Utrecht University

Master Thesis Urban & Economic Geography

A Study on How Transport Networks Spread Electrical Technology in 19th Century US

No Train, no Gain

Author:
Maarten Schellens

Supervisor:
Dr. Pierre-Alexandre Balland

June 3, 2018

Universiteit Utrecht

Abstract

Proximity between inventive actors is important for the spread of new ideas. However, in the current literature there is little attention to measures of accessibility - a form of functional proximity - which depend on infrastructural connections. This article tries to fill this gap by analysing the effect that railroad, canal and river connections have on the diffusion of electrical technology in US counties between 1850-1900. The results of the linear probability analysis show that both railroads and canals are linked to the probability of patenting in electrical technology within a county.

1 Introduction

In the past two centuries, technological diffusion has been a key ingredient shaping the difference in economic growth between countries (Baldwin, 2016). The great divergence in economic growth since 1820 between the Western world and the rest of the globe is partly attributable to the fact that areas can have geographical, environmental and social circumstances that impede the diffusion of new technologies (Bloom et al., 1998). Bloom et al. (1998) for instance, attribute Africa's dissapointing technological and economic growth to unfavourable geographical circumstances, such as having a relatively short coast line compared to the hinterland coupled with few navigable rivers inland. However this focus on geographical conditions doesn't explain the technological blossoming of the US in the same period, under comparable geographical circumstances as Africa. One of the reasons for the US's remarkable growth in the past two centuries is suggested to lie in the 'transportation revolution' (Taylor, 1951) of canals and railroads, that connected much of inland America to the coastal economic hubs. It is this key role of infrastructural connections in technological development that warrants further research, starting with this paper.

Proximity between actors is an important mediator for the success of knowledge diffusion. Cognitive, social, organisational and institutional proximity can link up pools of knowledge to create new knowledge (Boschma, 2005; Balland, 2012). Geographical proximity - simply being located close to one another in space and time - matters for innovative activity (Jaffe et al., 1993; Sonn & Storper, 2008; Feldman et al., 2015). Recently, Feldman et al. (2015) have shown that the Euclidean distance between MSAs is a significant factor in explaining how new recombinant DNA (rDNA) technology spreads between them. While being interesting in their own respect, the proximity and diffusion literature tends to focus on absolute geographical proximity, and leaves the importance of functional proximity, or accessibility (Andersson & Karlsson, 2004), untouched. There is little attention to the improvements in transportation and communication that have greatly increased accessibility of cities and towns in the past centuries, and subsequently facilitated greater possibilities for inventive flows between actors.

On the other hand, the literature that does focus on the effect that infrastructure has on diffusion of invention is scarce, and does not depart from a proximity framework or focus on the diffusion of a single technology. A topic that is often researched in economic papers, but less so in innovation papers, is that of the development of the railroad network in the US. This infrastructural revolution is a suitable object of study, since while road transport cost did not change from the early 19th century to the early 20th, railroads offered an alternative that connected cities on a longer distance than ever before, being cheaper than the transportation alternatives (Perlman, 2016). It has been shown by Perlman (2016) that areas that have been newly connected to the railroad network have seen an increase in innovation, as measured by the average amount of patents per 1,000 people.

To connect the loose ends of the proximity and transportation literature, I make a longitudinal analysis of historical transportation connections in 19th century US. Specifically, I will focus on US counties in the period between 1850-1900, when big transport innovations such as railways and canals changed the economic and technological landscape (Perlman, 2016). This historical approach is made possible by novel data initiatives on the geographical distribution of patented knowledge (Petralia et al., 2016) and infrastructural connections (Atack, 2016). Specifically, I will focus on the diffusion of patents in the Electrical & Electronic (E&E) class. To measure diffusion, the use of a single, albeit broad class of technology is an advantage over using the aggregate of all patents, which could be of all kinds of technologies and thus signify the diffusion of very unrelated knowledge. The E&E class of technology possesses the traits of a General Purpose Technology (GPT), making it a suitable subject of study when tracking diffusion of technology (Petralia, 2017). Since this technology is both new and suitable for general applications, it can serve as a realistic proxy for diffusion. More concretely, the aim of this paper is *to examine the effect of infrastructure accessibility on the diffusion of patented knowledge in the E&E class.*

The results of the Linear Probability Model analysis show that both railroads and canals have a positive effect on the diffusion of E&E technology, even when controlled for various contextual variables and state fixed effects. This result suggests, along with the earlier work on the diffusion of rDNA by Feldman et al. (2015), that the role of proximity deserves more attention as a mediator for technological diffusion. More specifically, the focus on accessibility as a distinct form of geographical proximity shows that transport links, rather than just being a stimulant for 'hard' economic variables, can mediate the 'invisible' process of knowledge flows.

The structure of the paper is as follows: first, theory on General Purpose Technologies, diffusion, proximity and transportation will be outlined; second, the data sources will be evaluated, thereby assessing their strengths and weaknesses; third, I will explain the choice for a linear probability model and the inclusion of the variables in the methodology section; fourth, the results of the regression are analysed; fifth and last, the results are discussed and the paper is summarised.

2 Theory

Not all inventions are of equal technological and economical value (Balland & Rigby, 2017). Whereas most inventions offer incremental changes to an already existing technology, others have the power to change the economic landscape of a county or even the world. Inventions of this magnitude are considered to be General Purpose Technologies (GPTs). As the name suggests, GPTs can be applied to many parts of the economy and thus have a greater impact than regular technologies (Bresnahan & Trajtenberg, 1995; Lipsey et al., 2005). Focusing on the second half of the 19th century, recent research by Petralia (2017) has shown that inventions classified in the *Electrical & Electronic* (E&E) patent class between 1860 and 1930 possess the characteristics of a GPT. Petralia comes to this classification because E&E technology matches the essential characteristics of having a 'wide scope for improvement and collaboration', alongside a 'potential for use in a wide variety of products and processes', and 'strong complementaries with existing or potential new technologies' (Helpman & Trajtenberg, 1998a; Helpman & Trajtenberg, 1998b; Moser & Nicholas, 2004; Jovanovic & Rousseau, 2005; Lipsey et al., 2005).

Although E&E technology eventually managed to have an astounding worldwide impact, it originated in a certain place at a certain time. Whereas the initial geographical area in which a new technology is known can be 'lumpy' (Mensch, 1975), a technology can diffuse to other places and induce inventors to find new applications. The fact that invention diffuses slowly across space and time means that for invention, location matters. More specifically, it has been shown that co-location of inventive actors in space fosters the transfer of innovative ideas (Jaffe et al., 1993). Jaffe et al. (1993) and Thompson (2006) proved that patents in the US cite other patents from the same country, state and even metropolitan statistical area in a quantity that is above average.

While the - empirically proven (Ellison et al., 2010) - agglomeration theory of Alfred Marshall (1920) focused on knowledge spillovers that occurred between actors in close proximity, later research has suggested that linkages to less geographically proximate actors are also of great importance. Bathelt et al. (2004) coined the idea of *global pipelines*, through which clusters have access to knowledge that fits their specialisation. By establishing linkages to non-local actors, technological lock-in can be prevented (Boschma, 2005). Consequently, acquiring new technology is a mix between the local and the non-local: 'In any economic system, the accumulation of knowledge depends on the economys internal capacity to produce innovation and also on its ability to acquire the stock of knowledge generated in other areas and put it to work' (Paci et al., 2014, p. 10).

How well a region can make use of external knowledge, is partly dependent on whether it proximate in a cognitive sense with regards to the knowledge that is being transmitted. It is argued that cognitive proximity indicates the way different actors in regions and industries share knowledge structures (Feldman et al., 2015). The higher the cognitive proximity, the greater is the overlap between routines, skills, institutions and knowledge, creating potential for absorptive

capacity (Cohen & Levinthal, 2000; Nooteboom, 2000). It is also argued that higher levels of cognitive proximity lead to 'enhanced collaboration as well as knowledge sharing' (Feldman et al., 2015). When viewing new technology as the recombination of old knowledge, it is argued that high levels of cognitive proximity facilitate the integration of different elements of knowledge into new knowledge (Weitzman, 1998; Fleming & Sorenson, 2001; Kogler et al., 2013; Rigby, 2013).

Although contact between actors can be established over long geographical distances - involving other proximities in the form of cognitive, social, organisational and institutional proximity -, geographical proximity is still very important (Boschma, 2005). One of the reasons for this, as is suggested by Audretsch & Feldman (1996) is that, 'Although the cost of transmitting information may be invariant to distance, presumably the cost of transmitting knowledge rises with distance' (p. 630).

Usually, geographical proximity is seen in an absolute, Euclidean sense, but there is also a functional interpretation of geographical proximity. As Lagendijk & Lorentzen (2007) argue, geographical proximity is partly 'a product of the historically accumulated construction of transport infrastructures' (p. 460). By shortening time and financial cost to connect to other actors, this functional geographical proximity influences the abilities of actors to become more cognitvely, socially, organisationally and institutionally proximate (Shaw & Gilly, 2000).

According to Coenen et al. (2004), this kind of functional proximity requires a different approach to the concept of proximity, as '... it would be more valid to understand functional proximity as accessibility rather than distance' (p. 1010). Andersson & Karlsson (2004) further elaborate on accessibility as a form of relative geographical proximity, by linking it to Hägerstrand's (1970) idea of time being a constraint to geographical proximity. Furthermore, borrowing from Weibull (1980), it can be argued that accessibility indicates, amongst others, the 'ease of spatial interaction', the 'potential of opportunities of interaction', and the 'potentiality of contacts with activities or suppliers' (p. 54). Based on these theories, Andersson & Karlsson (2004) hypothesise that:

> It is possible to claim *ceteris paribus* that a region characterised by high accessibility to face-to-face-contacts is likely to produce and diffuse new knowledge at a higher speed than a similar region with lower accessibility. Such a region is able to develop a dense human interaction network. Also, frequent contacts between regional actors imply that they are prone to developing common norms and bilateral understanding. The regional actors are likely to develop reciprocal understanding of codes essential for the sharing of tacit knowledge. Taken together, regions with high accessibility to relevant opportunities should, ceteris paribus, have a higher innovation potential and a higher innovation rate (p. 13).

Following this line of thinking, when geographical proximity facilitates invention, and transportation matters for geographical proximity, it follows that

transportation links affect the inventive output of spatial entities.

While the literature on the effect that physical infrastructure has on economic growth is quite large, the literature connected to innovation and invention is significantly smaller. Agrawal et al. (2016) show that interstate highways in the US increase patenting in nearby cities and make knowledge 'travel' over longer distances between cities. As a result of these findings, they are able to state that: 'In addition to facilitating the flow of human capital into cities (agglomeration), transportation infrastructure, such as interstate highways, lowers the cost of knowledge flows within regions between local innovators' (p.1). Crucially, the effect that roads have on invention is found to be especially large in newer fields of technology (Agrawal et al., 2016).

Returning railroad to connections, which shaped the great transport revolution of the 19th century, Perlman (2016) has shown using historical patent data that being connected to a railroad track increase innovatived output in US counties. The two suggested mechanisms for this are connectivity and increased market access. While support for the latter theory was not found, the first theory has yielded significant results: newly connected counties have a higher patent output than before the establishment of the connection, suggesting a greater spatial diffusion of knowledge (Perlman, 2016). Phillips (1992), who researched the railroad network in Virginia, further elaborates on the effect that this railroad connectivity might have, by stating that:

> [. . .] improved rail connections made it easier for prospective inventors to receive information on the viability of their ideas and engage in the patent application process. The improved access to ideas outside the region stimulated the inventive minds in the affected areas, prompting them to think about solutions to new problems or to think about old problems in new ways (p. 395).

With Agrawal et al. (2016) and Perlman (2016) being the only authors that have researched the impact of physical infrastructure on diffusion of technology on a US wide scale, there is still a great wealth of knowledge left to discover, especially since these authors didn't differenciate between different technology classes. Summarising the existing literature on proximity and knowledge diffusion, it can be stated there is still an empirical gap connecting the concepts of proximity and accessibility to diffusion of general purpose technologies. What this paper tries to find is whether absolute geographical proximity (distance), accessibility (infrastructure connections) and cognitive proximity (knowledge relatedness) determine the spatial diffusion of E&E technology across US counties. To test for this relation, I investigate the E&E patent output in all US counties between 1850 and 1900 using a linear probability model.

3 Data

This study concerns the US because it played a major role in the transportation revolution that I am assessing here, but also because there is a lot of histori-

cal data available. One of the extensive data sources that I use, the HistPat dataset (Petralia et al., 2016), contains all data on patented inventions in the US between 1790 and 1975. The inclusion of patents as indicator of inventive activity in this paper, stand within a longer tradition in economic and geographical research. Patents, defined by Griliches (1998) as 'a document, issued by an authorized governmental agency, granting the right to exclude anyone else from the production or use of a specific new device, apparatus, or process for a stated number of years' (p. 288), have been a favoured proxy for invention in economic research. Patent statistics have many applications in studies on knowledge and diffusion (Scherer, 1984; Griliches, 1998; Jaffe & Trajtenberg, 2002), but are not without its flaws either; patents are more of a juridical concept signifying an invention, rather than an indicator of all knowledge produced, as knowledge often goes unpatented (Pavitt, 1985; Griliches, 1998). For this reason, this paper prefers to call it the diffusion of invention, rather than the diffusion of knowledge. However, patents still carry valuable information, of which the assigned technology class of an invention is something that has only recently been utilised in research (Fleming & Sorenson, 2001; Nesta, 2008; Quatraro, 2010; Strumsky et al., 2012). It is this focus on technology class and the interaction between classes, that this paper will make further use of.

For the geographic analysis, I use geographically located patents, which are coded by technology class and assigned to US counties. The patents in this HistPat database (Petralia et al., 2016) are linked to shapefiles of US counties' historical borders. The border changes are updated every tenth year, which means that a patent filed between 1840 and 1849 will be assigned according to the county borders in 1840. As a fire destroyed much of the USPTO patent archives in 1836, and the next census update and establishment of county borders took place in 1840, the period under research starts in 1850 (as the covariates are lagged and recorded in 1840). As E&E patents are the variable of interest here, all patent classes that are supposed to belong to this supraclass (consisting of 54 existing classes) as indicated in the HistPat dataset have been recoded to a common class according to the taxonomy made by Hall et al. (2001).

Datawise, the counties used are identified by their FIPS code, rather than their name, since many duplicate names exist. In the dataset, some counties were not coupled to a FIPS code because they did not have one, and have subsequently been dropped. As a result, inventions made in these counties are not used in this research. This is however not the only example of a sizeable reduction of the available data. The fact that a time lag is used, has implications for the base of counties that is used. Since between 1850 and 1900 the number of counties in the US greatly increased, it is not unusual to find a county that existed in time period t, but did not yet exist in t-1. Given that the covariates are all lagged by one period, this means that only use those counties that exist in both t and t-1 are used. A result of this approach is that, for instance, of the 1618 counties that existed in the 1850-1859 period, only 1277 are used in

the analysis, since only these counties also existed between 1840-1849.[1] One could also opt for an analysis in which the county borders are held constant at the earliest border configuration. Although this method is used by Perlman (2016), I have chosen not to use it here since keeping these borders stable does no right to the historical process of constantly shifting and splitting counties. Furthermore, in the HistPat dataset, the patents have been assigned to the counties according to their actual borders at the time, making it difficult to assign patents according to an earlier border configuration.

The geographical information on railroads, canals and navigable rivers is taken from the extensive work of Atack (2016) on this subject, whose work covers all years between 1776 and 1911. Atack's (2016) data consists of geocoded lines that represent railroad, canal or river connections, which can be selected by years of operation. Canals and rivers are included since they formed the most efficient way of transport before the arrival of the railroad (Perlman, 2016), thus being a suitable control to assess the impact of the railroad. Canals reached their all-time highest mileage in 1851, right at the beginning of the period under research (Grübler, 1996). According to Grübler (1996), '...after reaching its maximum size, the canal network declined rapidly because of vicious competition from railways (p. 27-28)'. These rapidly growing railroads then experienced peak growth in 1891, right at the end of the research time frame, meaning that the period under research covers the crucial phase of emergence of the railroad and the replacement of canals as main mode of transport. The transformation of this geographical data into a variable is as follows: when a linear infrastructure segment intersects a county polygon within a ten-year bracket, a county gets coded value 1, and 0 otherwise.

The historical transportation data is not without its limitations: while trajectories of railroads, navigable rivers and canals are known, there is no large scale information on the location of train stations or ports in rivers and canals. The solution of coding counties either 1 or 0 whenever one of these connections are located within their borders is not ideal, as it is not known whether the local people actually could transport themselves or their produce via these connections. This binary approach has however been used in previous research that made use of this data source (Atack et al., 2010) which is why I am opting for the same approach.

As for the control variables, US population census data is gathered from the NHGIS project,which is updated every tenth year. Besides US census data, this dataset also includes variables that are the results of earlier scientific research. The census data is aggregated at the county level, so no further computation was needed, besides transforming some variables from absolute values to percentages

[1] An alternative approach would be to link the independent variables of time t-1 to the county borders in period t, e.g. link the 1840-1849 observations to 1850 borders. On the plus side, this results in more counties to be used in the analysis. On the downside, counties might wrongfully be considered to (not) have patented in time t-1 or to (not) have been connected because they have split from another county or merged into a new one. Both approaches have been tried, and resulted in very similar outcomes. The current approach is used in this paper, because it is more theoretically waterproof, but results wise the difference is negligible.

(Table 1), to prevent multicollinearity. Since not all variables have been collected regularly, some have been extrapolated between two existing collection dates.

4 Methodology

The analysis is made by using a Linear Probability Model (LPM). This model is a regression with a binary dependent variable, indicating whether the event - patenting in E&E - will happen or not. As a result, the probability of this event happening is always between 0 and 1. In this paper, I will only use a repeated measures approach to monitor diffusion. This means that everytime a county shows to patent in E&E, it is given value 1, rather than only giving value 1 to the very first time a county patents. The approach of using a binary dependent variable to indicate diffusion is borrowed from the approach used by Feldman et al.'s (2015) paper on the diffusion of rDNA technology and Boschma et al.'s (2014) on relatedness and technological change, as in their research an MSA gets value 1 for producing at least one patent in a specific class.[2] An alternative model, such as the Cox Hazard model, as utilised by Feldman et al. (2015) is not feasible given the quality of the available data: Hazard model analyses start at the very first occasion of a certain event, and in the case of E&E patenting, this would be before the earlier mentioned year of 1836. This would mean less reliable information on patents and their place of origin, plus the fact that the earlier in the 19th century we go, the fewer control variables are available.

To guarantee accuracy of statistical significance inf the model (Table 2), standard errors have been White clustered to control for heterogeneity. All independent variables have been centered to allow for an estimation of their relative effect on the propability of producing at least one patent in the E&E class.

The period under research, 1850-1900, is segmented into five decade-long parts. To prevent endogeneity, all covariates are lagged by one period. The reasons why these periods are ten-year, rather than shorter periods often used in research, are twofold. Firstly, ten-year periods reflect the process of developing a patent, which is often the work of years of research, trial and error. If there would be one-year or three-year brackets for instance, developing a patent in these periods could be unrelated to the work on these patents and the variables that influenced them in much earlier periods. The longer the period under research, the smaller this negative effect will be; however never zero. Too long periods on the other hand, could reduce the explanatory power of this research, because the relation between two very long periods is much less obvious (i.e. logically, it is much more likely that something in 1850 will influence something in 1860, rather than in 1900). Secondly, data availability and quality is an issue

[2]This stands in contrast to the earlier research done on railroads and the production of patents by Perlman (2016). Perlman's research focuses on the average production of patents per person regardless of technology class. The latter approach would not be accurate in this paper, since I am tracking the diffusion of a certain GPT, rather than measuring the total shift in inventive production in general.

of importance when using historical databases. Despite the work done by Atack (2016), tracing down the opening and closing years of railroad connections, there is always a risk of having a measuring error, leading to wrongly assuming that railroad X started operations sooner or later than was actually the case. Using longer time periods reduces this chance (i.e. assuming a one-year average misestimation of the opening of a connection, a ten-year period is much more accurate than a one-year period). This also accounts for the census data borrowed from the NHGIS project, which is only measured every tenth year, meaning the years in between cannot be estimated with precision.

The econometric equation that is used in the model can be written as follows:

$$PatentEE_{c,t} = \beta_1 AC_{c,t-1} + \beta_2 GP_{c,t-1} + \beta_3 CP_{c,t-1} + \beta X_{c,t-1} + \phi_c + \epsilon_{c,t}$$

The dependent variable, 'patentEE', is binary, indicating whether in period t, a county has at least one patent in class E&E. When a county patents in E&E it is assigned value 1, and 0 otherwise.

The most important independent variables in this research are the accessibility variables, 'AC', consisting of the variables 'RRconnect', 'Riverconnect' and 'Canalconnect', indicating whether a county is connected to a railroad, navigable river or canal in period t-1. Like the dependent variable, these accessibility variables are binary, giving value 1 when a county is connected to a form of infrastructure, and 0 otherwise.

Three measures of absolute geographical proximity, 'GP', have been tried in the model, and are calculated the same way as in the Feldman et al. (2015) paper, by measuring the distance in kilometers between the centroids of the county boundaries. The first measure is the average distance from any county to all other counties that patented in E&E in period t-1. If the county itself patented in E&E, it is included, thus reducing the average distance to all counties that have patented in this class. The second measure is the distance between a county and the closest county that has patented in E&E in period t-1. Again, the county itself is included, assigning value 0 to this county if it has indeed patented in this class. The third measure is measuring the mean distance from a county to all other counties, regardless of whether these counties have patented in the class under research. Due to multicollinearity issues and the fact that the minimum distance between a county and a patenting county is the most robust under different model specifications, only this measure is included in the final model. Interestingly, the 'mindist' variable was also the most stable of the three geographical proximity variables in the Feldman et al. (2015) paper, perhaps suggesting that this measure has predictive power both in 19th and 20th century diffusion of GPTs.

In its operative form, the variable 'CP' shows the relatedness of a county's inventive portfolio to the class of E&E. Cognitive proximity is an index number based on the co-occurence of any class with the E&E class. This measure was constructed the same way as in the Feldman et al. (2015) paper, besides the fact that I have recoded multiple existing classes into the E&E class. Mathematically,

this measure consists of some simple steps. Firstly, $F_{ip} = 1$ if a patent record p lists technology class i, and 0 otherwise. Here, the technology classes are consisting of the original ones as coded by the USPTO, but with 54 classes recoded to the E&E class. Secondly, for each of the five decades under research here, the total number of patents listing a specific class is given by $N_i = \sum_p F_{ip}$. Thirdly, I calculate how often two patent classes appear together on the same patent with the count $N_{ij} = \sum_p F_{ip}F_{jp}$. Lastly, to get the standardised co-occurence matrix, I use the following equation:

$$S_{ij} = \frac{N_{ij}}{\sqrt{N_i * N_j}}$$

What we have now is a symmetric matrix with a principal diagonal that is given value 1, showing the standardised co-occurence of every technology class. Now, to show how related a county's knowledge in period t is to E&E technology, I use calculate the following:

$$AR^{ct} = \frac{\sum_j S^t_{CBj} * D^{ct}_j}{N^{ct}}$$

Here, AR^{ct} is the measure that is a proxy for cognitive proximity: it is the average relatedness index value for a county c in decade t. S^t_{CBj} is the technological relatedness between E&E patents and patents in j other technology classes, including E&E itself. D^{ct}_j is the number of patents in technology j in county c in decade t. N^{ct} represents the total amount of patents in a county c in decade t. Since the E&E class is much broader than the rDNA class (which is only a subclass of one technology class) used by Feldman et al. (2015), the cognitive proximity to E&E that comes out of the above equation might be much less robust, as it will co-occur with many more patent classes than rDNA does. The variable name used for cognitive proximity is 'CogProx'.

The variable 'X' indicates the use of several time-lagged control variables, including the log of the population in a county, the percentage of people living in urban areas ($>$ 2,500 people), the percentage of people born outside the US, the average amount of patents per 1,000 people, the percentage of improved land, the percentage of people working in manufacturing, and the percentage of people in school. The control variables are based on the first year of each period, e.g. for the 1850-1859 period, the census data of 1850 is used.

ϕ is a state fixed effect, controlling for omitted variables relating to institutional circumstances. These institutional factors might include state laws, taxation and culture. ϵ is the error term.

5 Analysis

In multiple different model setups, the connection variables of interest, RRconnect and Canalconnect, remain positive and significant (Table 2). The first three setups of the model consist of regressing the 'patentEE' variable on just one transport variable at a time. The results show that both railroad and canal

Table 1: Descriptive statistics

Variables	Observed	Mean	Standard deviation	Min	Max
patentEE	9788	0.23	0.42	0	1
RRconnect	9788	0.55	0.50	0	1
Riverconnect	9797	0.34	0.47	0	1
Canalconnect	9791	0.08	0.28	0	1
mindist	9788	106.84	123.16	0	2763.78
CogProx	9788	0.01	0.05	0	1.04
pop	9779	9.41	1.03	1.39	14.23
urbpct	9778	0.08	0.18	0	1
forbornper	9778	0.09	0.12	0	0.80
relpatpp	9779	0.79	1.85	0	100
schoolpct	9175	0.20	0.14	0	0.95
implandshare	9727	0.49	0.23	0	1
manper	9545	0.02	0.03	0	0.34

Note: The 'pop' variable is log transformed because its original distribution contains a limited number of very high values that skew the results. The 'relpatpp' variable indicates the amount of patents per 1,000 people. 'Schoolpct' is calculated using extrapolation, as it is not observed every tenth year. Measures for 'manper' seem to differ between different censuses, as the minumum age of a worker is not consistent.

connections, when used as the sole independent variable, have a positive and significant effect on the probability that a county will patent in E&E technology. This is in line with the expectation that accessibility improves the flow of ideas. However, navigable river linkages show a negative coefficient. It could be that counties next to rivers - which were less efficient than railroads and canals - saw their inventive activity move to areas which were better connected to more modern modes of transport. Thus, patenting in the very novel technology of electricity would be lower in those counties.

In the fourth model, all transportation variables are used. Results show that railroad and canal linkages remain positively significant, with canals having a slighty larger coefficient (0.2968) than railroads (0.2585). Here, river connections show to be no longer of significance.

In the fifth model, the control variables are added. The first variable that is of interest, is the mindist variable, indicating how close a county is to the nearest county that patented in E&E. As expected from the earlier work by Feldman et al. (2015), this variable's coefficient is negative and significant, indicating that the further away a county is from a county that patented in E&E in period t-1, the less likely it is to patent itself in period t.

Cognitive proximity is added as another main variable of interest. Again, as expected, this variable is positive and significant, indicating that the higher cognitive proximity is in period t-1, the more likely a county will patent in period t. However, contrary to Feldman et al.'s (2015) findings, its effect on the r-squared is very small, and it takes away only a small bit off the coefficients of the other variables. This result was already hypothesised in the methodology

section, and likely is the result because of the technology of interest being a broad class, co-occuring with a lot more classes than a regular technology class would.

The control variables that have to do with a county's population are added, reducing as expected the coefficients of RRconnect and Canalconnect. The effect that this addition of population variables has, indicates that just being accessibile likely doesn't yield many patents if a county has few people living in it. The same accounts for the urban percentage: since it is argued that mainly people in urban environments patent, it is no more than expected that the urban percentage variable takes some away from the coefficients of the connection variables. The percentage of a county that is foreign born might indicate the flux of new people bringing new ideas, rather than the flow of the ideas themselves, which is what this paper tries to reveal.

Also added are the control variables that resemble the existing knowledge and economic expertise in a county. All added variables show significant and positive coefficients, reducing the coefficients of the variables of interest.

In the sixth model, state fixed effects are added to account for institutional factors that might yet be unobserved in the data. Institutional factors can be things such as legislation and taxation which might influence the output of patents in the E&E class. Now, it is the Canalconnect variable who's coefficient is reduced. Apparently, the institutional factors that are included are of a bigger influence to canals than to railroad connections. Most importantly, including these state fixed effects, it can be stated for now that the variables of interest have been robust in the fact that their coefficients remain positive and significant throughout all the alterations of the model.

After this last robustness check, the effects of the variables of interest can be estimated. The presence of a railroad in time t-1 increases the probability of patenting in E&E in period t by 23.7% (0.0402/0.1694). The presence of a canal does so by 30.7% (0.0520/0.1694). This result is unexpected, as canals have been argued to be mainly of importance in earlier decades (Atack, 2010), whereas the railroad is supposed to be more important in the second half of the 19th century that is under research here. A reason for the railroad being of lesser importance could be because the system had not yet reached its point of saturation, whereas canals had theirs in 1851 (Atack, 2010).

6 Discussion

In this article, I have shown that between 1850 and 1900, physical infrastructure connections in the form of railroads and canals facilitated a sizeable share of the diffusion of new Electric & Electronic technology in US counties. This result reinforces the idea of accessibility as used by Andersson & Karlsson (2004) as a mediator of possible contact between inventors, leading to more inventiveness. By showing the relevance of the concept of connectivity, these outcomes warrant a revival of interest in the concept of accessibility and other measures of relative geographical proximity that have been neglected in the literature.

This paper furthermore confirms the existing literature on physical infrastructure and diffusion by showing that infrastructural connections spread patented knowledge. However, in contrast to earlier work, by focusing on a single coherent class of technology and by taking cognitive relatedness into account, the results of this paper make it much more likely that this diffusion is a process of knowledge spread rather than a process in which entirely unrelated knowledge is spread by the railroad.

Connecting to the GPT diffusion literature, this paper has revealed some remarkable consistency between newer and older examples of technological diffusion. Up until the addition of state fixed effects, the results are congruent with Feldman et al.'s (2015) work on diffusion of rDNA, regarding the positive role of cognitive and geographical proximity in the diffusion of technology.

However, one should be cautious by stating a causal link between the arrival of transportation links and the development of E&E patents. It could be hypothesised that a new technology like electricity would be in higher demand in areas where people lived that could actually afford the technology. In fact, there could be not a direct link between connections and the diffusion of technology, but rather an indirect one, with transport networks increasing wealth, which in turn leads to a higher demand of new technology. In that case, the connectivity would be a measure of market access after all, and not solely of accessibility.

Adding to the notions of accessibility and the proximity literature is not without policy implications. This result of this paper help building the idea that physical infrastructure is not just important for 'hard' economic factors such as logistics and market access, but also for a harder to track concept such as knowledge flows. This result is especially important, since this paper concerns the diffusion of a GPT, a technology with the potential to transform the economy, which should be of special interest for the policy maker. Especially now, with potentially huge infrastructure investments in the US on the cards, policy makers should assess the effects that this will have on knowledge diffusion.

There are quite a few additions that future research could make to this paper. Firstly, as new statistical approaches emerge, other model specifications could be tried to see whether the results remain robust. Specifically, the use of a suitable instrumental variable (IV) could be very valuable here to disentangle the effects of market access and accessibility from each other. Hereby, one could think of using mineral resources as IV, since many infrastructural projects were initially built for transporting bulk goods, rather than for connecting people. This IV has been tried in the development phase of this paper, but did not behave as expected. Rather, the addition of mineral resources as an instrumental seemed to spring more from theoretical reasoning than from statistical reality: the IV violated the assumption of being both correlated with the transportation variables but uncorrelated to the independent variable, as in fact it was correlated to neither. This resulted in a very weak instrument that only watered down the findings, instead of reinforce the outcomes. It is up to future research to find a suitable IV, as I wasn't able to find a variable that was correlated to an accessibility variable, but uncorrelated to patenting in E&E.

Secondly, with the increasing availability of data, more detailed approaches

as to how knowledge flows through time and space could be made. It would be valuable to filter out all the E&E patents that are directly related to railroad technology, because these patents signify a spread of the railroad rather than a spread of knowledge because of the railroad. Furthermore, a more refined dataset could focus more on the link between accessibility and social proximity, since infrastructural connections might increase the likelihood that people from different geographical locations might come together to collaborate. Lastly, this experimental approach, making use of the concepts of both proximity and accessibility could be reproduced or expanded into different eras to see whether the results hold, and subsequently might show some universal pattern.

Acknowledgements

The author would like to thank the IPUMS National Historical Geographic Information System (2017), Jeremy Atack's Historical Transportation SHP Files (2016) and Petralia et al.'s (2016) HistPat Dataset for providing the research data for the analysis.

Table 2: Results

	Dependent variable:					
	patentEE					
	(1)	(2)	(3)	(4)	(5)	(6)
RRconnect	0.2839***			0.2585***	0.0322***	0.0402***
	(0.0076)			(0.0076)	(0.0081)	(0.0083)
Riverconnect		−0.0338***		−0.0128	−0.0081	0.0012
		(0.0087)		(0.0081)	(0.0073)	(0.0077)
Canalconnect			0.3691***	0.2968***	0.1080***	0.0520***
			(0.0178)	(0.0173)	(0.0156)	(0.0148)
mindist					−0.0001**	−0.0001
					(0.00004)	(0.00004)
CogProx					0.1616**	0.1135*
					(0.0723)	(0.0689)
pop					0.0946***	0.0859***
					(0.0055)	(0.0062)
urbpct					0.3714***	0.4600***
					(0.0357)	(0.0374)
forbornper					0.3138***	0.1093*
					(0.0376)	(0.0565)
relpatpp					0.0698***	0.0631***
					(0.0045)	(0.0048)
schoolpct					0.1376***	0.0938***
					(0.0301)	(0.0321)
implandshare					0.0732***	0.0343*
					(0.0188)	(0.0205)
manper					0.4119***	0.0807
					(0.1589)	(0.1688)
Constant	0.2253***	0.2253***	0.2253***	0.2253***	0.2244***	0.1694***
	(0.0040)	(0.0042)	(0.0041)	(0.0039)	(0.0034)	(0.0176)
State fixed effects	No	No	No	No	No	Yes
Observations	9,788	9,788	9,788	9,788	8,971	8,971
R^2	0.1145	0.0015	0.0598	0.1529	0.4011	0.4194
Adjusted R^2	0.1144	0.0014	0.0597	0.1526	0.4003	0.4154
Residual Std. Error	0.3932 (df = 9786)	0.4175 (df = 9786)	0.4051 (df = 9786)	0.3846 (df = 9784)	0.3303 (df = 8958)	0.3261 (df = 8909)
F Statistic	1,265.2040*** (df = 1; 9786)	14.4642*** (df = 1; 9786)	622.4026*** (df = 1; 9786)	588.5587*** (df = 3; 9784)	499.9864*** (df = 12; 8958)	105.5002*** (df = 61; 8909)
Note:						*p<0.1; **p<0.05; ***p<0.01

Table 3: Correlation matrix

	patentEE	RRconnect	Riverconnect	Canalconnect	mindist	CogProx	pop	urbpct	forbornper	relpatpp	schoolpct	implandshare	manper
patentEE	1												
RRconnect	0.3331	1											
Riverconnect	-0.04	-0.0346	1										
Canalconnect	0.2443	0.1394	-0.0821	1									
mindist	-0.3126	-0.4219	0.0445	-0.1709	1								
CogProx	0.1331	0.1036	0.0014	0.0459	-0.2223	1							
pop	0.4908	0.486	0.0168	0.3135	-0.4588	0.145	1						
urbpct	0.5006	0.3199	0.0259	0.2315	-0.2631	0.139	0.5383	1					
forbornper	0.2776	0.1446	-0.0465	0.0833	-0.0961	0.0818	0.0813	0.373	1				
relpatpp	0.5137	0.3312	-0.0705	0.1523	-0.3349	0.1456	0.4339	0.5339	0.3073	1			
schoolpct	0.2529	0.2541	-0.108	0.0828	-0.3167	0.0366	0.2211	0.1715	0.2118	0.3168	1		
implandshare	0.3466	0.3905	-0.1092	0.2011	-0.3897	0.0715	0.4191	0.296	0.2576	0.3949	0.3434	1	
manper	0.3918	0.2489	-0.0184	0.1987	-0.2294	0.1374	0.4172	0.5937	0.2741	0.4696	0.1843	0.1699	1

Bibliography

Agrawal, A., Galasso, A., & Oettl, A. (2017). Roads and innovation. *Review of Economics and Statistics*, 99(3), 417-434.

Andersson, M., & Karlsson, C. (2004). 10. The role of accessibility for the performance of regional innovation systems. *Knowledge Spillovers and Knowledge Management*, 283.

Atack, J. (2016) "Historical Geographic Information Systems (GIS) database of U.S. Railroads".

Atack, J., Bateman, F., Haines, M., & Margo, R. A. (2010). Did railroads induce or follow economic growth?: urbanization and population growth in the American Midwest, 18501860. *Social Science History*, 34(2), 171-197.

Audretsch, D. B., & Feldman, M. P. (1996). R&D spillovers and the geography of innovation and production. *The American economic review*, 86(3), 630-640.

Baldwin, R. (2016). *The great convergence.* Harvard University Press.

Balland, P. A. (2012). Proximity and the evolution of collaboration networks: evidence from research and development projects within the global navigation satellite system (GNSS) industry. *Regional Studies*, 46(6), 741-756.

Balland, P. A., & Rigby, D. (2017). The geography of complex knowledge. *Economic Geography, 93(1)*, 1-23.

Bathelt, H., Malmberg, A., & Maskell, P. (2004). Clusters and knowledge: local buzz, global pipelines and the process of knowledge creation. *Progress in human geography*, 28(1), 31-56.

Bloom, D. E., Sachs, J. D., Collier, P., & Udry, C. (1998). Geography, demography, and economic growth in Africa. *Brookings papers on economic activity*, 1998(2), 207-295.

Boschma, R. (2005). Proximity and innovation: a critical assessment. *Regional studies*, 39(1), 61-74.

Boschma, R., Balland, P. A., & Kogler, D. F. (2014). Relatedness and technological change in cities: the rise and fall of technological knowledge in US metropolitan areas from 1981 to 2010. *Industrial and Corporate Change, 24(1)*, 223-250.

Bresnahan, T. F., & Trajtenberg, M. (1995). General purpose technologies Engines of growth?. *Journal of econometrics, 65(1)*, 83-108.

Coenen, L., Moodysson, J., & Asheim, B. T. (2004). Nodes, networks and proximities: on the knowledge dynamics of the Medicon Valley biotech cluster. *European Planning Studies*, 12(7), 1003-1018.

Cohen, W. M. & Levinthal, D. A. (2000) Absorptive capacity: a new perspective on learning and innovation. In: *Strategic Learning in a Knowledge economy*, 39-67.

Ellison, G., Glaeser, E. L., & Kerr, W. R. (2010). What causes industry agglomeration? Evidence from coagglomeration patterns. *American Economic Review, 100(3)*, 1195-1213.

Feldman, M. P., Kogler, D. F., & Rigby, D. L. (2015). rKnowledge: The spatial diffusion and adoption of rDNA methods. *Regional studies*, 49(5), 798-817.

Fishlow, A. (1965). *American Railroads and the Transformation of the Antebellum Economy* (Vol. 127). Cambridge, MA: Harvard University Press.

Fleming, L., & Sorenson, O. (2001). Technology as a complex adaptive system: evidence from patent data. *Research policy, 30(7)*, 1019-1039.

Fogel, R. W. (1964). *Railroads and American economic growth* (p. 38). Baltimore: Johns Hopkins Press.

Friedman, T. L. (2005). *The world is flat: A brief history of the twenty-first century*. Macmillan.

Griliches, Z. (1998). Patent statistics as economic indicators: a survey. In *R&D and productivity: the econometric evidence* (pp. 287-343). University of Chicago Press.

Grübler, A. (1996). Time for a change: on the patterns of diffusion of innovation. *Daedalus*, 125(3), 19-42.

Hägerstrand, T. (1970, December). What about people in regional science?. In *Papers of the Regional Science Association* (Vol. 24, No. 1, pp. 6-21). Springer-Verlag.

Helpman, E., and M. Trajtenberg (1998a): "Diffusion of General Purpose Technologies,". In: *General purpose technologies and economic growth*, ed. by E. Helpman, p. 85. MIT Press.

Helpman, E., and M. Trajtenberg (1998b): "A Time to Sow and a Time to Reap: Growth Based on General Purpose Technologies,". In *General purpose technologies and economic growth*, ed. by E. Helpman, p. 55. MIT Press.

Jaffe, A. B. and M. Trajtenberg (2002), "Patents, Citations, and Innovations: A Window on the Knowledge Economy". MIT Press: Cambridge, MA.

Jaffe, A. B., Trajtenberg, M., & Henderson, R. (1993). Geographic localization of knowledge spillovers as evidenced by patent citations. *The Quarterly journal of Economics*, 108(3), 577-598.

Jovanovic, B., & Rousseau, P. L. (2005). General purpose technologies. In *Handbook of economic growth* (Vol. 1, pp. 1181-1224). Elsevier.

Kogler, D. F., Rigby, D. L. & Tucker, I. (2013). Mapping knowledge space and technological relatedness in U.S. cities, *European Planning Studies* 21, 1374-1391.

Lagendijk, A., & Lorentzen, A. (2007). Proximity, knowledge and innovation in peripheral regions. On the intersection between geographical and organizational proximity. *European Planning Studies*, 15(4), 457-466.

Lipsey, R. G., Carlaw, K. I., & Bekar, C. T. (2005). *Economic transformations: general purpose technologies and long-term economic growth.* OUP Oxford.

Manson, S., Schroeder, J., Riper, D. Van, & Ruggles, S. (2017). IPUMS National Historical Geographic Information System: Version 12.0. Minneapolis: University of Minnesota. http://doi.org/10.18128/D050.V12.0

Marshall, A. (1920). *Industry and trade: a study of industrial technique and business organization; and of their influences on the conditions of various classes and nations.* Macmillan.

Mensch, G. (1975). *Das technologische Patt: Innovationen berwinden die Depression.* Umschau Verlag.

Morgan, K. (2004). The exaggerated death of geography: learning, proximity and territorial innovation systems. *Journal of economic geography*, 4(1), 3-21.

Moser, P., & Nicholas, T. (2004). Was electricity a general purpose technology? Evidence from historical patent citations. *American Economic Review, 94(2)*, 388-394.

Nesta, L. (2008). Knowledge and productivity in the worlds largest manufacturing corporations. *Journal of Economic Behavior & Organization, 67(3-4)*, 886-902.

Nooteboom, B. (2000) *Learning and Innovation in Organizations and Economies.* Oxford University Press, Oxford.

Paci, R., Marrocu, E., & Usai, S. (2014). The complementary effects of proximity dimensions on knowledge spillovers. *Spatial Economic Analysis*, 9(1), 9-30.

Paullin, C. O., & Wright, J. K. (1932). *Atlas of the historical geography of the United States.*

Pavitt, K. (1985). Patent statistics as indicators of innovative activities: possibilities and problems. *Scientometrics, 7(1-2)*, 77-99.

Perlman, E. R. (2016). Dense enough to be brilliant: patents, urbanization, and transportation in nineteenth century America. *Working Paper*, Boston University.

Petralia, S. (2017). Unravelling the Trail of a GPT: The Case of Electrical & Electronic Technologies from 1860 to 1930.

Petralia, S., Balland, P.A., Rigby, D. (2016). "HistPat Dataset", https://doi.org/10.7910/DVN/BPC15W, Harvard Dataverse, V7.

Phillips, W. H. (1992). Patent growth in the Old Dominion: the impact of railroad integration before 1880. *The Journal of Economic History*, 52(2), 389-400.

Quatraro, F. (2010). Knowledge coherence, variety and economic growth: Manufacturing evidence from Italian regions. *Research Policy, 39(10)*, 1289-1302.

Rigby, D. L. (2013). Technological relatedness and knowledge space: entry and exit of US cities from patent classes. *Regional Studies*.

Scherer, F. M. (1984), Innovation and Growth: Schumpeterian Perspectives. MIT Press: Cambridge, MA.

Shaw, A. T., & Gilly, J. P. (2000). On the analytical dimension of proximity dynamics. *Regional studies*, 34(2), 169-180.

Sonn, J. W., & Storper, M. (2008). The increasing importance of geographical proximity in knowledge production: an analysis of US patent citations, 19751997. *Environment and Planning A*, 40(5), 1020-1039.

Strumsky, D., J. Lobo and S. Van der Leeuw (2012), Using patent technology codes to study technological change, *Economics of Innovation and New Technology*, 21, 267286.

Taylor, G. R. (1951). *The Transportation Revolution, 1815-1860* (New York).

Thompson, P. (2006). Patent citations and the geography of knowledge spillovers: evidence from inventor-and examiner-added citations. *The Review of Economics and Statistics*, 88(2), 383-388.

Weibull, J. W. (1980). On the numerical measurement of accessibility. *Environment and Planning A*, 12(1), 53-67.

Weitzman, M. (1998) Recombinant growth. *Quarterly Journal of Economics* 113, 331360.