	-14.591	5.0748	0.0061464

5.2. Network Interventions

In this section, we steer our attention to network interventions. Such interventions correspond to increasing social distancing or regulating travel. Clearly, decreasing travel rates across the whole country is prohibitively impractical and expensive. With this idea in mind, we evaluate two more practical types of social distancing using our model and estimates: decreasing the travel rate within a particular pair of states and decreasing the state-level travel rate.

Throughout the rest of this section, the following lemma will prove to be crucial.

LEMMA 1. *For any square matrix G,*

$$\frac{\partial G^{-1}}{\partial x} = -G^{-1}\frac{\partial G}{\partial x}G^{-1}.$$

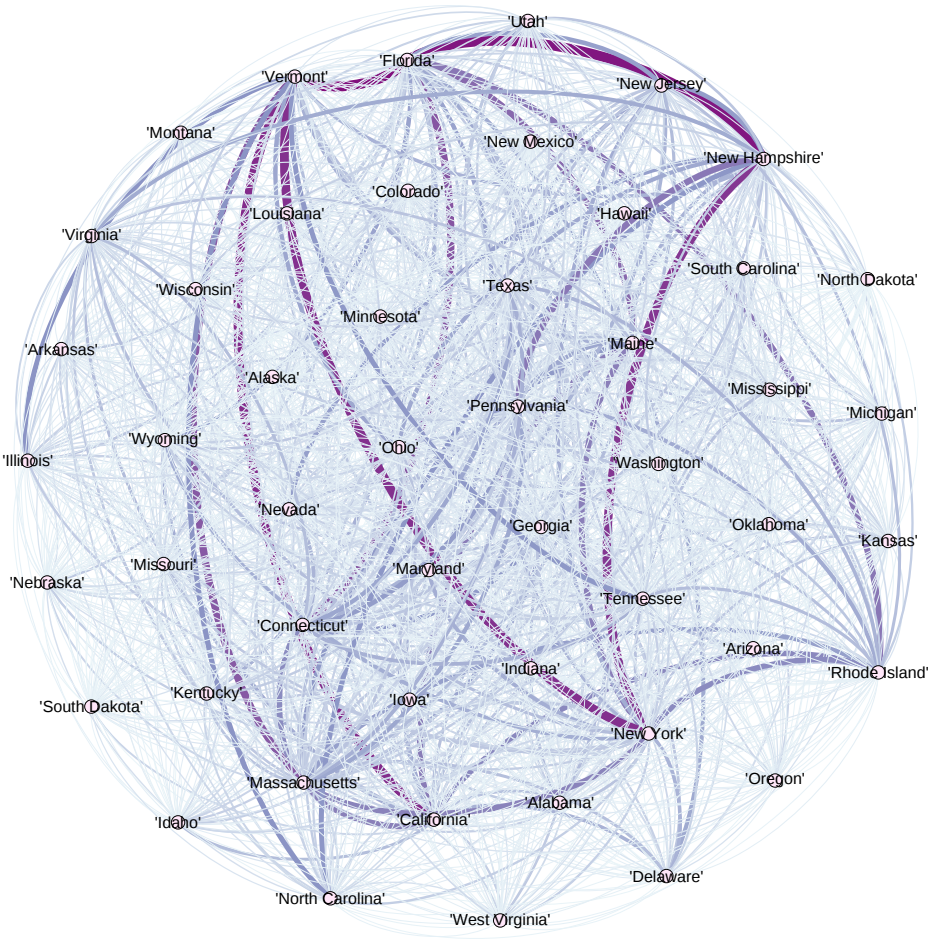


Figure 6 **Marginal improvement of objective from decreasing specific origin/destination travel.**

5.2.1. Decreasing travel rate for specific origin-destination pair We first study the intervention where the policy maker can intervene in any pair of states and decrease the travel intensity on that particular connection. This can be viewed as a perturbation to the travel matrix G, in which case we can write

$$T_{ij} = \frac{\partial U}{\partial g_{ij}}\Big|_{g_{ij}=\hat{g_{ij}}} = \mathbf{e}^T((z-\beta_3)I - \beta_4 G)^{-1}\beta_4 E_{ij}((z-\beta_3)I - \beta_4 G)^{-1} \\ \left(Y(1) - Y(T)z^{T-1} + \beta_0 \mathbf{1} Z(1) + \beta_1 H(z) + \beta_2 S(z)\right). \tag{18}$$

We present our findings for this case in Figure 6 by illustrating a weighted graph, where the thickness and the color of each edge are indicative of the marginal benefit of decreasing travel on that edge. Clearly, using this analysis, the policy maker can identify *central* edges that lead to the largest benefit per unit of decrease in travel.

Figure 6 shows that many *central* links involve the states in New England, which is consistent with the finding in Figure 8 that the states in New England are among those with the highest benefits. Figure 6 further explains that the high benefits come from these states' links to population and economic centers such as New York and Florida. One explanation is that given the environmental characteristics of the New England region, there is a time lag between the spread of the virus and the same process in the rest of the country, including many population and economic centers. Figure 7 is an example comparing the spread of flu in Vermont and Florida during the 2014-2015 season. The gap between the two curves indicates the time lag of the spread of flu in the two states. As a result, travel between the two states might have a higher effect on the diffusion of the virus across the population.

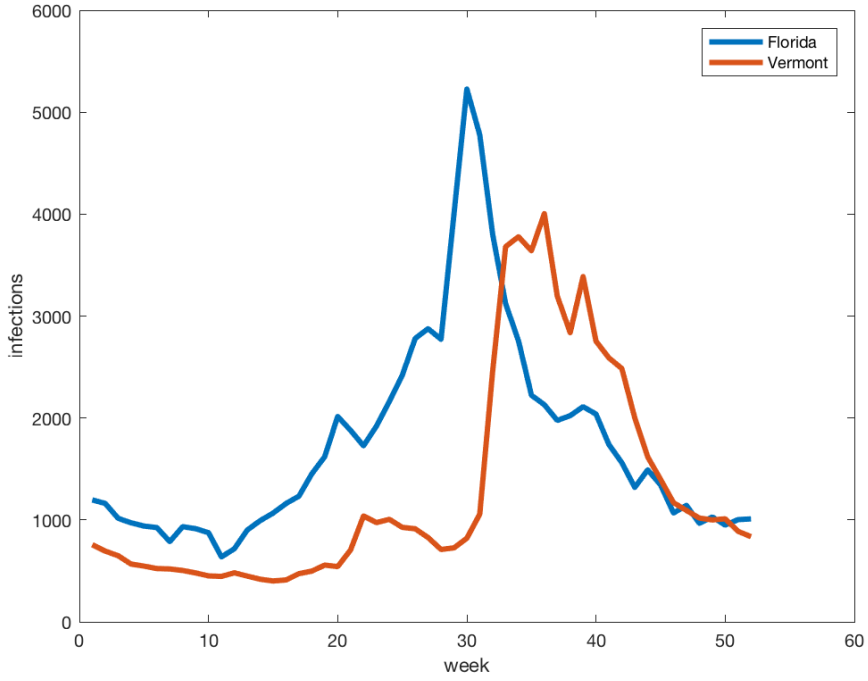


Figure 7 Infections in Florida vs. Vermont, 2014-2015 season.

5.2.2. Decreasing travel outflow from state i The second network intervention we consider is that of decreasing the outflow of travelers from a particular state i. In that case, the resulting performance of the intervention is equal to

$$T_i = \sum_j T_{ji},$$

where T_{ij} is defined in (18).

We present the results of such an intervention for different states in Figure 8. It shows that the intervention is most effective in states such as California, New York, Massachusetts, and Florida. Although different in magnitude, the the results are very similar to those in Figure 5. In other words, the intervention is more effective in states with larger population and higher network centrality.

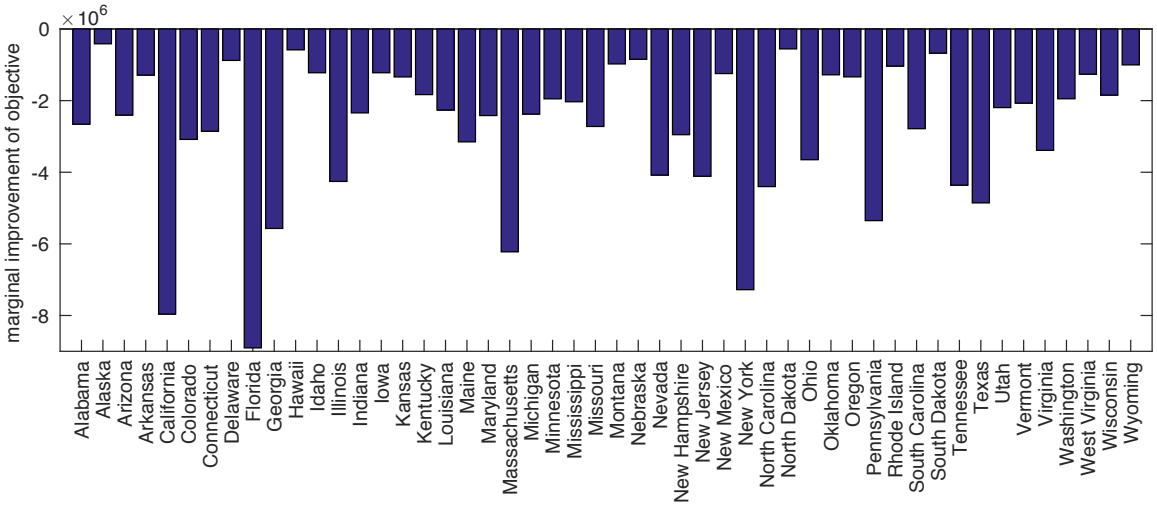


Figure 8 Marginal improvement of objective for state-focused social distancing.

6. Conclusions

We study network effects in contagion processes using the example of seasonal flu, i.e., the effect of travel on the spread of flu across the population. Since a large scale randomized experiment is infeasible in this setting, we use observational infection data in the U.S. in order to estimate the causal network effect. The main challenge in estimation is that the network effect is difficult to identify in observational data for two main reasons. First of all, because the data is generated from a contagion process, the observations are correlated across different time periods. This creates difficulties in the identification and estimation of the unknown parameters in the model. Second, the network effect is difficult to identify, even without the dynamics. The endogenous network effect, i.e., how Y_i affects Y_j if i and j are linked on the network, is difficult to be identified separately from the correlated effect, which arises from similarities in the determinants of Y_i and Y_j. We provide a method using instrumental variables to solve the identification problem. We show that after some transformation of the model, time- and network-lagged exogenous determinants of Y can be used as valid instrumental variables. Employing this method, we estimate that the effect of travel on the spread of flu is substantial. We then use the estimation result to evaluate the effects of medical and social interventions on the spread of flu. First, we show that taking into account the network

effect is key in designing effective vaccine distribution policies. In fact appropriately accounting for the effect of travel can lead to three times more effective policies. Second, we illustrate the effect of travel restrictions on the spread of flu using the estimation result. We show that there is large heterogeneity across geographic areas in terms of the effect of such policies. This implies that identifying the central nodes or central links on the network, which can be done using our estimation result, is key in designing social policies that are effective in controlling contagious disease.

From the methodological perspective, this paper is the first to introduce an identification method to estimate network effects in dynamic settings. The identification method combines methods used in panel data analysis and in static network effect estimation. It is not restricted to the specific application of studying epidemics, but rather can also be applied to studying many different operational decisions where understanding of the network effect is important. For example, the method can be applied to studying pricing and inventory decisions for network products, supply chain disruptions, and systemic risks spread across financial institutions. The wide range of potential applications is a big strength of our method. In contrast, the limitation of the method is that it requires having at least one truly exogenous determinant of the outcome of interest, which might not always be the case. However, without some true randomness in the system, the identification would be difficult to achieve, regardless of the method of choice.

Concluding, in this application, the travel intensities and hence the corresponding network structure is considered static due to the lack of high quality dynamic data. Clearly, the seasonal and spatial variability of travel intensities would not only allow us to obtain better estimates but would also pose a number of challenges and research directions of interest, in particular due to endogeneity issues in the evolution of the network structure.

References

[1] E. Adar and L. A. Adamic. Tracking information epidemics in blogspace. In *Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence*, WI '05, Washington, DC, USA, 2005. IEEE Computer Society.

[2] Hunt Allcott and Matthew Gentzkow. Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 2017.

[3] J. Leskovec, L. A. Adamic, and B A. Huberman. The dynamics of viral marketing. *ACM Trans. Web*, 1(1), May 2007.

[4] Michele Garetto, Weibo Gong, and Don Towsley. Modeling malware spreading dynamics. In *INFOCOM 2003. Twenty-Second Annual Joint Conference of the IEEE Computer and Communications Society*, volume 3. IEEE, 2003.

[5] E.M. Rogers. *Diffusion of Innovations, 5th Edition*. Simon and Schuster, 2003.

[6] Daron Acemoglu, Asuman Ozdaglar, and Alireza Tahbaz-Salehi. Systemic risk and stability in financial networks. *American Economic Review*, 105(2):564–608, February 2015.

[7] Monica Billio, Mila Getmansky, Andrew W Lo, and Loriana Pelizzon. Econometric measures of connectedness and systemic risk in the finance and insurance sectors. *Journal of Financial Economics*, 104(3):535–559, 2012.

[8] Nikolay Osadchiy, Vishal Gaur, and Sridhar Seshadri. Systematic Risk in Supply Chain Networks. *Management Science*, 62(6):1755–1777, 2016.

[9] Vasco M. Carvalho, Makoto Nirei, Yukiko U. Saito, and Alireza Tahbaz-Salehi. Supply Chain Disruptions: Evidence from the Great East Japan Earthquake. *working paper*, 2017.

[10] Raymond S. Koff. Infectious diseases of humans: Dynamics and control. *Hepatology*, 15(1):169–169, 1992.

[11] Daryl J. Daley and J. M. Gani. *Epidemic modelling : an introduction*. Cambridge University Press New York, 1999.

[12] Linda JS Allen, Fred Brauer, Pauline Van den Driessche, and Jianhong Wu. *Mathematical Epidemiology*. Springer, 2008.

[13] Louis Kim, Mark Abramson, Kimon Drakopoulos, Stephan Kolitz, and Asu Ozdaglar. Estimating social network structure and propagation dynamics for an infectious disease. In *Social Computing, Behavioral-Cultural Modeling and Prediction*, volume 8393 of *Lecture Notes in Computer Science*. Springer International Publishing, 2014.

[14] J. Leskovec, L. A. Adamic, and B. A. Huberman. The dynamics of viral marketing. In *Proceedings of the 7th ACM Conference on Electronic Commerce*, EC '06, New York, NY, USA, 2006.

[15] S. Aral and D. Walker. Identifying influential and susceptible members of social networks. *Science*, 337(6092):337–341, 2012.

[16] M. Gomez Rodriguez, J. Leskovec, and A. Krause. Inferring networks of diffusion and influence. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '10, New York, NY, USA, 2010.

[17] R. Cohen, S. Havlin, and D. Ben-Avraham. Efficient immunization strategies for computer networks and populations. *Physical Review Letters*, 91:247901, 2003.

[18] E. Gourdin, J. Omic, and P. Van Mieghem. Optimization of network protection against virus spread. In *8th International Workshop on the Design of Reliable Communication Networks (DRCN)*. IEEE, October 2011.

[19] F. R. K. Chung, P. Horn, and A. Tsiatas. Distributing antidote using pagerank vectors. *Internet Mathematics*, 6(2):237–254, 2009.

[20] V. M. Preciado, M. Zargham, C. Enyioha, A. Jadbabaie, and G. J. Pappas. Optimal vaccine allocation to control epidemic outbreaks in arbitrary networks. *CoRR*, abs/1303.3984, 2013.

[21] K. Drakopoulos, A. Ozdaglar, and J.N. Tsitsiklis. An efficient curing policy for epidemics on graphs. *IEEE Transactions on Network Science and Engineering*, 1(2):67–75, July 2014.

[22] Christian Borgs, Jennifer Chayes, Ayalvadi Ganesh, and Amin Saberi. How to distribute antidote to control epidemics. *Random Structures and Algorithms*, 37(2):204–222, 2010.

[23] World Health Organization. Influenza (seasonal). *Fact sheet No 211*.

[24] Jeffrey Shaman and Melvin Kohn. Absolute humidity modulates influenza survival, transmission, and seasonality. *Proceedings of the National Academy of Sciences*, 106(9):3243–3248, 2009.

[25] Charles F Manski. Identification of endogenous social effects: The reflection problem. *The review of economic studies*, 60(3):531–542, 1993.

[26] U.S. Congressional Budget Office. U.s. congressional budget office.a potential influenza pandemic: An update on possible macroeconomic effects and policy issues.

[27] U.S. Department of Health and Human Services. Pandemic Influenza Plan: 2017 Update.

[28] Jeffrey M. Drazen, Rupa Kanapathipillai, Edward W. Campion, Eric J. Rubin, Scott M. Hammer, Stephen Morrissey, and Lindsey R. Baden. Ebola and quarantine. *New England Journal of Medicine*, 371(21):2029–2030, 2014. PMID: 25347231.

[29] Center for Disease Control. Use of Group Quarantine in Ebola Control—Nigeria, 2014. *Morbidity and Mortality Weekly Report (MMWR)*.

[30] Eric Telmor and Emma Farge. Struggling Liberia creates 'plague villages' in Ebola epicentre. *Reuters*.

[31] Institute of Medicine. *Ethical and Legal Considerations in Mitigating Pandemic Disease: Workshop Summary*. The National Academies Press, Washington, DC, 2007.

[32] Yann Bramoullé, Habiba Djebbari, and Bernard Fortin. Identification of peer effects through social networks. *Journal of econometrics*, 150(1):41–55, 2009.

[33] David Godes and Dina Mayzlin. Using online conversations to study word-of-mouth communication. *Marketing Science*, 23(4):545–560, 2004.

[34] Christophe Van den Bulte and Gary L. Lilien. Medical innovation revisited: Social contagion versus marketing effort. *American Journal of Sociology*, 106(5):1409–1435, 2001.

[35] Sinan Aral, Lev Muchnik, and Arun Sundararajan. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences*, 106(51):21544–21549, 2009.

[36] Esther Duflo, Rachel Glennerster, and Michael Kremer. *Using Randomization in Development Economics Research: A Toolkit*, volume 4. North Holland, Amsterdam and New York, 2008. This file is the version posted by the Centre for Economic Policy Research, CEPR Discussion Papers: 6059, 2007.

[37] Bruce Sacerdote. Peer effects with random assignment: Results for dartmouth roommates. *The Quarterly journal of economics*, 116(2):681–704, 2001.

[38] Antoni Calvó-Armengol, Eleonora Patacchini, and Yves Zenou. Peer effects and social networks in education. *The Review of Economic Studies*, 76(4):1239–1267, 2009.

[39] Sinan Aral and Dylan Walker. Creating social contagion through viral product design: A randomized trial of peer influence in networks. *Management science*, 57(9):1623–1639, 2011.

[40] Sinan Aral and Dylan Walker. Tie strength, embeddedness, and social influence: A large-scale networked experiment. *Management Science*, 60(6):1352–1370, 2017/11/27 2014.

[41] Bryan S Graham. Identifying social interactions through conditional variance restrictions. *Econometrica*, 76(3):643–660, 2008.

[42] Paul Goldsmith-Pinkham and Guido W Imbens. Social networks and the identification of peer effects. *Journal of Business & Economic Statistics*, 31(3):253–264, 2013.

[43] Gal Oestreicher-Singer and Arun Sundararajan. The visible hand? demand effects of recommendation networks in electronic markets. *Management Science*, 58(11):1963–1981, 2017/11/27 2012.

[44] Sinan Aral and Christos Nicolaides. Exercise contagion in a global social network. *Nature Communications*, 8:14753 EP –, 04 2017.

[45] Michael Trusov, Randolph E. Bucklin, and Koen Pauwels. Effects of word-of-mouth versus traditional marketing: Findings from an internet social networking site. *Journal of Marketing*, 73(5):90–102, 2009.

[46] Gary Chamberlain. Multivariate regression models for panel data. *Journal of econometrics*, 18(1):5–46, 1982.

[47] Jeremy Ginsberg, Matthew Mohebbi, Rajan Patel, Lynnette Brammer, Mark Smolinski, and Larry Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457:1012–1014, 2009. doi:10.1038/nature07634.

[48] David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani. The parable of google flu: traps in big data analysis. *Science*, 343(6176):1203–1205, 2014.

[49] Justin R Ortiz, Hong Zhou, David K Shay, Kathleen M Neuzil, Ashley L Fowlkes, and Christopher H Goss. Monitoring influenza activity in the united states: a comparison of traditional surveillance systems with google flu trends. *PloS one*, 6(4):e18687, 2011.

[50] Avinash Patwardhan and Robert Bilkovski. Comparison: Flu prescription sales data from a retail pharmacy in the us with google flu trends and us ilinet (cdc) data as flu activity indicator. *PloS one*, 7(8):e43611, 2012.

[51] Ozgur M Araz, Dan Bentley, and Robert L Muelleman. Using google flu trends data in forecasting influenza-like–illness related ed visits in omaha, nebraska. *The American journal of emergency medicine*, 32(9):1016–1023, 2014.

[52] Mauricio Santillana, André T Nguyen, Mark Dredze, Michael J Paul, Elaine O Nsoesie, and John S Brownstein. Combining search, social media, and traditional data sources to improve influenza surveillance. *PLoS computational biology*, 11(10):e1004513, 2015.

[53] Jeffrey M Wooldridge. *Econometric analysis of cross section and panel data*. MIT press, 2010.

[54] Gary Chamberlain. Panel data. *Handbook of econometrics*, 2:1247–1318, 1984.

[55] Jeffrey Shaman, Virginia E. Pitzer, Cécile Viboud, Bryan T. Grenfell, and Marc Lipsitch. Absolute humidity and the seasonal onset of influenza in the continental united states. *PLOS Biology*, 8(2):1–13, 02 2010.

[56] Jeffrey Shaman and Melvin Kohn. Absolute humidity modulates influenza survival, transmission, and seasonality. *Proceedings of the National Academy of Sciences*, 106(9):3243–3248, 2009.

[57] Ethan R. Deyle, M. Cyrus Maher, Ryan D. Hernandez, Sanjay Basu, and George Sugihara. Global environmental drivers of influenza. *Proceedings of the National Academy of Sciences*, 113(46):13081–13086, 2016.

[58] John G Cragg and Stephen G Donald. Testing identifiability and specification in instrumental variable models. *Econometric Theory*, 9(2):222–240, 1993.

[59] James H Stock and Motohiro Yogo. Testing for weak instruments in linear iv regression. *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, page 80, 2005.