Social Network and Economic Development

By Thao Trang Nguyen[1]

Abstract

I use anonymized data from Facebook to construct social diversity and social connectedness weight matrix of counties in the U.S to evaluate the relationship between social network and economic development. I find that there is a negative relationship between social diversity and MDI rate, which further confirms the theoretical understanding that diverse contact helps improving socio-economic performance at individual and community levels. However, once I estimate this relationship for only the 10 percent counties having highest social diversity, this relationship disappears which implies that the relationship between having high social diversity does not hold true for better economic development performance for counties with highest social diversity level. I further ask the question of social and spatial spillovers in economic development between counties and try to answer by estimating an econometric specification for geographical and social connection weight matrix. My findings suggest that there are spillover effects among counties in terms of geographical and social connection. However, I cannot conclude these effects as causal relationships. Nevertheless, this paper can be considered as another step in understanding the population-level relationship between social network and economic development based on online network data to proxy for real life relationships.

Keywords: social connectedness, social network, economic development

Date: 5 February, 2021

[1] PhD fellow – UNU-MERIT, Maastricht University
I thank Prof. Robin Cowan for his contribution to help me with the idea of this paper and his comments on building a regression model.

Content

List of tables and figures

List of tables

List of figures

1 Introduction

Social network exists in our daily life and affects every aspect of our social and economic development. Theoretically, social network structure is said to have an impact on economic performance of individuals and the whole community. While the debate about which network structure supports economic development as a whole the most or how network forms is still ongoing, the methods and datasets for evaluating the impact of social network on population's economic development are expanding with the existence of big data. If Eagle et al (2010) was the first one to use an aggregate level data of individual mobile phone network data to estimate the relationship between network diversity and the whole community's economic development, more data is now available for researchers to dig further in this topic, including satellite data, online social network data, etc. However, in the case of online social network data, most of the research do not use public data, but mostly private data from a specific company by cooperating with some of the people working in that company and those datasets are only available for that researcher for a limited amount of time.

In this paper, I will go further to explore the relationship between network diversity and economic development by using a public dataset from Facebook to estimate network diversity through Social Connectedness Index as published in their website. I use the multidimensional deprivation rate as a proxy for the level of economic development in each county in the US. Therefore, the aim of this paper is (i) to evaluate and quantify the relationship between network diversity and socio-economic development of a community (ii) to estimate the extent of social and spatial socio-economic spillovers between U.S. counties and highlight the importance of geographical networks as well as social network between counties.

I first explain the method that I am using throughout this paper. The first method is adopted from Eagle et al (2010) to build a social diversity index for all counties in the U.S. I then explain the regression model that I use to quantify the relationship between social diversity and MDI rate. My results suggests that there is a negative relationship between social diversity and MDI rate. In applying my data to estimate a regression model for social diversity and MDI rate, I also check with OLS assumptions and spatial dependence of my dataset. My regression models

suffer from heteroskedasticity and spatial dependence among counties, which I try to adjust by using a spatial lag model. The second method is adopted from Amarasinghe et al (2018) to estimate how one county's MDI rate depends on its social and geographical connectivity with other counties in their network. I estimate this econometric model by using a dataset of MDI rate from US census, the social connectedness index from Facebook, and the distance data from NBER. My measure of socio and economic performance is MDI rate, and my interested independent variable is social connectedness. In order to estimate for the spillover effects, I need to build a matrix for geography and social connectedness. My results suggest that there are spillover effects for being geographical and social connected with other counties. I then use robustness check to evaluate my results. The results for social diversity, in overall, are robust, however, there is no evidence in whether there is an impact of social diversity on MDI rate for top 10% counties with highest social diversity. The results for robustness check for social connectedness are not reported due to being perfectly fit error. Therefore, even though my regression models are statistically significant, I cannot say that this relationship is causal as the social network itself suffers from being endogenous, but due to the time limitation, I cannot conduct a method to adjust for endogeneity.

Nevertheless, I think this paper could contribute to the literature in two ways. First, I further confirm the existence of the relationship between network diversity and economic development, that a community which is more socially connected will have better performance in social and economic aspects. Second, I also find that this relationship seems not true for the top 10% counties with highest social diversity, as shown in the map and the regression results. Third, I also further confirm the existence of spillover effects in terms of social and spatial dependence in economic development between counties. And finally, I contribute to a recent literature which use online data to understand various aspects of social and economic performance and social network.

The remaining of this paper is organized as follows. Section 2 summarizes some literature about social network, poverty, and the relationship between social network and socio-economic performance. I then explain the methodology that I use in this paper in section 3. Section 4 lists some dataset and explains in details which variables that I will use in my analysis.

Section 5 discusses some important results and evaluates my results with robustness check. Then in section 6, I discuss some limitations in my methodology and set out some ideas for future work. Finally, I conclude the paper in section 7.

2 Literature Review

The classical economic model sees humans as homoscedasticity and there is no interaction between human. However, since 1990s, the neoclassical model has looked into a different lens seeing people as heteroskedasticity and they can interact with each, and then influence each other's decision. Studies of network have now more expanded in understanding economic behaviors. Granovetter (2005) explains that by carrying information social network could help to diffuse the information and then can affect economic outcomes. He also points out other mechanisms that makes social network affect economic outcomes which are reward and punishment. A final mechanism is through trust embedded in the social network and believe that "others will do the right thing". Jackson (2014) explains these mechanisms more in details in terms of network characteristics, including "network-based notions of density", "distribution of connections", "segregation patterns", and the "positions of key nodes". He also emphasizes on the "non-unidirectional" relationship between social network and economic outcomes. Social network does not only determine economic outcomes, but vice versa, it is partly determined by economic outcomes. It is important to understand the difference between macro level characteristics of network, such as density of links or segregation patterns, and micro level characteristics, for instance whether a person's friends are friends with each other (Jackson, 2014).

Researchers have tried to use empirical evidence to explain these mechanisms. Previous studies have found the importance of social network in many areas. In terms of diffusion, social network is reported to play an important role in spreading ideas, information, shaping behaviors and even spreading diseases. Barr (2000) finds that entrepreneurial networks are a determinant of Ghanaian manufacturing enterprise performance, in which entrepreneurs with larger and more diverse contacts will have better performance. In analyzing the importance of knowledge flows between enterprises, he presents that the knowledge flow in Ghana is not

sufficiently complementary with their own knowledge to achieve endogenous growth. In learning about technology adoption, there's research from Conley and Christopher (2001) and Bandiera and Rasul (2006) who analyze the process of social learning and conclude that a farmer's initial decision to adopt a new technology depending on the decisions of others in their social network. Banerjee et al (2013) expand this idea further and analyze the diffusion of microfinance in rural Indian villages. Their findings suggest that households that have more friends participated in microfinance are more likely to hear about microfinance than households who have lower friends participated in microfinance. In terms of information flow and performance, Aral and Van Alstyne (2008) find that diverse networks drive economic performance where individuals in the network can access to novel information. These findings lie on some important aspects of network such as segregation, network density, or the adapted behaviors of network after an individual's reaction to the changes in the circumstances. For example, highly clustered networks are believed to limit access to economic opportunities from outside networks, while diverse networks give more opportunities for individuals in the network to access new information. Burt (2004) is a pioneer in this area where he concludes that individuals having a network low in cohesion will have better performance. Another opposite idea is started from Coleman (1988) where he concludes that dense networks are important for social capital. His reasons lie on that dense groups will have common language and can be a base to create a "critical mass" for knowledge generation. Granovetter (2005) also gives explanations on why dense networks might be better for performance as denser network can help overcome free-rider problems and emphasize trust between individuals in the network. These explanations are all relevant in understanding how network affects economic behaviors, social capital and economic performance, like Woolcock and Narayan (2000) concludes social capital in terms of network is "a double-edged sword" that involves both benefits and cost.

Some other areas also witness an increasing number of research on social network including analyzing criminality behaviors, labor market and trade. Criminal behaviors are found not to be at an individual level and do not happen at isolation, but happen in a social context, depend on social networks of individuals. Patacchini and Zenou (2008) find that young people have

higher degrees of social interactions happen to commit more crime. Ludwig et al (2013) find that relocating families from high- to low-poverty neighborhood can reduce violence by 30%-50% and imply an importance of social interactions in criminal behaviors. In labor market, social networks also play an important role. Starting from Granovetter's ideas of the strength of weak ties, Gee et al (2015) use six million Facebook users' data and find an evidence that most people find jobs through one of their weak ties relationships, however, she also finds that a single of strong tiles is more valuable for job seekers at the margin. Social networks also exist in trade patterns. Empirical findings from Chaney (2014) and Morales et al. (2019) find that social clusters influence trade, which means if two countries are in the same social cluster, it is likely that there is a bilateral trade between them. This pattern is found because being in the same social clusters help them to reduce information asymmetries and improve contract enforcement (Bailey et al, 2020), which agrees with the findings from Coleman. Guiso et al. (2009) also finds that trust facilitates trade and influence the flows of goods.

For literature on poverty, research observes a persistent pattern on inequality in wages, health and other economic and well-being dimensions. The reason for this persistent pattern might be that poor or disadvantaged people can only interact with other similar people like them, so it is hard for them to climb to upward situations due limited contacts. The peer effects from social network of poor people bring more negative feedbacks rather than a positive role model. Bertrand et al (2000) report the importance of networks (measured by language spoken at home) in welfare participation. For example, if an individual has a friend who is in welfare program, they can benefit by reducing the cost of applying for welfare, learning more about welfare program with a cost of hearing less about job opportunities or other opportunities. Harrison et al (2019) also find that communities with higher social capital will have lower poverty rates, and policies in reducing poverty will be more helpful if combining with supporting social capital formation, particularly more important for communities in persistent poverty.

Most of these above listed empirical evidence lie on analyzing a sub-population's social network, except for the research by Gee et al (2015) and Bailey et al (2020). With the help of big data, the research on social network and a population's economic well-being have been

easier. Several research has focused on mobile phone data to predict economic development, social mobility as well as estimate the relationship between network diversity and economic development. A pioneer in this area is from Eagle et al (2010) who use communication network data in August 2005 in the UK which covers more than 90% of the mobile phones in the country to find a relationship with socioeconomic opportunity. They find that economic development of communities is highly correlated with network diversity of that community. Satellite data is also becoming more popular in economics to predict poverty and estimate economic well-being (Engstrom et al, 2017; Jean et al, 2016; Donaldson and Storeygard, 2016; Amarasinghe et al, 2018). Another promising dataset is using social connectedness data from online social network, for example Facebook, LinkedIn, and Twitter to analyze social network and individual's economic well-being as well as other relevant economic perspectives, for example knowledge flow, housing market, international trade, or labor market. Bailey et al (2018a) find evidence of social networks in housing market in the U.S. by using anonymized data from Facebook. As mentioned above, Gee et al (2015) also use anonymized data from Facebook to evaluate the importance of social contacts in finding jobs. Bailey et al (2020) continues with the anonymized data from Facebook to build a social connectedness index for 180 countries and 332 European regions to pattern the relationship between social network and international trade. Diemer (2020a, 2020b) also builds on this social connectedness index to map and find the relationship between social network and the geography of knowledge flows in the US, and spatial diffusion of economic shocks in networks.

3 Methods

3.1 Social Diversity

By applying a similar method from Eagle et al (2010), I will calculate a variable called Dsocial(i) which measures how connected different counties to each other in the US.

$$\text{Hi} = -\sum_{j=1}^{k} pSCij * \log\left(pSCij\right)$$

Where k is the number of counties in the US and pSCij is the proportion of county i's total normalized social connectedness that involves county j, or

$$\text{pSCij} = \frac{normalized\ SCij}{\sum_{j=1}^{k} normalized\ SCij}$$

Where normalized SCij is the social connectedness index normalized by the product of the total population of two counties. I then define social diversity, Dsocial(i) as the Shannon entropy associated with county i's communication behavior, normalized by k:

$$\text{Dsocial(i)} = \frac{Hi}{\log(k)}$$

I then plot on the map the geographical distribution of social diversity and MDI rate to evaluate the relationship between economic development outcomes (MDI data) and social diversity.

3.2 Econometric Specification

Regression model with Social Diversity

My next step is to quantify the relationship between MDI rate and social diversity by regressing MDI rate on social diversity. The econometric specification is as follows:

$$\text{MDIi} = \beta_o + \beta_1 * \text{Social Diversity}_i + \text{Xi} + \varepsilon_i \qquad \text{(I)}$$

Where MDIi is the multi-dimensional deprivation index rate for county i (the MDI rate is only calculated in 2017), Social Diversity$_i$ is the social diversity calculated from the above part, Xi is the set of control variables (described below). β_1 is the interested coefficient in this econometric model.

Regression model with Social Connectedness

The method from Eagle et al (2010) gives us a brief understanding on the correlation between social network and economic development. However, I also want to explore whether there are social and spatial spillover effects in economic development. Therefore, I will apply the methods from Amarasinghe et al (2018) to answer this question. They use an econometric model to estimate spatial spillovers of economic activities in districts in 53 African countries. I will also build an econometric model based on their econometric model to evaluate the social and spatial effects of MDI rate among counties.

The econometric equation is described as:

$$\mathrm{MDI}_{it} = \beta_0 + \beta_1 * \sum_{j=1}^{J} w_{1,i,j} * MDI_{jt} + \beta_2 * \sum_{j=1}^{J} w_{2,i,j} * MDI_{jt} + X_{it} + \varepsilon_{it} \quad \text{(II)}$$

Where MDI_{it} is the multidimensional deprivation index of county i at time t (in this study, I only evaluate MDI in 2017), $w_{1,i,j}$ is the (i,j) cell of the adjacency matrix based on geographic connectivity, $w_{2,i,j}$ is the (i, j) cell of the adjacency matrix based on social connectedness, X_{it} is the set of county level characteristics as control variables of county i at time t (2017). This econometric model assumes that the MDI local spillovers to other counties only happen by the geographical and social connected lag of the dependent variable. It should be noted that in the econometric model by Amarasinghe et al (2018), they also introduce the local spillover effects due to spatial lag of the explanatory variables. However, due to the time constraint of the paper, I only consider lag of the dependent variable, and the lag of the explanatory variables may be interesting as a follow-up for the future study.

Endogeneity of network formation is a problem when working with network. In this case, there could be a reverse causality, in which MDI rate affects the network of that county by the movement of poor people between counties, rather than the network itself affects MDI. Diemer (2020a) adjusts this concern by creating a new measure of social connectedness which is the relative probability of friendship that controls migration and distance between counties. However, due to the time constraint of the paper, I will assume that the movement of poor people will take time and social connectedness in the short term do not fluctuate much and is not affected by the movement of poor people in the short term. Though the MDI is measured in 2017 and the social connectedness is calculated in 2020, I will assume that social connectedness does not change much within this time frame.

4 Data

The unit of observation in this study is at county level. For answering the research questions proposed, I will need data for economic development (proxy by the multidimensional deprivation index), data for social connectedness, data for geography and data for control variables.

4.1 Dependent variable: Multidimensional Deprivation Index

To measure economic development for each county, I use Multidimensional Deprivation Index (MDI) as a proxy for economic development. According to the U.S. Bureau of Census, MDI is a complement for the Official Poverty Measure and has six dimensions: standard of living, education, health, economic security, housing quality, and neighborhood quality.

The dataset is acquired through the website of U.S. Bureau of Census, in which the most updated version of this dataset is 2017. In the report, U.S. Census Bureau also mentions that each dimension was weighted equally (though it is not necessary). The reasons include that it is easy to understand and the dimension can be easy to stand on its own; that to weight one dimension more important than another needs a valid justification which is case by case depend on individuals and related to the robustness of the results.

Specifically standard of living is defined as poverty according to the official poverty measure; education is measured as aged 19 or older without a high school dilemma; health is calculated as poor health status; economic security includes at least two of the three conditions which are lack of health insurance, unemployed for 12 months, cumulative hours worked per week was less than 35 hours; housing quality is defined as lack of complete kitchen, plumbing, overcrowded housing unit or high cost burden; and neighborhood quality is measured as living in a neighborhood which has high crime, poor air quality, or poor food environment.

Since the MDI already adjust for the total population of a county, I will use the MDI rate without any modifications in this study.

4.2 Data for social connectedness

To measure social network, I use Social Connectedness Index (SCI) developed by Bailey et al (2018b). They use an anonymized snapshot of active Facebook users and their friendship networks to measure the social connectedness index across locations. To get their location, they use the locations based on their information listed on their Facebook account and their IP address when they log in their Facebook account. They are defined as:

$$Social\ Connectedness\ Index_{i,j} = \frac{FB\ Connections_{i,j}}{FB\ Users_i * FB\ Users_j}$$

Where FB $Users_i$ and FB $Users_j$ are the number of Facebook users in locations i and j, and FB $connections_{i,j}$ is the total number of Facebook friendship connections between individuals in the two locations i and j. The public data measures the relative probability of a Facebook friendship between a given Facebook user in location i and a given Facebook user in location j, which has the value from 1 to 1,000,000,000. In the US counties Social Connectedness Index, the maximum value is Los Angeles County-Los Angeles County connection with the value of 1,000,000. From this dataset, I then have a social network with 3,136 nodes and 9,462,485 edges.

Users of Facebook is mostly unchanged since 2018, which about 70% of adults in the U.S. use the platform (Pew Research, 2019). In the same report, it is reported that people from 18-49 use Facebook the most, around 70-80% for each age group. It decreases by age group, however, for people 65 and more, there are still 46% of people use Facebook. Therefore, the online Facebook friendship could be a good proxy for social connection of US friendship network in real life. (Gee et al, 2015). I use the most updated version (till the time this research is written) of US counties SCI from the website of Facebook Data for Good which is September 2020.

To measure social connectivity between counties, I build a social connectivity matrix. First, I normalize the SCI by the product of the population of two counties connected. This is to adjust that counties which have larger population will have larger facebook users, and the SCI will be larger. This normalization is also done by Bailey et al (2018b) and Diemer (2020a). Second, after I have the normalized SCI, I create a matrix between counties, in which the value between county i and county j is 0 ($w_{ic,jc}$) if the normalized SCI below the median level of all normalized SCI, and the value between county i and county j is 1 ($w_{ic,jc}$) if the normalized SCI greater or equal the median level of all normalized SCI.

4.3. Data for Geography

Bailey et al (2018b) specify that people tend to be friends with each other if they are located near each other, so geography data is important in my study, as also specified in the method. In this study, I use the data for Geography extracted from NBER which has data on the distance between counties.

To measure geographic connectivity between counties, I will build a geographic connectivity based on the distance between counties. I construct a spatial weight matrix as the value between county i and county j is 1 ($w_{ic,jc}=1$) if the distance between two counties is lower than 200 miles, and equals 0 ($w_{ic,jc} = 0$) if the distance between two counties is greater than 200 miles (Bailey et al (2018b) also defines the concentration of a friendship network as the share of friends live within 200 miles). Therefore, I take 200 miles as a cut-off value, however, it is more convincing if we could compare between different cut-off value and to see how it will affect the magnitude of the coefficients and their statistically significant. This could be an expansion for the future research.

4.4. Control variables

I select control variables from those is shown to be important in previous studies. I adopt the three control variables from Harrison et al (2019) including the total non-white population, total population under 19, and total population above 65. They mention that poverty tend to concentrate for some racial and ethnic minority groups in the U.S because discrimination happen in jobs and in hiring process for some ethnic minority groups as well as lower levels of education among these groups make them easier to be in poverty. Total population under 19 and above 65 are included as control variables to represent the population who is not in the work force, therefore, may affect the poverty rate of that county.

5 Results

5.1 Main results

Comparable maps between highest MDI rate and lowest MDI rate counties

As an illustrating example to show the difference in the social connectedness map between county with highest MDI rate and county with lowest MDI rate, I compare the social

connectedness maps of Cape May County, NJ and Hamilton County, IN respectively. To construct the map, I use the Social Connectedness normalized by dividing the Social Connectedness Index by the product of county-level population. I show the heat map in figure 1 which shows the relative probability social connectedness map in Cape May County, NJ (Figure 1A) – which has the highest MDI rate, and Hamilton County, IN (Figure 1B) – which has the lowest MDI rate.

Overall, for both Cape May County and Hamilton County, they are more socially connected to counties that are geographically close to them. For example, for Hamilton County, IN, the darkest color is around the Indiana state. Likewise, for Cape May County, NJ, the darkest color is around their New Jersey state and around North East region. In addition, the main difference for the two counties social connectedness map is on their connection with two different regions. While Hamilton County, IN is mostly connected with the nearest geographically counties to them which is the Mid-West region, Cape May County does not have strong connection with this area. The similar pattern applies for the North East and South Atlantic regions. While Cape May County, NJ is mostly connected to these two regions, Hamilton County does not have strong connection with this area. For the rest of the regions in the US, these two counties have quite similar pattern in social connectedness.

To illustrate more clearly in their difference in terms of social connectedness level, I illustrate with linear interpolation in figure 2 while for figure 1, it is deciles interpolation. With linear interpolation, we can see that Cape May County, NJ does not have much difference in their social connectedness between counties, except for their own county. For Hamilton County, their social connectedness is not only strong with their nearest neighborhood in terms of geography, but also in some counties in Mid-West, West South Central and West Mountain regions who are not geographically close to them. Hence the social connectedness level in Hamilton County is more diverse compared to Cape May County, who is mostly strong in their own neighborhood. This may reflect the idea of Burt (2004) which concludes that highly clustered communities limit their opportunities, and communities having more diverse network have better performance.

Figure 1. County-Level Social Connectedness Maps (Deciles Interpolation)

Figure 1A. Social Connectedness Map of Cape May County, NJ

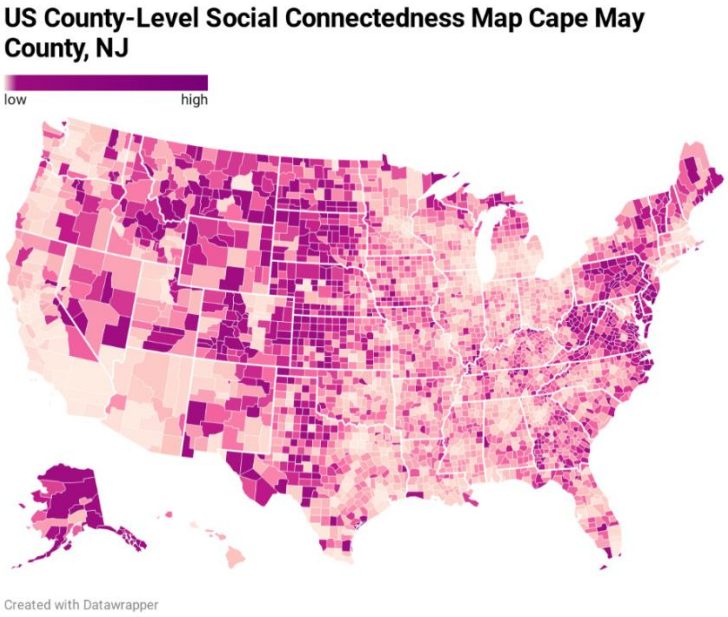


Figure 1B. Social Connectedness Map of Hamilton County, IN

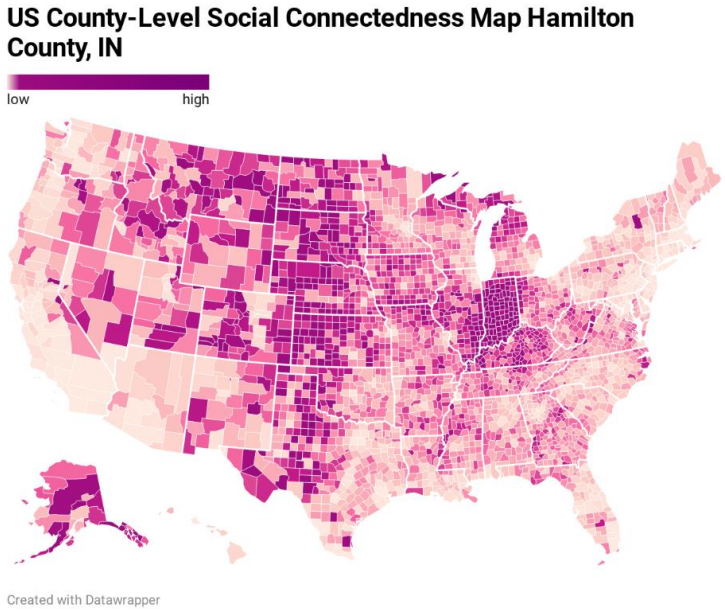


Figure 2. County-Level Social Connectedness Maps (Linear Interpolation)

Figure 2A. Social Connectedness Map of Cape May County, NJ

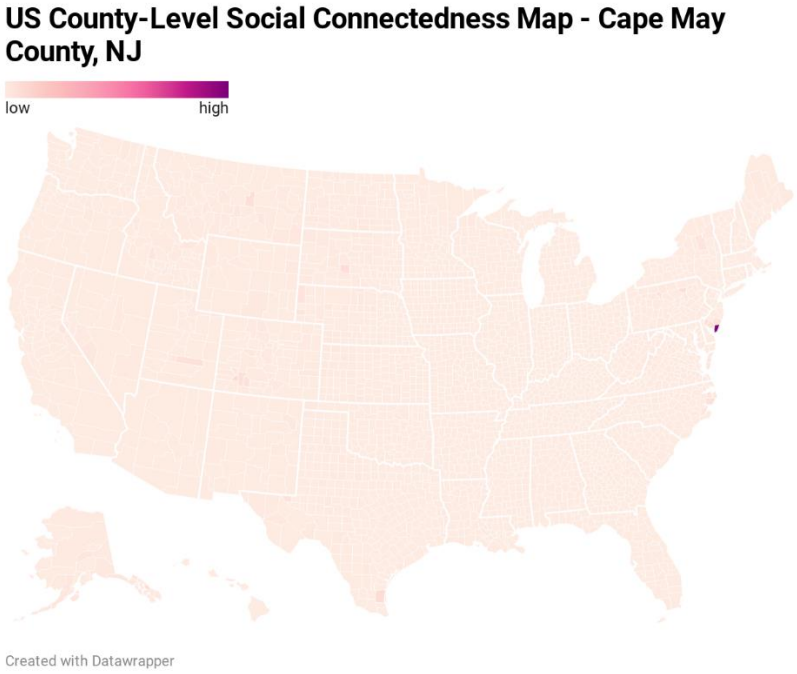
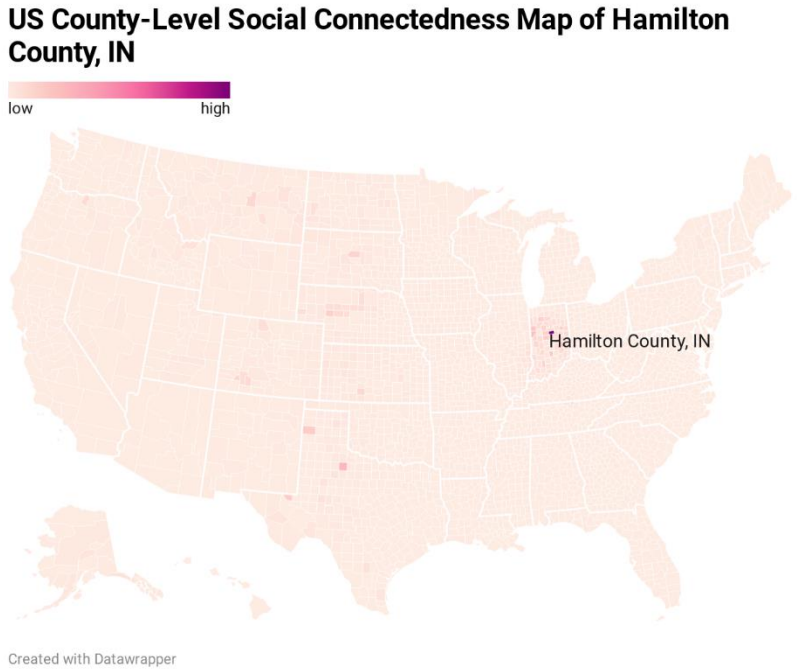


Figure 2B. Social Connectedness Map of Hamilton County, IN



Social Diversity and Economic Development

The above comparable maps between two counties with highest and lowest MDI rate show a difference in their social connectedness map. To conduct a more general view on all counties in the US, in this section, I apply the method from Eagle et al (2010) as discussed in the section 4. Figure 3 shows a comparable map between MDI rate and social diversity score of counties in the US. To construct the map, I use the Social Diversity score which ranges from 0 to 1 for all counties, instead of Social Connectedness Index. I show the heat map in figure 3 which shows the social diversity map (Figure 3A) and the MDI-inverse map (Figure 3B). To have a better visualization and find the similarity between the two maps, I use deciles interpolation and use an MDI-inverse rate rather than the MDI rate – which means the darker the color, the better performance that this county has. There are some similar patterns between the two maps which indicate that counties with diverse communication patterns tend to have better performance, for example in the North East region, a part of the Mid-West region, a part of the West Mountain region, and a part of the West South and East South-Central region. However, some regions show opposite pattern in social connectedness and economically healthy performance. For example, the West Coast region, a part of the Mid-West region, and a part of the South Atlantic region. These regions have diverse communication patterns (darker color for social diversity), but worse economically healthy performance or vice versa, have low social diversity score, but higher overall economic and health performance. This implies that social diversity is correlated to economic development, but only have a low correlation. There are other unexplained reasons for economic development that are not covered by social diversity.

To be more precise on the correlation level, I conduct a correlation test between social network diversity and socioeconomic performance. I use MDI rate for socioeconomic performance, so it is predicted that the two variables should show negative relationship. I found a weak negative correlation between social diversity and MDI rate ($r = -0.175$) in line with the illustrative maps of these two variables in figure 3. Figure 4 shows a correlation map between social network diversity and MDI rate. A line was fit to the data which is a downward slope straight line to show the negative relationship between these two variables.

Figure 3. County-Level Social Diversity and MDI-inverse Maps (Deciles Interpolation)

Figure 3A. Social Diversity Map

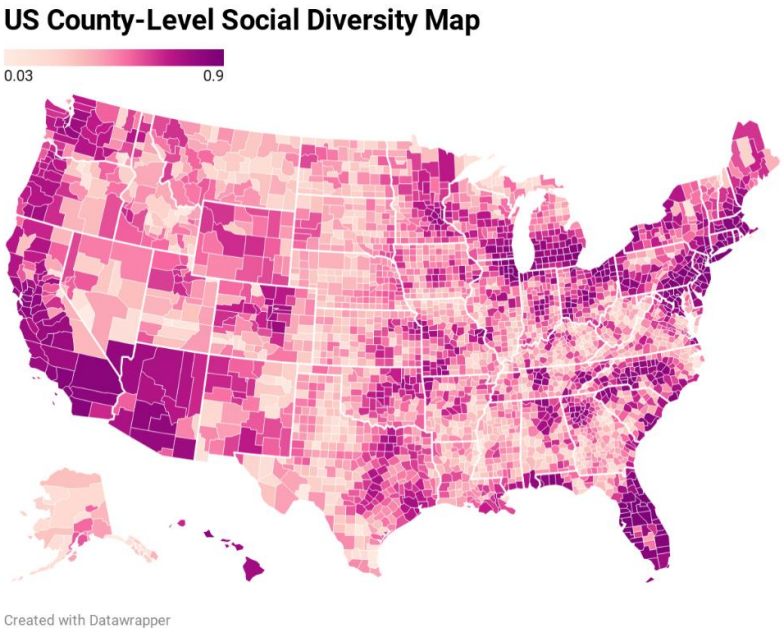


Figure 3B. MDI-inverse Map

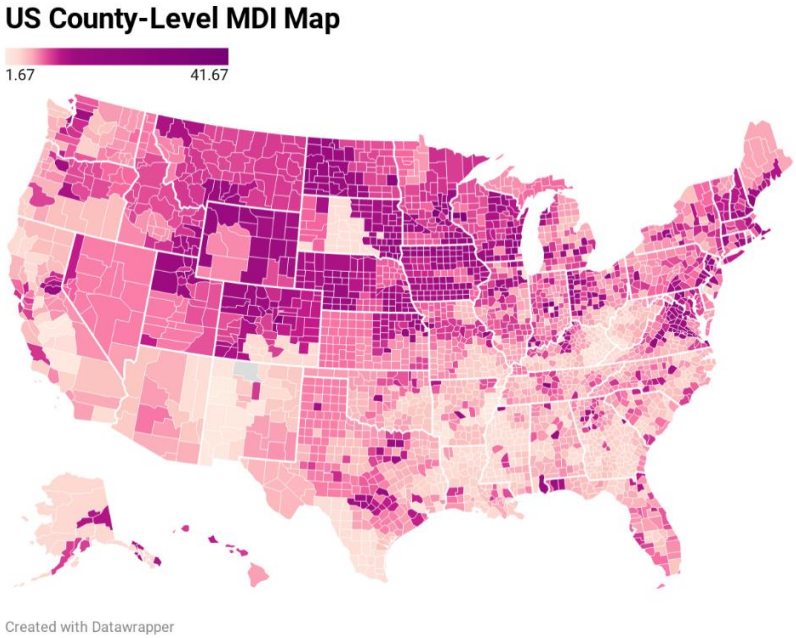
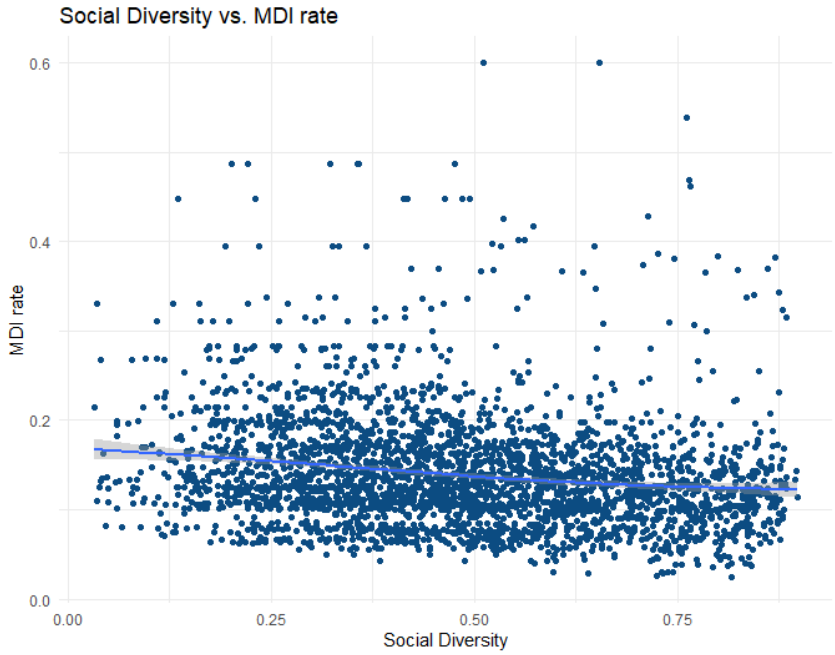


Figure 4. A correlation map between social diversity and MDI rate



Regression Results with Social Diversity

Table 1: Regression model for MDI rate with Social Diversity

	Dependent variable: MDI rate	
	(1)	(2)
Panel A. OLS Regression		
Social Diversity	-0.065 (***) (0.006)	-0.055 (***) (0.006)
Panel B. Spatial Regression		
Social Diversity		-0.033 (direct) (***) -0.043 (indirect) (***) -0.075 (total) (***)
State fixed effects		x
County Demographics	x	x

I conduct a regression model for MDI rate with social diversity as an independent variable (model I) to quantify the relationships between social diversity and MDI rate. As described above, some control variables are added into the model, including non-white population, total population under 19 and total population above 65. These control variables are important in prior research about poverty rate. The model (1) in table 1 has social diversity as an independent variable and county demographics as control variables. The coefficient for social

diversity is -0.065 and statistically significant at 1% level. I can interpret as one increase in social diversity level is correlated with a 0.065% decrease in the MDI rate when all other explanatory variables are held constant. The model (2) in table 1 also has similar variables like in model (1), except that I include state fixed effects into the model to control for state-variant characteristics. The coefficient for social diversity decreases a bit in absolute value from -0.065 to -0.055, and it is still statistically significant. I select the model (2) as my preferred model and will conduct other diagnostic tests based on this model.

I continue to check for the model assumptions for linear regression models, including linearity, independence, equal variance, and normality. To check for these assumptions, I use residual plot and QQ-plot, as shown in figure A1 and A2. Figure A2 shows that the Q-Q plot has a departure from the straight line at the end which means there is right skewness in the residuals. This also reflects in figure A1 with the residual plots where some of the residuals at the larger values of fitted values show the right skewness, while other remains to be scattered randomly and around the 0 value. To test for heteroskedasticity, I perform Breusch-Pagan test and the result confirms that there is heteroskedasticity in my model. There are many ways to fix heteroskedasticity in the model, and the most common way is to use robust standard errors. I ran coeftest function in R to use robust standard errors and the results are included in table A3.

Until now, I have not considered geography into my model. Eagle et al (2010) when performing social diversity, they also construct a measure for spatial diversity. So I will check whether there is a spatial characteristic of my data. As shown in figure 3B, MDI rate seems to concentrate on some areas which have higher MDI than the average of the national MDI rate. I perform Moran's I test for the model (2) in table 1 as shown in table A1. The Moran's I statistic is 0.584 which seems that there is a positive spatial dependence in MDI rate between neighboring counties. I also apply local Moran's I test to check for local spatial dependence within each county. Using this test, I could check for spatial dependence within each county, and for illustration, I show 6 first counties in the table A2. The results from the table show that there is no local spatial dependence in MDI rate within each county as their E.Ii are 0.

Since the test statistics show that there is a spatial dependence among neighboring counties in MDI rate, I try to fit the model with spatial lag model. The results are reported in panel B of table 1. I only perform spatial lag model for model (2). For spatial lag models, there are three coefficients for social diversity. One shows the direct effect (or the local effect), another shows indirect effect (or the spillover effect), and the last one shows total effect (or sum of local and spillover effect). The total effect of social diversity in spatial lag model is -0.075 which means that an increase in social diversity level decrease the MDI rate by 0.075% - which is greater than coefficients of both models in OLS regression.

Regression Results with Social Connectedness

In this part, I will try to answer the question about the spillover effects of MDI rate across counties that are social connected with each other, which means to estimate the extent of spatial and social MDI rate spillovers between the U.S counties. The above spatial lag model only considers spatial dependence and base on social diversity, rather than direct from social connectedness. The econometric specification model is illustrated in part 3.2 (model II). The result from this econometric model is reported in table 2.

Table 2: Regression model for MDI rate with Social Connectedness

	Dependent variable: MDI rate		
	(1)	(2)	(3)
Panel A. OLS Regression			
Geography W MDI_{jc}	0.899 (***)	0.898 (***)	0.906 (***)
	(0.005)	(0.005)	(0.006)
Social Connectedness W MDI_{jc}	0.059 (***)	0.059 (***)	0.051 (***)
	(0.004)	(0.004)	(0.004)
State fixed effects			x
County Demographics		x	x

Table 2 presents the baseline results for the equation (2). First, in column (1), I show the effect of both social connectedness weight matrix and geographical weight matrix included in the model without any covariates. The coefficient for geographical weight matrix MDIjc is 0.899 and statistically significant at 1% level. The coefficient for social connectedness weight matrix

MDIjc is 0.059 and statistically significant at 1% level. In column (2), I include county demographics as control variables, as described in part 4. The coefficients for both weight matrix do not change much. In column (3), I include state fix effects. The coefficient for geography W MDIjc increases from 0.899 to 0.906 and statistically significant at 1% level. The coefficient for social sconnectedness W MDIjc decreases from 0.059 to 0.051 and also statistically significant at 1% level. I will use model (3) as my preferred model to check for other OLS assumptions.

To interpret the magnitude of the above coefficients, I apply the interpretation from Amarasinghe et al (2018). I illustrate with an example with three counties: county 1, 2, and 3. I assume the social connectedness within a county equals 0, and the social connectedness between 1 and 2, 1 and 3, and 2 and 3 is 1 (means that the normalized social connectedness between these counties are greater than the median). For example, with the coefficient of the social connectedness W MDIjc is 0.051, it means that a 1% increase of the MDI rate of county 2 (or county 3) will increase the MDI rate by 0.051%.

To check for OLS assumptions, I conduct several tests. First, for the assumption that residuals are not related with the predicted value of outcome variable or to the value of independent variables (homoscedasticity), I plotted residuals versus the outcome variable (MDIi) and values of independent variables (the geographical weight matrix*MDIj and the social connectedness weight matrix*MDIj). Both the residuals versus fitted values and the Q-Q plot (figure A3 and figure A4) show some patterns. The Q-Q plot shows that the standardized residuals depart from the straight line at the end and at the beginning of the line, which seems to have some right skewness in the residuals. To confirm with statistics for the existence of heteroskedasticity, I also use Breusch-Pagan test to check for heteroskedasticity. The result shows that the p-value is smaller than 0.05 which means statistically significant, so I can reject the null and can conclude that there is heteroskedasticity in my model.

5.2 Robustness check

Robustness check for regression with Social Diversity

I did robustness check for regression with social diversity as an independent variable. It might be the case that some of the counties with high social diversity drive the results. Therefore, in panel A of table 3, I did the same regression models as table 1, but exclude the top 10 percent counties with highest social diversity. The coefficients for social diversity do not change much compared with the table 1. Both coefficients are statistically significant. Therefore, these coefficients are robust even when I exclude the top 10 percent counties with highest social diversity. In panel B of table 3, I also performed the same regression as in table 1, but I only include in the models the top 10% counties with highest social diversity. Both coefficients are not statistically significant, it means that I cannot reject the null hypothesis that there is no effect of social diversity on MDI rate. It means my results are not robust if only include the top 10% percent counties in social diversity. However, it also explains the pattern we see in figure 3A that some counties have darker color in social diversity, but lighter color in MDI rate, which means they are more social connected, but perform worse in economic and health indicators. Therefore, being social connected only explains a part in economic development, there are more variables deciding on the level of economic development.

Table 3: Robusness check for regression models with Social Diversity

	Dependent variable: MDI rate	
	(1)	(2)
Panel A. Exclude top 10 percent counties in social diversity		
Social Diversity	-0.071 (***)	-0.056 (***)
	(0.008)	(0.007)
Panel B. Only include top 10 percent counties in social diversity		
Social Diversity	-0.004	0.101
	(0.111)	(0.137)
State fixed effects		x
County Demographics	x	x

Robustness check for regression with Social Connectedness

I also conduct robustness check for regression models with Social Connectedness for two groups excluding the top 10 percent counties with highest social diversity, and only including the top 10 percent counties with highest social diversity. However, as reported in the results

by R, the model is essentially perfect fit, and the summary may be unreliable. The reason for this maybe that I overfit the data or the estimated effect of social connectedness weight MDIjc equals 0 for these two subgroups. Due to this problem, I do not report the results for my robustness check for regression models with Social Connectedness.

6 Limitations and Future Work

First, a problem with the above models is endogeneity. Jackson (2014) discusses the importance to understand network formation and the endogeneity of network as people tend to form relationships with others that are similar to them or to share economic benefits. This homophily formation may be driven by unobserved characteristics that could be hard for researchers to measure these characteristics. One way to adjust for endogeneity of networks as mentioned by Jackson et al (2016) is to use instrumental variables. Acemoglu et al (2019) use neighbors' colonial history as an instrumental variable for the network of municipalities in Colombia. Some instrumental variables are used by Harrison et al (2019) to estimate a spatial, simultaneous model of social capital and poverty in the US are ethnic heterogeneity and same county variables. The problem of endogeneity in the model also comes from the reverse causality where depreciation creates incentives for poor people to move between counties, rather than being socially and geographically connected with other counties affect MDI rate of that county. Diemer (2019) in evaluating the spatial diffusion of local economic shocks in the US also uses the Facebook social connectedness index and in adjusting for endogeneity, he uses migration flows between all county pairs. In the context of my model, using migration flows between counties could be a potential instrumental variable or the ancestry historical immigration. The migration flows data between counties could be obtained from the Internal Revenue Service (IRS) Statistics of Income Division. They have the available data at state and county level from 1991 to 2018. The ancestry historical immigration can be obtained from the datasets collected by Bailey et al (2018). The ancestry historical immigration could a good instrument as it was determined a long time ago and may no longer affect MDI. However, the migration flow as an instrument could also raise some concerns as whether it is truly exogenous or not, and if it is the case where the migration flow is endogenous with the error

terms, the instrumental variable estimation may even more biased than the OLS. Within the limited time, I could not conduct 2SLS for these two instrumental variables to compare between both, and to compare with the OLS and spatial models. However, in the future work, it is essential to adjust for endogeneity in the models.

Second, in evaluating the spill over effects of being socially connected, I use a weight matrix for social connectedness, with having normalized social connectedness greater than the median equals 1, and smaller than the median equals 0. This division may make us difficult to understand the spill over effects of counties who have social connectedness close to the cut-off level, but in different sides of the cut-off level. Therefore, in the future work, one could, instead of building a weight matrix of 1 and 0, for each county, divide into 20 bins of 100 social connected counties with 5 counties for each bin. In this way, we could avoid using the cut-off level. The same could apply for geography weight matrix. It is also more convincing if we could compare between different cut-off value and to see how it will affect the magnitude of the coefficients and their statistically significant. In addition, one could also build a geography weight matrix base on being a neighbor with each other. Another limitation in my geography and social connectedness weight matrix is that I did not row normalize my matrix. In the future work, it is important in creating these matrix to row normalize ensure consistency.

Finally, in my paper, I have not evaluated with other network diversity metrics, for example Eagle et al (2010) also use Burt's measure of "structural holes" or Amarasinghe et al (2018) use centrality measures, for example betweenness centrality, eigenvector central city, Katz-Boncich centrality to measure key player centrality or to answer the question which district, once removed, will reduce total nighttime lights the most (their dependent variable is nighttime lights). This same method could also be applied into my paper to answer the question of the key-player could affect other counties' MDI rates most. This could be an interesting question to answer in the future work.

7 Conclusion

This paper has studied a population-level social network to estimate the relationship between social network and community economic development. My empirical evidence has again confirmed that network diversity is associated with economic development, which means being more socially connected is correlated with better economic development performance. However, a causal relationship cannot be inferred from my models. Despite this limitation, it is still worthy to note that on population level, the structure of network does affect socio-economic performance, expanding the conclusion from Eagle et al (2010) by using an online network data to proxy for real life relationships.

I first build a map of social diversity and MDI rate, then perform a correlation map between these two variables to understand their relationship. I go further to quantify this relationship by regressing MDI rate on social diversity and can find the negative relationship between these two variables held other variables constant, and this coefficient is robust with different models and with a subset of excluding top 10 percent counties with highest social connectedness score. However, this coefficient is not robust if I only include top 10 percent counties with highest social connectedness score in my model, and this can explain why some counties who are strongly social connected with other counties, but have higher MDI rates, for example some West Coast counties. It means that being socially connected only answers a part of the question about economic development, and there are more variables should be considered.

I go one step further by using another regression model to estimate the social and spatial spill over effects of MDI rate. The results show that there are spillover effects of MDI rate for counties who are being geographically connected, or socially connected with each other. Results from both regression models cannot be assumed as causal relationship due to endogeneity problem of being socially connected with each other. Future work could try to establish the causal mechanism between network diversity and economic development as well as the social and spatial spillover effects between counties.

8 References

Acemoglu, D., García-Jimeno, C., & Robinson, J. A. (2015). State capacity and economic development: A network approach. *American Economic Review*, *105*(8), 2364-2409.

Amarasinghe, A., Hodler, R., Raschky, P., & Zenou, Y. (2018). Spatial diffusion of economic shocks in networks.

Aral, S., & Van Alstyne, M. (2008). Networks, information & social capital.

Bailey, M., Cao, R., Kuchler, T., & Stroebel, J. (2018a). The economic effects of social networks: Evidence from the housing market. *Journal of Political Economy*, *126*(6), 2224-2276.

Bailey, M., Cao, R., Kuchler, T., Stroebel, J., & Wong, A. (2018b). Social connectedness: Measurement, determinants, and effects. *Journal of Economic Perspectives*, *32*(3), 259-80.

Bailey, M., Gupta, A., Hillenbrand, S., Kuchler, T., Richmond, R., & Stroebel, J. (2020). International trade and social connectedness. *Journal of International Economics*, 103418.

Bandiera, O., & Rasul, I. (2006). Social networks and technology adoption in northern Mozambique. *The economic journal*, *116*(514), 869-902.

Banerjee, A., Chandrasekhar, A. G., Duflo, E., & Jackson, M. O. (2013). The diffusion of microfinance. *Science*, *341*(6144).

Barr, A. (2000). Social capital and technical information flows in the Ghanaian manufacturing sector. *Oxford Economic Papers*, *52*(3), 539-559.

Bertrand, M., Luttmer, E. F., & Mullainathan, S. (2000). Network effects and welfare cultures. *The Quarterly Journal of Economics*, *115*(3), 1019-1055.

Burt, R. S. (2004). Structural holes and good ideas. *American journal of sociology*, *110*(2), 349-399.

Chaney, T. (2014). The network structure of international trade. *American Economic Review*, *104*(11), 3600-3634.

Coleman, J. S. (1988). Social capital in the creation of human capital. *American journal of sociology*, *94*, S95-S120.

Conley, T., & Christopher, U. (2001). Social learning through networks: The adoption of new agricultural technologies in Ghana. *American Journal of Agricultural Economics*, *83*(3), 668-673.

Diemer, A. (2020a). Spatial diffusion of local economic shocks in social networks: evidence from the US fracking boom.

Diemer, A., & Regan, T. (2020b). No inventor is an island: social connectedness and the geography of knowledge flows in the US (No. dp1731). Centre for Economic Performance, LSE.

Donaldson, D., & Storeygard, A. (2016). The view from above: Applications of satellite data in economics. *Journal of Economic Perspectives*, *30*(4), 171-98.

Eagle, N., Macy, M., & Claxton, R. (2010). Network diversity and economic development. *Science*, *328*(5981), 1029-1031.

Engstrom, R., Hersh, J., & Newhouse, D. (2017). Poverty from space: using high-resolution satellite imagery for estimating economic well-being.

Gee, L. K., Jones, J. J., Fariss, C. J., Burke, M., & Fowler, J. H. (2017). The paradox of weak ties in 55 countries. *Journal of Economic Behavior & Organization*, *133*, 362-372.

Granovetter, M. (2005). The impact of social structure on economic outcomes. *Journal of economic perspectives*, *19*(1), 33-50.

Guiso, L., Sapienza, P., & Zingales, L. (2009). Cultural biases in economic exchange?. *The quarterly journal of economics*, *124*(3), 1095-1131.

Harrison, J. L., Montgomery, C. A., & Jeanty, P. W. (2019). A spatial, simultaneous model of social capital and poverty. *Journal of behavioral and experimental economics*, *78*, 183-192.

Jackson, M. O. (2014). Networks in the understanding of economic behaviors. *Journal of Economic Perspectives*, *28*(4), 3-22.

Jackson, M. O., Rogers, B., & Zenou, Y. (2016). Networks: An economic perspective.

Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., & Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. *Science*, *353*(6301), 790-794.

Ludwig, J., Duncan, G. J., Gennetian, L. A., Katz, L. F., Kessler, R. C., Kling, J. R., & Sanbonmatsu, L. (2013). Long-term neighborhood effects on low-income families: Evidence from Moving to Opportunity. *American economic review*, *103*(3), 226-31.

Morales, E., Sheu, G., & Zahler, A. (2019). Extended gravity. *The Review of Economic Studies*, *86*(6), 2668-2712.

Patacchini, E., & Zenou, Y. (2008). The strength of weak ties in crime. *European Economic Review*, *52*(2), 209-236.

Pew Research (2019). Share of U.S. adults using social media, including Facebook, is mostly unchanged since 2018.

U.S. Census Bureau (2019). Multidimensional Deprivation in the United States: 2017. American Community Survey Report.

Woolcock, M., & Narayan, D. (2000). Social capital: Implications for development theory, research, and policy. *The world bank research observer*, *15*(2), 225-249.

9 Appendix

Table A1. Moran's I test

```
        Moran I test under randomisation

data:  merge_sc_counties_sf$`MDI rate`
weights: listw_scB  n reduced by no-neighbour observations


Moran I statistic standard deviate = 55.86, p-value < 2.2e-16
alternative hypothesis: two.sided
sample estimates:
Moran I statistic       Expectation          Variance
     0.5840912819     -0.0003195909      0.0001094546
```

Table A2. Local Moran's I test

	Ii	E.Ii	Var.Ii	Z.Ii	Pr(z > 0)
01001	0.030	0	4.980	0.010	0.490
01003	-2.010	0	5.970	-0.820	0.790
01005	14.580	0	7.960	5.170	0
01007	4.240	0	5.970	1.730	0.040
01009	-2.790	0	5.970	-1.140	0.870
01011	12.760	0	4.980	5.720	0

Table A3. Robust Standard Errors

	Dependent variable:
Dsoc	-0.055*** (0.007)
non_white_sum	0.00000*** (0.00000)
under19	0.00000 (0.00000)
above65	-0.00000*** (0.00000)

Note: I do not include the results for dummy variables for state fixed effects in table A3

Figure A1. Residual Plot with social diversity as independent variable

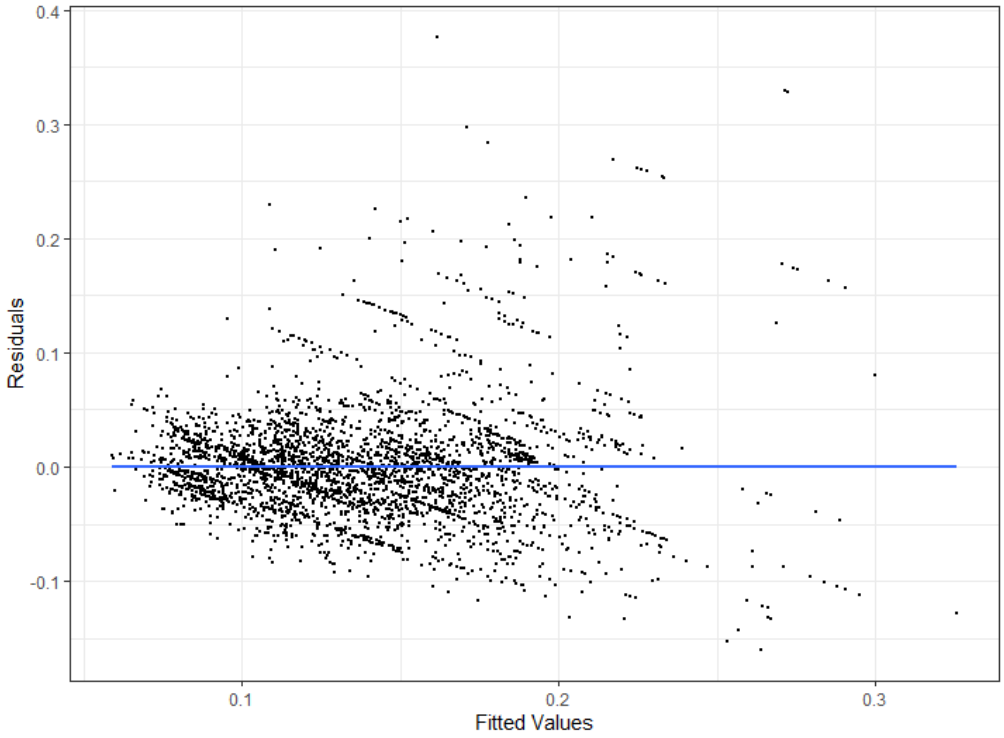


Figure A2. Q-Q Plot with social diversity as independent variable

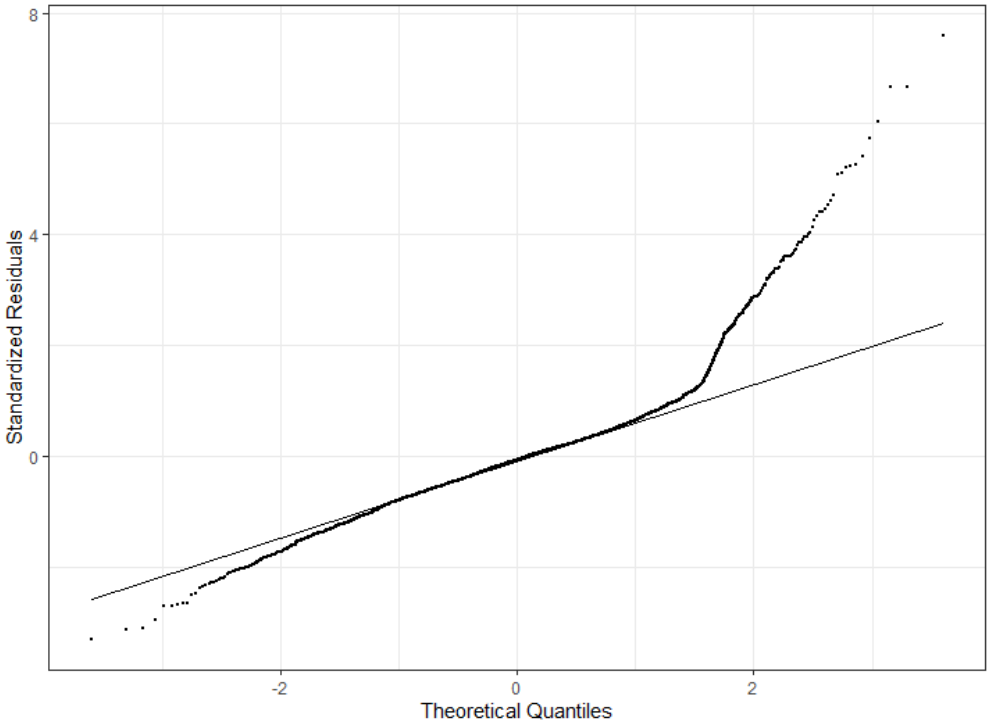


Figure A3. Residual Plot with Social Connectedness as independent variable

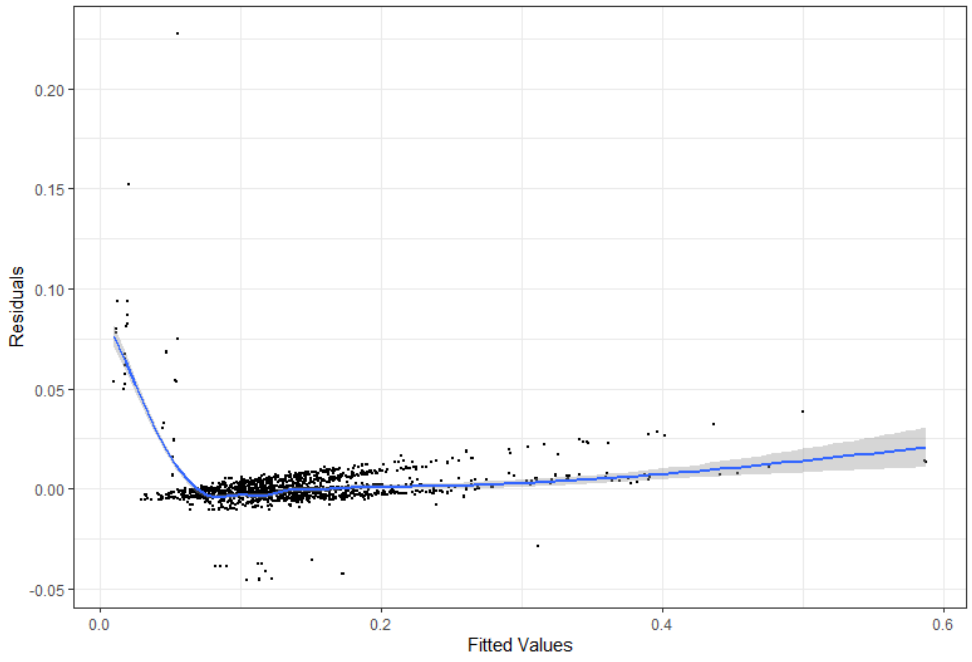


Figure A4. QQ Plot with Social Connectedness as independent variable