

Information and Innovation: Evidence From Railroad Expansion in the Nineteenth Century ^{*}

Shuting Li[†]

July 2026

I study how access to information shapes invention by exploiting the staggered expansion of the U.S. railroad network and the mail it carried between 1840 and 1900. Using a new dataset that links the railroad mail routes recorded by the U.S. Post Office Department to the underlying rail network, I find that counties gaining railroad access produced about eleven percent more patents over the following five years than comparable counties without it, and that the gains were larger where mail service was more frequent and more extensive. I find no improvement, on average, in patent quality, though patent quality declined in counties without prior water access and rose where waterways already existed. Using large-language-model embeddings of patent text, I further show that connected counties' patents grew more similar in content, and more so to those of ordinary counties than to those of the established innovation centers. The patenting advantage widened as the postal reforms of 1845 to 1863 lowered the cost of mailing, and patenting fell where the Post Office discontinued Southern mail service during the Civil War. Together, these findings point to the flow of information as a channel through which railroad expansion raised patenting in the counties the mail reached.

Keywords: Innovation, Information Diffusion, Railroad, Knowledge Spillover, Patent Similarity, Large Language Model

^{*}This paper is based on my doctoral dissertation at Georgia State University. I am grateful for the invaluable guidance and continuous support of my dissertation committee members: Baozhong Yang (chair), Vikas Agarwal, Omesh Kini, and Jiekun Huang. I thank Gerald D. Gay and Sophia (Jing) Xue, as well as conference and seminar participants at the 2025 FMA Doctoral Student Consortium, the 2026 AFA PhD Student Poster Session, the Georgia State University Brown Bag, California State University San Marcos, and San Diego State University, for helpful comments and suggestions. All errors are my own.

[†]College of Business Administration, California State University San Marcos. Email: sli@csusm.edu

1 Introduction

How does access to information shape innovation? Since Schumpeter, theories of economic growth have treated the diffusion of knowledge as a fundamental input into invention: each innovator builds on ideas produced by others.¹ More recent studies provide empirical evidence on the effects of knowledge flows on innovation outcomes.² Yet rigorously identifying the causal impact of information flows on innovation remains challenging. Information flows are inherently difficult to observe directly, making it necessary to identify their contribution through settings in which changes in information access can be plausibly isolated. Modern economies offer limited opportunities for such identification. When a place becomes better connected, multiple channels of information improve simultaneously, while many other economic conditions that influence innovation change at the same time. Consequently, separating the effect of information access from the broader consequences of improved connectivity is difficult.

In this paper, I address this challenge within the historical context, the expansion of the railroad network across the United States between 1840 and 1900. The historical context provides a particularly well-suited setting for tackling this challenge, mainly for three reasons. First, unlike today, when information can be exchanged through many competing means, the nineteenth century offered far fewer such channels. Long-distance correspondence and printed matter moved primarily through the mail: the Post Office held a legal monopoly on the carriage of letters, newspapers and magazines traveled at preferential rates, and the mail in turn rode the railroads under the Post Office's contracts. Steamboats and canal

¹[Schumpeter \(1950\)](#) conceptualized “creative destruction” as the process by which innovation transforms the economy by replacing obsolete industries and technologies. Based on the idea, [Aghion, Antonin, and Bunel \(2021\)](#) define the creative destruction as “the process by which new innovations continually emerge and render existing technologies obsolete, new firms continually arrive to compete with existing firms, and new jobs and activities arise and replace existing jobs and activities.” Also see [Romer \(1990\)](#), [Aghion and Howitt \(1992\)](#), [Grossman and Helpman \(1993\)](#), [Aghion, Howitt, Brant-Collett, and García-Peñalosa \(1998\)](#), [Helpman \(1998\)](#), [Weitzman \(1998\)](#), [Acemoglu \(2008\)](#), [Mokyr \(2002\)](#), [Jones and Vollrath \(2013\)](#), [Aghion, Akcigit, and Howitt \(2014\)](#), among others.

²See [Matray \(2021\)](#) and [Bai, Jin, and Zhou \(2024\)](#).

packets also carried mail extensively, but they reached only places accessible by water.³ The telegraph, in commercial use from the 1850s, transmitted only short dispatches priced by the word, unsuited to the drawings, specifications, and technical periodicals from which inventors worked.⁴

Second, the emergence of railroads significantly improved the efficiency of information exchange. Prior to the railroad era, mail and goods were transported primarily by horseback or horse-drawn vehicles, and by steamboat where water allowed. Compared with horses and stagecoaches, trains were not only faster and more cost-effective, but also capable of carrying much larger volumes and delivering mail and goods in inclement weather. Compared with steamboats, trains were faster and ran year-round, while frozen rivers halted water traffic in winter. Third, the expansion of the railroad unfolded in several stages, beginning in the Northeast and Mid-Atlantic regions, then spreading across the rest of the East, and ultimately pushing westward. This gradual expansion resulted in a staggered introduction of railroads, creating variations in the efficiency of information exchanges. As the railroads were constructed in different counties at different times, the frequency of information exchanges varied based on the density of the rail network. I therefore exploit this staggered expansion to identify the causal effect of information flows on innovative activity.

To capture variation in information exchange across counties, I construct two datasets at the county and route levels. At the county level, I combine a digital map of the U.S. railroad network with historical county boundary maps to construct a county-year panel spanning 1840 to 1900. The railroad data include all constructed segments and their opening years, allowing me to identify the year in which each county first received railroad access and to track its patenting activity before and after railroad arrival. At the route level, I assemble a

³The *Annual Reports of the Postmaster General* record steamboat mail service alongside railroad mail service. Steamboat mail frequency therefore enters the analysis as a control, and Section 5.4 compares the effect of railroad access between counties with and without prior water access.

⁴Priced by the word, a telegram cost roughly ten times a letter yet carried only ten words, so the wire was reserved for terse, high-value dispatches such as commodity prices and market reports (Du Boff, 1980, 1983). Rates fell steeply over the century, but even so it could convey little of the detailed technical material on which inventors relied (Nonnenmacher, 2001; Blondheim, 1994; John, 2010).

new dataset linking nineteenth-century postal mail routes, collected from the *Annual Report of the Postmaster General*, to the physical railroad network between 1843 and 1880. This dataset allows me to move beyond a binary measure of railroad access and quantify the intensity of mail service. Specifically, it captures the frequency of mail service, the number of mail routes serving each county, and the counties connected through those routes, thereby providing county-level proxies for the intensity and breadth of a county’s access to the circulation of mail, newspapers, and other printed information.

To study the effect of information flows on innovation, I exploit the staggered timing of county-level railroad expansion in an event-study design. Specifically, I define the construction of the first railroad in a county as the treatment event. Using this event, I apply a difference-in-differences framework that compares changes in innovation outcomes for treated counties with contemporaneous changes in control counties. A potential concern with this framework under staggered adoption is that the effect of railroad access may vary across counties or evolve over time. To address this issue, I employ a stacked difference-in-differences approach, which reorganizes the data by treatment cohorts. For each treatment year, I construct a cohort with a five-year window before and after treatment, and then “stack” these cohorts to form a panel aligned by relative treatment timing. This design helps mitigate concerns about treatment-effect heterogeneity across counties and periods under staggered adoption. Within each cohort, counties that receive their first railroad in the treatment year are designated as treated units, whereas counties without railroad access serve as potential controls.

While the stacked data provide a favorable setting for identification, a reasonable concern is that confounding factors, such as underlying economic activity, may affect both the decision to build the rails and the level of innovation output. In other words, systematic differences between treated and untreated counties could simultaneously drive railroad construction and inventive activity. To address this concern, I employ propensity score matching to construct a matched sample in which treated and control counties have a similar likelihood of receiving

a railroad. Specifically, for each treated county within a cohort, I select a counterfactual control county from the same cohort that is matched on observable characteristics, including population size, urban population share, and white population share. The key identification assumption is that, absent the arrival of the railroad, counties that first received a railroad and their matched counterparts would have followed comparable trends in innovation output.

The innovative output is measured by patents, and the patent-based measures allow me to not only observe the quantity of a county’s innovation output but also characterize the quality and nature of the output. Specifically, I follow [Kelly, Papanikolaou, Seru, and Taddy \(2021\)](#) and assess patent quality, constructed based on the text of patent documents, across three dimensions: novelty, impact, and importance.⁵ A patent is considered novel if it is distinct from prior patents, impactful if it influences subsequent patents, and important if it is both novel and impactful.

My hypothesis is that more efficient information exchange raises innovation. Consistent with this view, counties that gained railroad access produced significantly more patents over the subsequent five years than their matched counterparts, with patenting increasing by approximately 11 percent in the OLS specifications and 23 percent in the Poisson specifications. The effect is concentrated on innovation quantity rather than innovation quality: Patents produced in connected counties are, on average, no more important or impactful than those produced elsewhere and, if anything, modestly less novel. This pattern suggests that improved access to information raised the quantity of invention rather than its quality, prompting more incremental invention in connected counties rather than raising the importance of the typical patent. The gains are larger where railroad mail service was more intensive—measured by the number of weekly mail trips or the number of distinct mail routes serving a county—and where a county occupied a more central position in the railroad network, measured by its degree or betweenness centrality.

⁵A common quality measure in the literature is based on a patent’s citation. The citation-based measure, however, is not feasible in the 19th century as patent citations are consistently recorded by the USPTO in patent documents only after 1947.

If these effects operate through information diffusion along the railroad, then counties connected by the same line should exhibit greater similarity in their knowledge base, and consequently the patents generated in those counties should become more alike after the connection. To test this mechanism, I conduct a county-pair analysis in which the outcome variable is the average patent similarity between two counties. To measure patent similarity, I employ the Stella language model, an advanced large language model designed to capture nuanced semantic meaning, to convert the text of each patent into a numerical embedding. I then calculate cosine similarity for each pair of patents and aggregate patent-level similarity to county-pair level. I find that the patents of two counties become significantly more similar once they are joined by a common mail route, relative to pairs that remain unconnected. This finding helps to further pin down the underlying mechanism: The connection itself diffused knowledge across counties, aligning their inventive activity. Splitting this effect by the partner county’s innovative standing shows where the shared knowledge originated: the convergence is driven by ordinary, non-hub partners rather than by the established innovation centers. The information that traveled along the mail network thus circulated among peripheral counties rather than diffusing outward from the established centers.

I then conduct four additional analyses to probe the information channel and the form of knowledge driving the results. First, I exploit a series of mid-nineteenth-century postal reforms, including the Postal Acts of 1845, 1851, and 1863, which substantially reduced the cost of communication by lowering postage rates on letters. Combined with the preferential rates historically granted to newspapers and periodicals, these reforms reduced the cost of disseminating codified information and expanded access to knowledge across locations. The patenting advantage of rail-connected counties widens as the reforms progress, consistent with information carried by mail and print driving the results. Second, I use the U.S. Post Office Department’s termination of mail service to the Confederate states as a reverse shock: Southern counties that lost their mail routes experienced a sharp decline in patenting, reinforcing the same channel from the opposite direction. Third, to address the concern that

routes were built where invention was already likely, I restrict attention to counties that railroads merely passed through en route to demand-selected endpoints; the effect persists among these intermediate counties. Fourth, I find no significant spillover onto control counties bordering treated counties, indicating that the estimates are not contaminated by local diffusion.

My paper contributes to several strands of literature. First, it is most directly related to the growing body of work in economics and finance on how information flows shape innovation. A central theme is that greater exposure to knowledge raises innovation, and that such exposure operates through several channels: mobility and interpersonal interaction, as inventor mobility diffuses knowledge across firms and regions ([Hombert & Matray, 2017](#); [Bernstein, Diamond, Jiranaphawiboon, McQuade, & Pousada, 2022](#)) and workplace interaction improves knowledge flows ([Sandvik, Saouma, Seegert, & Stanton, 2020](#)); reductions in the cost of contact, as new direct flights and geographic proximity foster collaboration and invention ([Bernstein, Giroud, & Townsend, 2016](#); [Bai et al., 2024](#); [Bahar, Choudhury, Kim, & Koo, 2023](#); [Chikis, Kleinman, & Prato, 2025](#));⁶ corporate events that recombine ideas, such as mergers and acquisitions ([Bena & Li, 2014](#); [Seru, 2014](#)); disclosure regimes and the availability of public data ([Hegde, Herkenhoff, & Zhu, 2023](#); [Bryan & Ozcan, 2021](#)); and networks of inventors ([Kerr, 2008](#); [Zacchia, 2020](#)).⁷ Two features of this literature motivate my contribution. It is built almost entirely on contemporary settings, in which information is abundant and reaches inventors through many channels at once; and it studies episodes in which information access expands, rarely ones in which it contracts. In these settings, moreover, improved connectivity has been found to raise not only how much is invented but also the quality of what is invented ([Bernstein et al., 2016](#)).

The nineteenth-century setting lets the paper put a sharper question to the data than

⁶Relatedly, [Atkin, Chen, and Popov \(2022\)](#) show that face-to-face interactions, measured using smartphone geolocation data, improve innovation outcomes.

⁷Related work highlights the role of professional intermediaries, such as patent attorneys, in facilitating knowledge diffusion ([Lamoreaux & Sokoloff, 2001](#); [De Rassenfosse, Jensen, Julius, Palangkaraya, & Webster, 2023](#)).

contemporary work can: not whether connection raises invention, but what greater access to mailed information does to how much, and what kind of, invention a place produces.⁸ I show that counties gaining mail access invented more, but the average importance of their patents did not rise, and that null is not uniform. Patent importance declines in counties without prior water access and rises where waterways already existed, and even the counties with the most intensive service produced no more important patents. Together these patterns are consistent with wider access to mailed information raising the volume of invention in the counties it reached, while leaving the average importance of what a place patented unchanged. The setting also supplies what modern settings rarely do, a contraction of information access. When the Civil War suspended Southern mail service, patenting fell more in the counties that lost their routes than in otherwise similar Southern counties facing the same wartime conditions. This reversal, mirroring the paper’s main results from the opposite direction, reinforces the causal interpretation.

Second, my paper contributes to the literature on text-based measurement of innovation, by showing what modern language models make possible in historical settings. A growing body of work uses text as data to capture economic relationships, from product-market similarity (Hoberg & Phillips, 2016) to the content of innovation in patent text (Kelly et al., 2021; L. Dai, Monroe, Naiker, & Qin, 2024). Two limits bind in the nineteenth century: citations, the usual complement of these measures, are not recorded, and word-frequency representations such as TF-IDF understate the relatedness of patents that describe the same idea in different words, a first-order problem in a sample spanning six decades of shifting technical vocabulary. I therefore apply the quality framework developed and validated by Kelly et al. (2021) to this citation-free setting, and replace word-frequency representations with language-model embeddings in the similarity analysis (Zhang, Li, Zeng, & Wang, 2025). The same toolkit builds the treatment data: a vision-capable model reads the scanned route

⁸Relative to county-level records that document the presence of the post (Acemoglu, Moscona, & Robinson, 2016; Blevins, 2021) and to work relating transportation access to patenting in this era (Perlman, 2015), the mail-route panel described above lets me study the service the railroad carried, its frequency and reach, and, through patent text, the content of what counties invented.

tables of thirty-eight *Annual Reports of the Postmaster General* into structured form, and language-model and text-matching steps link each mail route to the physical railroad network. Together, these steps make nineteenth-century sources, patent text and administrative records alike, usable at scale.

Third, my paper contributes to the literature on the agglomeration of innovation and knowledge spillovers. A large body of research shows that inventive activity clusters geographically and that spillovers are strongest among nearby locations (Jaffe, Trajtenberg, & Henderson, 1993; Ellison, Glaeser, & Kerr, 2010; Carlino & Kerr, 2015; Hunt, Cockburn, & Bessen, 2024), and that these spillovers raise firms' innovation (Matray, 2021). Other work shows that knowledge also travels through non-geographic channels, including mergers and the firm and industry networks that link distant locations (Ahern & Harford, 2014; Giroud, Liu, & Mueller, 2026; Chikis et al., 2025). Much of this evidence infers spillovers indirectly, from the co-location of inventors or from patent citations, rather than from what is actually invented in specific places. The county-pair analysis offers a closer look: the text of two counties' patents grows more alike once the counties share a mail route, relative to matched pairs that remain unconnected, and the convergence runs toward the prior patents of ordinary partner counties rather than the established centers. That direction, visible in the content of what is invented, cuts against the concentration around leading clusters that this literature usually emphasizes, and it is there that my evidence adds. I also test directly for spillovers onto neighboring counties and find little evidence of them, which bounds local diffusion in this setting and indicates that the baseline comparisons are not contaminated by it.

2 Historical Context and Hypothesis Development

2.1 Railroad network expansion in the nineteenth century

The expansion of the railroad provides an ideal setting for studying the effect of information flow on innovation activities. The first key factor is the railroad’s central role in facilitating information exchange during the nineteenth century. At that time, railroads served as the primary infrastructure for transmitting information, a degree of dominance rarely seen in modern economies, where information flows through multiple channels. A large part of this dominance stemmed from the U.S. Post Office Department’s contracts with railroad companies. These agreements allowed mail to be transported via railroad, greatly accelerating the circulation of letters, newspapers, and magazines. Moreover, the Post Office Department held a monopoly on letter delivery and enjoyed significant cost advantages in distributing printed materials.⁹ Low or near-zero postage rates for newspapers and magazines further increased their circulation.¹⁰

The main electric alternative, the telegraph, did not substitute for the mail as a carrier of detailed information. In commercial use from the 1850s, it was priced by the word and far dearer than postage. In the early 1850s a ten-word message cost about twenty-five cents per hundred miles: fifty cents from Pittsburgh to Washington and \$1.55 from New York to Chicago, roughly ten times the cost of a letter that could itself enclose several pages (Wang, 2025; Nonnenmacher, 2001; Kennedy, 1957). Priced this way, the wire was reserved for terse, high-value dispatches such as commodity prices and market reports (Du Boff, 1980, 1983).

⁹For example, In the 1850s, sending a standard domestic letter of up to a half-ounce over a distance such as Baltimore–New York (well under 3,000 miles) cost only about 3–5 cents, following the Postal Acts of 1845 and 1851. At the same time, contemporary evidence on mid-nineteenth-century wages suggests that common laborers earned on the order of 50–100 cents per day, implying that a long-distance letter cost only a small fraction of a day’s wage, while the value of traded goods (for example, a barrel of flour) was typically several dollars, far exceeding the postage required to arrange the transaction.

¹⁰The 1792 postal law provided for the carriage of newspapers to subscribers for the relatively low rate of 1 cent for up to 100 miles and 1.5 cents for more than 100 miles. However, many subscribers failed to pay the postage until prepayment of newspaper postage became mandatory in 1875. Kennedy (1957) estimates that 40 percent of newspapers and magazines traveled essentially postage-free in 1874 because postmasters failed to collect the postage.

Rates declined steeply over the century: the New York to Chicago message dropped to forty cents by 1890, and the average message fell from \$1.09 in 1867 to thirty cents by 1900. Even then a telegram cost many times a letter and, as a channel for coded text, could carry little of the detailed technical material on which inventors relied ([Blondheim, 1994](#); [John, 2010](#)).

Water transport was the other established channel: steamboats, coastal steamers, and canal packets carried mail and printed matter cheaply and in bulk, but only to places reachable by water, whether along the seacoast or on navigable rivers, lakes, and canals. The railroad therefore added the most where water did not already reach, extending fast and inexpensive mail to inland counties, and this contrast between counties with and without prior water access is examined in [Section 5.4](#).

Second, the mail delivery efficiency improved significantly with the use of railroads and the in-transit distribution approach. Prior to the railroad age, mail was carried by post riders and horse-drawn vehicles.¹¹ By contrast, the railroad offered clear advantages in speed, cost, capacity, stability, and, most importantly, extended scope or coverage. Compared to horses and stagecoaches, trains are significantly faster, cheaper, could carry much larger volumes of mail, and able to carry mail in bad weather. For example, as the railroad's role in carrying the mail grew, delivery from New York City to Raleigh, North Carolina fell from 94 hours in 1835 to 55 hours in 1837 ([John, 1995](#)). The railroad was also more dependable across seasons, since ice closed many water routes each winter, whereas trains ran year-round. For instance, the Ohio River could be shut to navigation for weeks at a time, and the Erie Canal for roughly five months. [Fogel's \(1964\)](#) accounting of the antebellum transportation system explicitly debits water routes for the time lost when the canals and rivers froze.

Meanwhile, the mail sorting and distribution improvement sped service. Initially, mails were sorted only at the major distributing Post Offices such as Chicago or New York. Later in the 1830s route agents were appointed to the route to accompany the mail and, gradually,

¹¹Recognizing the potential of the railroads to speed up mail transportation, Post Office Department officials began to grant contracts for carrying the mail on the railroad. The first contract was awarded in 1832 on a route from Philadelphia to Lancaster, Pennsylvania, with an allowance of \$400 per year. Eventually, all railroads in the U.S. were declared to be post routes.

the agents began to sort mail in transit.¹² This sorting-in-transit approach was further standardized with the introduction of the Railroad Post Office (RPO) cars, the specially equipped cars manned by the agents who sorted and exchanged mails in transit.¹³ Moreover, along the route, passenger trains did not stop at every station. Rather, the postal agent tosses out the mail pouch for that station and picks up mail using a hook to snatch the mail pouch of outgoing mail hanging on a track-side crane.

Third, the expansion of the railroad network in the nineteenth century progressed through various stages, creating variation in information flow across counties. In 1837 twenty-nine railroads were under construction, and by 1850 most of the large eastern cities were connected by rail. Panels A and B of Figure 1 show the state of U.S. railroads between 1840-1860. By the end of the 1850s, most railroad construction was concentrated around New England, Mid-Atlantic, Southeast, and Midwest. In the next decade, the railroad network expanded to most of the areas east of the Mississippi River. Starting in the 1850s, national railroad expansion began pushing westward, leading to the establishment of overland mail routes to California.¹⁴ Panel C of Figure 1 shows the first overland railroad. When the railroad mail services were introduced into counties with the expansion of railroad networks, the counties with the services enjoy periodic information with fast and cheap delivery. The time-saving effect through railroad is magnified when looking at the transcontinental route. Prior to the completion of the transcontinental railroad in 1869, the average travel time from the Missouri River to California ranged between 16 and 20 days, and one can imagine that it should take longer to travel from the west to east coast. By contrast, the transcontinental railroad decreases the average travel time to 7 or 8 days from San Francisco to Washington, DC. I use the arrival of the railroad network in each county to investigate the impact of this

¹²Specifically, the agents opened the bags they picked up along the route and distributed the mail to the bags that were being dropped off further down the line.

¹³The majority of RPO service consisted of one or more cars at the head end of passenger trains. This new method of sorting the mail decentralized the postal service operations by sorting much of the growing volume of mail while it was being carried on the nation's rail lines. In 1869, the use of RPO cars became widespread with the inauguration of the Railway Mail Service.

¹⁴In 1862, the Pacific Railway Act authorized the Union Pacific Railroad Company and Central Pacific Railroad Company to build railroads to connect California to the Missouri River.

improvement in information exchange.

Fourth, the heterogeneity in railroad track gauge and the Post Office Department’s railroad mail routes provide me with a setting to further explore the underlying mechanism by conducting county-pair similarity analysis. During the nineteenth century, railroads in the U.S. were constructed and operated by numerous independent companies, and many adopted different track gauges ranging from 3 feet, 0 inches to 6 feet, 0 inches (Fishlow, 1965).¹⁵ The incompatibility of different gauges and minor gaps in physical networks led to a lack of interoperability, and interchange was time-consuming and labor-intensive because it had to be performed manually (Taylor & Neu, 1956; Puffert, 2009). In contrast, shipments that could remain on a continuous line without transfer avoided unloading and reloading freight, changing wheelsets, or using laborers to move cargo across cars. Therefore, one can assume that newspapers and goods were transported more easily on connected railroad lines operated by the same company and using the same gauge.

2.2 Hypothesis development

In this section, I argue that railroad construction facilitated innovation by enhancing information diffusion and accelerating the accumulation of knowledge. The expansion of railroad network leads to a periodic flow of information with elevated efficiency. The cheaper, faster, and larger volume information flow allows inventors to learn the demand of the new ideas or process and the potential value of their ideas, thereby providing powerful demand-based incentives. Moreover, the supply of knowledge serves as supply-side channel of innovation. Inventors gain technological knowledge and incorporate the knowledge to the problem they are facing, generating ideas for new products or processes. In particular, inventors gain information and knowledge via railroad mail services or through traveling. By the middle of the

¹⁵These regional and company-level choices reflected engineering traditions, cost considerations, and local geography. For example, narrower gauges were often favored in mountainous regions because they reduced construction costs on steep terrain, while broader or standard gauges were used in flatter areas to facilitate higher speeds and heavier loads. By the late 19th century, standard gauge (4 feet 8.5 inches) became dominant in the U.S., especially after the 1886 Southern gauge conversion, when thousands of miles of track were shifted from 5 feet to standard gauge almost overnight (Puffert, 2002; Gross, 2020).

century, a number of science newspapers and journals reporting inventions had emerged. The most famous examples include *Scientific American* and *American Inventors*. These journals featured articles on new inventions and the specifications of recently issued patents, and this information was disseminated through the railroad mail delivery service, which facilitated its spread along the expanding rail network.¹⁶ Additionally, inventors can gain new knowledge from the publications of patent offices. Specifically, railroads allow inventors to journey to the Patent Office, where they can check the patent list that was published on a regular basis as well as detailed specifications for those patents.¹⁷

Previous economic theories emphasize that the diffusion of knowledge is a key driver of new discoveries. In endogenous growth models (Romer, 1990; Grossman & Helpman, 1993; Aghion & Howitt, 1992), knowledge is treated as a non-rival and partially excludable input: an inventor's use of an idea does not prevent others from using the same idea, and although inventors can sometimes restrict access through patents or secrecy, they cannot fully exclude others. As a result, the dissemination of knowledge generates increasing returns to inventive effort at the aggregate level. As the stock of knowledge grows and becomes more accessible, inventors can build on a larger base of prior ideas, which raises the productivity of research, accelerates knowledge accumulation, and leads to a higher rate of innovation. Historical analyses by Mokyr (2002) support this view, arguing that the Industrial Revolution became self-sustaining only once scientific ideas and practical know-how circulated openly and reinforced each other. The interaction between theoretical knowledge and hands-on experimentation created a feedback loop in which new inventions built cumulatively on prior discoveries, generating sustained technological progress. In short, freer information diffusion expands the collective knowledge stock, enabling innovators to learn from others

¹⁶In addition to providing knowledge to inventors, the journals also provide market based incentive for inventors. For example, these journals cover news on inventors who profit from the inventions under headlines, such as "Fortunes in Patents," "Patents that Pay," and "Many Little Things Which Have Made Patentees Rich"(Nicholas, 2010). Also, the journals listed invention for sale by inventors.

¹⁷When applying for a patent, the inventor has to prove that the innovation is "first and true" by providing a detailed description to the Patent Office. The procedure is a knowledge production process, and the Patent Office serves as the storage of information.

and combine existing insights in novel ways, thereby enhancing innovative output. Therefore, I hypothesize:

Hypothesis 1 (H1a) - Quantity: *Counties with railroads generate more patents than those without.*

Nonetheless, a higher quantity of patents does not necessarily translate into higher quality. On the one hand, accumulated information and knowledge often form the foundation for breakthrough discoveries, as radical innovations typically build on a sequence of incremental advances. Rich information environments can therefore foster high-quality innovation by enabling complementary knowledge to be shared and recombined (Weitzman, 1998; Bloom, Schankerman, & Van Reenen, 2013). When ideas circulate freely, innovators can avoid duplicating existing work and instead pursue more ambitious or refined projects, while firms facing well-informed competitors must continually improve their products to stay ahead.¹⁸

On the other hand, exposure to frequent and familiar information flows may also have biased inventors toward incremental ideas, as they became more receptive to recombining existing knowledge rather than pursuing entirely novel directions (R. Dai, Donohue, Drechsler, & Jiang, 2023). Indeed, Akcigit and Kerr (2018) show that easier access to existing knowledge fosters incremental innovation but can crowd out radical innovation. Taken together, this literature suggests that while increased information flow may raise the importance and impact of patents, it may also reduce their novelty, even as it builds the knowledge base necessary for eventual breakthroughs. Therefore, I hypothesize:

Hypothesis 1 (H1b) - Quality: *Patents in counties with railroads are more likely to be important but less novel compared to those in counties without railroads.*

A county that was connected to more other counties is likely to exchange information

¹⁸The “escape competition” mechanism predicts that firms facing stronger or more informed competitive pressure have greater incentives to raise the quality of their innovations in order to “escape” their rivals. When competition intensifies, leading firms invest more in quality-improving innovation to stay ahead, while lagging firms innovate to catch up (Aghion, Bloom, Blundell, Griffith, & Howitt, 2005; Aghion & Howitt, 2008; Aghion, Akcigit, & Howitt, 2015). This dynamic is especially strong near the technological frontier, where successful innovation yields substantial rents. Mechanisms such as entry, imitation, and knowledge diffusion therefore amplify innovation incentives for frontier firms by increasing the relevance of rivals’ technological progress.

and gain knowledge more rapidly and cost-effectively. Based on that, I hypothesize:

Hypothesis 2 (H2a): *Counties with greater network connectivity tend to have a greater number of patents than those with lower connectivity.*

Hypothesis 2 (H2b): *Patents produced in more connected counties are more likely to be significant but less novel compared to those in counties with lower connectivity.*

The information channel and knowledge channel imply that inventors located in counties along the same railroad are likely to be exposed to similar information and knowledge and generate their ideas upon the received knowledge or respond to similar demands. Consequently, innovation should be more relevant to each other. Therefore, I hypothesize:

Hypothesis 3 (H3): *Patents from inventors located in counties along the same railroad tend to be similar or related to each other after the railroad’s opening.*

3 Data Collection and Variable Construction

3.1 Historical railroad and mail route data

My railroad data come from the historical GIS database constructed by [Atack \(2015, 2018\)](#). The database covers railroad lines constructed between 1830 and 1916 at the segment level. For each segment, it records the year in which the segment was completed together with its geometry, which allows me to determine whether a county contained railroad track in any given year. From these, I can determine whether and when each county gained access to the railroad network.

To trace how information flowed through the railroad network, I first identify the counties connected by each railroad line. This cannot be done using the railroad GIS data alone because the network is recorded as tens of thousands of short geometric segments rather than as named end-to-end routes. I collect railroad mail-route data from the *Annual Report of the Postmaster General*, published annually by the U.S. Post Office Department starting from 1843. To transport mail, the Department contracted with railroad companies to carry mail

along their lines, with each company responsible for one or more routes operating at specified delivery frequencies. For each active route, the reports record its origin and destination, route length, contracting railroad companies, delivery frequency, and compensation paid.¹⁹ Figure B.1 presents the first page of the 1850 annual report as an illustration of the report format. I prepare the mail-route data for matching by extracting the route tables from the digitized reports with a vision-capable large language model Google Gemini Pro 3 and geocoding each route’s origin and destination.

Next, I match postal mail routes to the railroad network to construct route-level data. This task is challenging because the two sources share no common identifier and are organized at fundamentally different levels of aggregation. I provide a detailed description of the matching procedure in Appendix A.1. Through a series of matching and validation steps, I successfully link the two datasets for the period 1843 to 1880. Consequently, the route-level analysis is restricted to these years. By matching each mail route to the underlying railroad geometry, I recover the counties traversed by the route and construct measures of information flows along the railroad network.

From the matched route-level dataset, I capture three variables at two levels of analysis: the county and the county pair. At the county level, I use two measures: *Trips*, the total number of weekly mail trips serving a county, and *Routes*, the number of distinct mail routes that pass through it. *Trips* reflects how frequently mail moved through the county, and *Routes* the number of separate channels through which it arrived; together they capture how intensively a county was integrated into the rail mail network, beyond the binary fact of rail access. At the county-pair level, I define an indicator variable $Connected_{f,p}$ that equals one if county f and county p are directly connected by railroad mail route.

¹⁹Aneja and Xu (2024) use the same reports to construct city-level measures of postal service quality and communication intensity, highlighting their value as a systematic source on nineteenth-century information infrastructure.

3.2 Historical patent data

3.2.1 Measuring innovation outcome

My main data source of patent data is PatentCity created by [Bergeaud and Verluise \(2024\)](#). The database covers the information of all utility patents during my sample period, including the name of inventors and their locations. I therefore restrict my sample to utility patents. This data set allows me to calculate my first key outcome variable patent volume, the number of utility patents published each year at county level.

My second outcome variable is patent quality. A common measure of patent quality is based on citations. This measure is not feasible in my historical setting, as citation practices were not standardized in the nineteenth century and citation data are largely unavailable.²⁰ Instead, I adopt the text-based quality measures developed by [Kelly et al. \(2021\)](#), which capture a patent’s importance, impact, and novelty. These measures provide a consistent way to evaluate patent quality across my sample period.

I obtain patent quality dataset from [Kelly et al. \(2021\)](#). The dataset covers patents over my sample period at patent level. Specifically, a patent’s novelty is defined as its dissimilarity with the existing patents at the time it was filed, and a patent’s impact is defined as its similarity with the patents filed following the patent. The idea is that a patent deviating from the existing patents is more likely to be novel, and a patent greatly influencing the subsequent patents is more likely to be impactful.

To calculate a patent’s similarities with its prior patents and following patents, [Kelly et al. \(2021\)](#) convert each patent description into a numerical representation using a natural language processing methodology called Term Frequency-Inverse Dense Frequency (TF-IDF). TF-IDF measures how important a word is in a patent description compared to the description of a whole universe of the patents, and it highlights words that are frequent in a given patent description but rare across the whole universe of the patents.²¹ The resulting

²⁰The USPTO began publishing citation data only in 1872.

²¹TF measures how often a word appears in a document, emphasizing words that are important within that specific text. IDF reduces the weight of common words by measuring how rare a word is across all fields

numerical representation, a vector, is used to calculate the cosine similarity between each pair of patents. Then they aggregate the pair-wise patent similarities into individual patent level by summing up the similarities over the patent pairs. Specifically, patent i 's similarity with the existing patents, *Backward Similarity*, and similarity with the following patents, *Forward Similarity*, are defined as

$$BwdSim_{i,T} = \sum_{j \in B_{i,T}} \rho_{i,j}, \quad FwdSim_{i,T} = \sum_{j \in F_{i,T}} \rho_{i,j}, \quad (1)$$

where $B_{i,T}$ is the set of patents filed in the T years prior to i 's filing and $F_{i,T}$ is the set of patents filed over the next T years following i 's filing. $\rho_{i,j}$ is the similarity between patent i and patent j where patent j is in the set of $B_{i,T}$ or $F_{i,T}$.

Based on this framework, the importance of a patent is measured as the ratio of its forward similarity to its backward similarity. I adopt two measures of importance: $Importance_{i,[-5,5]}$, the ratio of five-year forward similarity to five-year backward similarity, and $Importance_{i,[-5,10]}$, the ratio of ten-year forward similarity to five-year backward similarity:

$$Importance_{i,[-5,5]} = \frac{FwdSim_{i,5}}{BwdSim_{i,5}}, \quad Importance_{i,[-5,10]} = \frac{FwdSim_{i,10}}{BwdSim_{i,5}} \quad (2)$$

Intuitively, patents that differ substantially from prior work (low backward similarity) yet strongly shape future inventions (high forward similarity) are classified as more important. More detailed information about the procedure for calculating the patent quality measures can be found in the Appendix.

3.2.2 Measuring county-pair similarity

To perform the county-pair analysis, I construct text-based similarity measures at the patent-pair level in order to capture county-pair similarity.²² The main data sources are the

in the entire universe of patents.

²²Kelly et al. (2021) do not provide similarity measures at the patent-pair level.

USPTO’s historical patent files and Google’s Patent Search engine. From the USPTO, I collect each patent’s number and grant date. Because USPTO records for patents granted prior to 1976 are stored in non-standardized text formats, I rely on Google’s Patent Search engine to retrieve the corresponding patent descriptions using the USPTO patent numbers.

To calculate county-pair similarity, I begin at the patent level by computing patent-pair similarity. Following the approach of [Kelly et al. \(2021\)](#), I measure the cosine similarity between patent descriptions. Each description is first converted into a 1024-dimensional embedding vector using the Stella language model developed by [Zhang et al. \(2025\)](#). Stella is a large-scale embedding framework trained on diverse text corpora and designed to capture nuanced semantic meaning. Compared to frequency-based representations such as TF-IDF, which treat documents as collections of word counts and ignore context, Stella embeddings capture both the semantic meaning and the contextual relationships between words. This allows them to identify conceptual similarity between patents even when different vocabulary is used to describe similar ideas.

For example, consider two historically important steam patents. George Henry Corliss’s 1849 patent for the Corliss valve gear introduced rotary valves and variable cutoff mechanisms that dramatically improved steam engine efficiency.²³ Three decades later, John Tregoning’s 1879 patent for an improved steam pump valve advanced valve design to enhance the reliability and performance of pumping engines. While both patents concern innovations in steam valve mechanisms, their descriptions employ different technical terminology. A TF-IDF representation would register little overlap between terms such as “rotary valve” or “variable cutoff” in Corliss’s specification and “valve chest” or “grooved piston rod” in Tregoning’s. In contrast, the Stella embedding model captures the underlying semantic similarity: both patents represent advances in steam valve technology aimed at improving engine efficiency.

Using the embedding vectors generated by the Stella model, I compute the cosine simi-

²³George Henry Corliss received his most famous U.S. patent (Patent No. 6162) on March 10, 1849, covering the Corliss steam engine valve gear. He later obtained additional patents improving the mechanism — including the one granted in 1859 (Patent No. 24618)

ilarity between patent i and patent j , denoted as $Similarity_{i,j}$. I then aggregate patent-level similarities into county-level measures. Specifically, the similarity between a focal county f and partner county p in year t is defined as the average similarities between patents published in year t in county f and patents granted in county p during the preceding window $[t - 10, t - 3]$:

$$Similarity_{f,p,t} = \frac{1}{N} \sum_{[t-10, t-3]} Similarity_{i_{f,t}, j_{p, [t-10, t-3]}}, \quad (3)$$

where $Similarity_{f,p,t}$ is the average similarity between patent i from county f in year t and patent j from county p in years $[t - 10, t - 3]$, and N is the total number of such patent pairs.

I adopt this restricted time window for two reasons. First, excluding contemporaneous patents $[t - 2, t]$ reduces the risk of capturing simultaneous discoveries that do not reflect true knowledge diffusion across counties. Second, using a 3-to-10-year backward window balances recency and lag: patents from the recent past are most likely to influence current innovations, but diffusion and adoption often took several years, especially in the nineteenth century when communication and transportation were slower. The $[t - 10, t - 3]$ window thus focuses on relevant prior knowledge while avoiding simultaneity bias.

3.3 County boundary and county-level control variable

I collect U.S. historical county boundary data from the IPUMS National Historical Geographic Information System (NHGIS), assembled by [Manson et al. \(2024\)](#), for the years 1850, 1860, 1870, 1880, 1890, and 1900. Following [Hornbeck \(2010\)](#) and [Hornbeck and Rotemberg \(2024\)](#), I standardize county-level variables across decades by using 1890 county boundaries as the consistent geographic unit. This adjustment accounts for frequent changes in county borders during the nineteenth century and ensures comparability of data over time. When a county from another decade overlaps with more than one 1890 county, I allocate its data proportionally by multiplying the county’s value by the share of its area contained within

each 1890 county.²⁴ For each decade, the values assigned to an 1890 county are obtained by summing all overlapping pieces.²⁵

I collect county-level population data from the U.S. Censuses of Agriculture and Population. The three main variables of interests are total population, total urban population, and total white population. Based on that, I calculate the urban share, defined as the total urban population over total population, and the white share, defined as the total white population over total population. Table 1 shows that between 1840 and 1900 the average population within a county-year is 9,976, the average urban share is 2%, and the average white share is 85%. I further collect three measures of local economic conditions that I use as additional controls in the robustness analysis: manufacturing capital, the total capital invested in a county’s manufacturing establishments from Census of Manufactures; farm value, the cash value of farms in a county from Census of Agriculture; and steamboat mail frequency, the number of steamboat mail trips per week serving a county from *Annual Report of the Postmaster General*.

4 Empirical Specification

4.1 Stacked difference-in-difference and propensity score matching

To identify the effect of railroads on innovation, I implement an event study design around the staggered timing of county-level railroad construction. Specifically, I define the construction of the first railroad in a county as the treatment event. Using this event, I apply a difference-in-differences (DID) framework that compares changes in innovation outcomes for treated counties with contemporaneous changes in control counties. A concern with this framework under staggered adoption is that the effect of railroad access may evolve over time or differ

²⁴Hornbeck and Rotemberg (2024) provide the dataset reporting these area shares.

²⁵This procedure assumes that data are evenly distributed across county area, though in practice most counties in each decade overlap with only a single 1890 county, so adjustments are typically minor.

across counties. In such cases, standard two-way fixed effects DID estimators can be biased.²⁶

To address this concern, I employ a "stacked" difference-in-differences approach. The stacked approach re-organizes the data by treatment cohorts. For each treatment year, a cohort is formed with a fixed time window before and after treatment, and these cohorts are then "stacked" to create a panel aligned by relative treatment timing. This method isolates the treatment effect by comparing changes in outcomes for treated counties with changes in outcomes for ex ante similar control counties. As shown in [Gardner \(2022\)](#), the stacked approach is equivalent to estimating an average treatment effect within each cohort and then taking a weighted average of the cohort-specific estimates, where the weights reflect cohort sizes. In this sense, the stacked design parallels the approaches of [Sun and Abraham \(2021\)](#) and [Callaway and Sant'Anna \(2021\)](#), which also estimate treatment effects separately for different groups and then aggregate these estimates into an overall treatment effect. Therefore, this approach is robust to heterogeneity across counties and periods when treatment event is staggered ([Gardner, 2022](#)). Recent applied studies have begun to adopt this approach, such as [Gormley and Matsa \(2011\)](#), [Gormley and Matsa \(2016\)](#), [Cengiz, Dube, Lindner, and Zipperer \(2019\)](#), [Deshpande and Li \(2019\)](#), [Gormley, Jha, and Wang \(2026\)](#).

To create a stacked sample for the test, I first identify 24 treatment years in which any of the counties has the first railroad built during the sample period. For each treatment year listed in [Figure 2](#), I analyze county-year level observations in an 11-year window around the year of treatment (five years before and five years after). Counties with the first railroad built in the treatment year are flagged as treated counties, and counties with no railroad built during the 11-year window are the potential control counties. A potential control is therefore either never connected to a railroad during the sample period (never-treated) or connected only after the window closes (not-yet-treated); counties already served by a railroad at the start of the window are excluded from the comparison group, so that treated counties are

²⁶Recent econometric research demonstrates that with staggered treatment timing and heterogeneous treatment effects, conventional DID estimates may even take the opposite sign of the true effect ([A. C. Baker, Larcker, & Wang, 2022](#); [A. Baker, Callaway, Cunningham, Goodman-Bacon, & Sant'Anna, 2026](#))

never compared with already-treated ones.

Figure 2 plots the share of counties that are treated, not-yet-treated, never-treated, or already connected at the start of the event window for each cohort. The figure highlights two features central to my identification strategy. First, railroad access expanded steadily over time. The share of counties already connected at the start of a cohort’s event window rose from a small minority in the 1840s to nearly the entire sample by the 1890s, with the largest waves of new connections occurring around 1872, 1881, and 1889. Second, because already-connected counties are excluded from the comparison group rather than used as controls, the design avoids the already-treated comparisons that can bias conventional two-way fixed-effects estimators under staggered treatment adoption (Goodman-Bacon, 2021; Sun & Abraham, 2021; A. C. Baker et al., 2022).

Because federal census coverage expanded westward over time, many trans-Mississippi counties are not observed until after 1850.²⁷ I therefore require a county to be observed in all eleven years of a cohort’s event window in order to enter that cohort. This balanced-window requirement ensures that every event-time comparison is made within the same set of counties while still allowing western counties to enter the sample. Specifically, because the requirement is imposed within each cohort’s event window rather than across the entire panel, the estimation sample is not limited to long-settled eastern counties and can incorporate western counties as they appear in the census. As a result, identification in later cohorts is based on a national, rather than predominantly regional, cross-section of counties. Figure 2 illustrates this expansion of census coverage in the trans-Mississippi West: The number of

²⁷These cases reflect the historical expansion of federal census coverage. In most instances, a county’s absence from the data coincides with periods during which it lay outside the censused territory of the United States. Counties that enter the panel only partway through the sample period are concentrated in the trans-Mississippi West, including 468 counties in Texas, California, Oregon, the Mountain West, and the Pacific Coast. These areas were either foreign territory or remained unorganized in 1830 and entered federal census coverage only after the territorial acquisitions of the 1840s, including the annexation of Texas in 1845, the Oregon Treaty of 1846, and the Mexican Cession of 1848. Most of these counties first appear in the 1850 census. A second group consists of seven counties in northeastern Texas along the Red River, including Bowie, Cass, Red River, and Titus Counties. These counties were enumerated in 1830 as part of Miller County in the Arkansas Territory, but subsequently became part of the independent Republic of Texas between 1836 and 1845 and therefore fell outside the scope of the 1840 U.S. Census. Following annexation, they re-entered federal census coverage and appear again in the 1850 census.

observable counties increases from approximately 2,300 before 1850 to nearly 2,800 thereafter. Next I pool these cohorts into a single dataset and re-center so that event time is measured relative to each cohort’s treatment year.²⁸

Another major concern about my analysis is that railroad construction may not be exogenous: counties that received railroads likely differed from those that did not in ways that affect both where railroads were built and local innovation activity. To mitigate this concern, I employ propensity score matching to construct a sample in which treated and control counties have a similar estimated probability of railroad construction. I select control counties separately for each cohort. Eligible controls are counties that did not receive a railroad at any point during the cohort’s eleven-year event window. Within each cohort, I match every treated county to the control county with the most similar propensity score.

The matching is based on three covariates measured in the year before treatment: log population, the urban population ratio, and the white population ratio. I use these covariates because they plausibly influence both selection into treatment and the outcome: County population captures market size, which influenced both railroad investment and the returns to invention.²⁹ Urbanization proxies for agglomeration economies, which attracted rail connections and fostered innovative activity.³⁰ The white population ratio captures differences in human capital, access to the patent system, and the broader economic conditions that shaped regional development and infrastructure investment. During the nineteenth century, enslaved individuals and much of the Black population faced substantial barriers to participating in formal invention and patenting activities.³¹ Within each cohort, I estimate a logistic regression of treatment status on these three covariates and use the predicted values as propensity scores. I then implement one-to-one nearest-neighbor matching without

²⁸Because counties that serve as controls in one cohort may receive a railroad and become treated in a later cohort, the same county can contribute observations as both a control and a treated unit across different cohorts. The stacked specification accommodates this structure through cohort-specific fixed effects, while standard errors are clustered at the county level.

²⁹See [Fogel \(1964\)](#), [Atack, Bateman, Haines, and Margo \(2010\)](#), [Donaldson and Hornbeck \(2016\)](#), and [Akçigit, Grigsby, and Nicholas \(2017\)](#).

³⁰See [Higgs \(1971\)](#).

³¹See [Khan \(2005\)](#) and [Sokoloff and Engerman \(2000\)](#).

replacement, pairing each treated county with the control county that has the most similar propensity score.³²

Panel A of Table 1 reports covariate balance before and after matching. Before matching, treated counties differ markedly from the pool of potential controls, consistent with railroads reaching larger, more developed counties. Propensity score matching pairs 2,122 of the 2,394 treated counties with controls and brings the three matching covariates into close balance: all statistically indistinguishable across groups. Panel B of Table 1 reports summary statistics for the matched, stacked estimation sample, deduplicated to unique county-years so that a control county appearing in several cohorts is counted only once; the sample comprises 2,540 counties and 41,826 county-years. County patenting is sparse and right-skewed: the average county-year produces 0.60 patents, yet the median is zero and 61 percent of county-years record no patent at all. The median county had a population of approximately 8,200 residents. The urban share was zero at the median and only 0.02 on average, indicating that the typical county in the sample was rural. The average white population share was 0.85, suggesting that county-year observations were predominantly white during the study period.

4.2 Regression specification

My difference-in-differences specification compares changes in innovation outcomes for treated counties with contemporaneous changes in control counties following the first railroad completion. Because the analysis employs a stacked event-study design, counties are grouped into treatment cohorts based on the year in which they first receive a railroad. The baseline specification is:

$$Y_{c,t,g} = \beta \textit{Treat}_{c,g} \times \textit{Post}_{c,t,g} + \alpha_1 \textit{Treat}_{c,g} + \alpha_2 \textit{Post}_{c,t,g} + \textit{Controls}_{c,t} + \gamma_c + \gamma_t + \gamma_g + \epsilon_{c,t,g}, \quad (4)$$

³²Each treated county is paired with the untreated county whose propensity score is closest, provided that county lies within the caliper of 0.2 standard deviations; treated counties with no control inside the caliper are left unmatched. Because matching is carried out separately by cohort, each control county is used at most once within a cohort but may serve as a control in more than one cohort.

where $Y_{c,t,g}$ is the outcome of interest for county c in year t within cohort g , including patent volume and patent quality. $Treat_{c,g}$ is an indicator equal to one if county c receives its first railroad in cohort g . $Post_{c,t,g}$ is an indicator that equals one for years on or after the railroad is completed in county c within cohort g . I cluster standard errors at the county level to account for any heteroskedasticity and possible county-level correlations among observations.

The county-level control vector $Controls_{c,t}$ includes population, urban share, and white share. These control variables help absorb the time-varying socioeconomic conditions that were potentially associated with the innovation activity. Since the county-level control variables are drawn from the decennial census data, I also include county fixed effect γ_c , year fixed effect γ_t , and cohort fixed effects γ_g to mitigate potential omitted variable bias. County level fixed effect γ_c controls for time-invariant omitted county characteristics. Year fixed effect γ_t accounts for transitory economy-wide factors, such as macroeconomic conditions, and any time trend in innovation activity. Cohort fixed effects γ_g absorb systematic differences across treatment cohorts. In additional specifications, I further include county-by-cohort fixed effects and year-by-cohort fixed effects to account for unobserved heterogeneity within each stacked cohort.³³ These more saturated fixed-effect structures ensure that identification comes solely from within-cohort variation between treated and control counties over time, thereby providing a more stringent test of the causal effect of railroad access on innovation.

The key coefficient of interest is β , which captures the effect on innovation output of the improvement in information access that followed the completion of the first railroad. If the improved information exchange efficiency leads to more innovation, the change of innovation output in treated counties should be more than the change of those in the counterfactual control counties. Therefore, I expect to see a positive and significant β .

Given the fact that some counties do not generate any patents in some years, the county-based patent quantity used in the regression exhibits right-skewed distributions with many

³³County-by-cohort fixed effects absorb all time-invariant characteristics of a county within each treatment cohort, while year-by-cohort fixed effects control for cohort-specific time shocks that may affect innovation outcomes.

zeros. I adopt two approaches to handle the right-skewed nature. First, I use the natural logarithm of one plus the number of patents.³⁴ Second, I estimate Poisson regressions on the number of patents, as Poisson regression tends to generate consistent and reasonably efficient estimates (Cohn, Liu, & Wardlaw, 2022).

The key assumption underlying the baseline empirical approach is parallel trend. This assumption requires that, in the absence of railroads, the difference in innovation output between the treated and control groups would have been constant over time. To provide support for this assumption and show the trends for the treated counties change after the treatment year, I estimate:

$$Y_{c,t,g} = \sum_{k \neq -1} \beta_k Treat_{c,g} \times D_{k,g} + \alpha_1 Treat_{c,g} + \sum_{k \neq -1} \alpha_k D_{k,g} + Controls_{c,t} + \gamma_c + \gamma_t + \gamma_g + \epsilon_{c,t,g}, \quad (5)$$

This specification is equivalent to the baseline DID model, except that I replace the post-treatment indicator with a set of event-time indicators that capture the dynamics of the treatment effect. Specifically, $D_{k,g}$ equals one if year t is k periods away from the treatment year for cohort g , where $k = -1$ (the year immediately preceding treatment) is omitted as the reference period. The coefficients β_k therefore trace the evolution of innovation outcomes before and after the arrival of the railroad. To assess the parallel-trends assumption, I examine whether the coefficients on the interactions between the treatment indicator $Treat_{c,g}$ and the pre-treatment event-time indicators are statistically indistinguishable from zero.

I also adopt three continuous measures of treatment intensity. Let $I_{c,t}$ denote an intensity measure for county c in year t : $Trips_{c,t}$, the total number of weekly railroad mail trips passing through county c across all mail routes (the intensive margin, capturing the frequency of service); $Routes_{c,t}$, the number of distinct railroad mail routes serving county c (the extensive margin, capturing the breadth of connection); or the county's centrality in the railroad

³⁴Studies that adopt this log-transformation practice include Hirshleifer, Low, and Teoh (2012) and Bernstein, McQuade, and Townsend (2021), among others.

network, described below. For each measure I estimate

$$Y_{c,t,g} = \beta \left(Treat_{c,g} \times Post_{t,g} \times I_{c,t} \right) + \delta \left(Treat_{c,g} \times Post_{t,g} \right) + \alpha_1 Treat_{c,g} + \alpha_2 Post_{t,g} + Controls_{c,t} + \gamma_c + \gamma_t + \gamma_g + \epsilon_{c,t,g}. \quad (6)$$

where $Treat_{c,g}$ equals one if county c received its first railroad in cohort g , $Post_{t,g}$ equals one in the years on or after the treatment year, and γ_c , γ_t , and γ_g are county, year, and cohort fixed effects (the saturated specification additionally absorbs cohort-by-year and cohort-by-county fixed effects). The key coefficient of interest is β on the triple interaction: if the innovation gains from railroad access are larger in counties with more intensive rail mail service, β should be positive and significant. I estimate Model (6) separately for $Trips_{c,t}$ and $Routes_{c,t}$.

Beyond the intensity of mail service, I relate the patent response to a county's *position* in the railroad network. From the physical rail network I compute each county's degree centrality, $Degree_{c,t}$, the number of counties to which it is directly connected, and its betweenness centrality, $Betweenness_{c,t}$, which measures how often the county lies on the shortest paths linking other counties. I enter each as $I_{c,t}$ in Model (6) after converting it to a within-year percentile rank among connected counties. To ensure that a county's network position does not itself respond to its post-connection patenting, I freeze this rank at its value in the connection year and hold it fixed over the event window. Unlike the mail-route measures, which are observed only for 1843–1880, railroad-network centrality is defined throughout the full 1840–1898 sample. For these specifications I also report the effect for the most central connected county, obtained as $\beta + \delta$ evaluated at the top of the rank distribution.

5 Empirical Results

This section reports the findings of my analysis. I first explore how the completion of the first railroad in each county altered innovation output, including patent quantity and quality.

I then provide evidence on treatment intensity, examining how the intensity of railroad mail service, measured by the frequency of mail trips and the number of mail routes, affects innovation output.

5.1 The effect of information flow on patent quantity

I estimate the effect of the arrival of the first railroad on county-level patenting between 1840 and 1900 using the baseline specification in Model 4. The results are reported in Table 2. Across all four columns, the coefficient on the interaction term $Treat \times Post$ is positive and statistically significant at the 1 percent level, indicating that counties receiving their first railroad experienced a subsequent increase in patenting relative to matched control counties. Columns (1) and (3) report OLS estimates using $\ln(1 + Patents)$ as the dependent variable. The estimated coefficients imply an increase in patenting of approximately 11 percent following railroad arrival. Columns (2) and (4) report Poisson estimates, which suggest a larger increase of roughly 23 percent and similarly indicate that treated counties generated more patents after gaining railroad access. The results are economically meaningful, stable across specifications, and remain statistically significant after controlling for population, urbanization, and racial composition, as well as after incorporating cohort-by-year and cohort-by-county fixed effects.

In terms of economic magnitude, the estimates in Table 2 indicate that the arrival of the first railroad increased county-level patenting by approximately 11 percent in the OLS specifications and 23 percent in the Poisson specifications. Given that the average county-year in the estimation sample produced about 0.6 patents, these estimates correspond to roughly 0.07 to 0.14 additional patents per county per year following railroad completion. Patenting activity in this period was both low and highly skewed: 61 percent of county-years recorded no patents, the median county-year produced zero patents, and the most innovative county-year generated 35 patents. Against this backdrop, the estimated effects are economically meaningful, suggesting that railroad access substantially increased innovative

activity in the typical county, despite the relatively small absolute increase in a setting where patenting was rare.

The coefficients on the controls are consistent with the economic geography of patenting in this period. Population is positive and significant, so larger counties are associated with more patents. The coefficient on white share, by contrast, is negative and significant. This may seem to indicate a negative relation between white population and innovation, but white share moves little within a county and the matched counties are demographically similar, so the estimate rests on limited within-county variation and means only that, holding population and urbanization fixed, county-years with a higher white share record somewhat fewer patents.³⁵

The results above are based on the matched sample, which helps rule out some confounding factors that could bias the estimates. To further strengthen identification, I next examine the dynamics of the treatment effect using Model 5, which allows me to test whether treated counties exhibited differential pre-treatment trends in innovation. Figure 3 presents the results. The lead coefficients are small and statistically insignificant prior to treatment, providing support for the parallel-trends assumption. Following railroad arrival, the coefficients become positive and increase steadily, rising from approximately 0.03 in the treatment year to about 0.14 five years later. These estimates indicate a sustained and statistically significant increase in patenting after counties gain railroad access.

A natural concern is that railroads were not built at random. Railroad lines were often extended toward counties that were already industrializing, agriculturally prosperous, or commercially active, and these characteristics may themselves promote patenting, thereby confounding the estimated effect of railroad access. To address this concern, I augment the baseline specification (Model 4) with three county-level proxies for economic activity that are

³⁵I read it as a descriptive demographic association rather than a causal effect of racial composition on innovation for two reasons. First, white share enters as a level rather than in interaction with treatment. Second, the matched sample pairs treated and control counties with similar population, urbanization, and racial composition, so the comparison runs among demographically comparable counties, and the county fixed effects remove the time-invariant differences in white share that distinguish regions. The coefficient is identified instead from variation within a county over time.

plausibly correlated with both railroad placement and innovation: manufacturing capital, farm value, and steamboat mail frequency, which captures preexisting access to waterborne transportation.

Because these variables become available at different points in time, their inclusion requires restricting the sample to cohorts for which the controls are observed throughout the event window. Columns (1) and (2) of Table A.2 add manufacturing capital and restrict the sample to cohorts with treatment years beginning in 1845. Columns (3) and (4) further add farm value and steamboat frequency and restrict the sample to cohorts beginning in 1855. Across all four specifications, the coefficient on $Treat \times Post$ remains positive and statistically significant at the 1 percent level and changes little relative to the baseline estimates. The estimated effect is approximately 0.11 in the OLS specifications and ranges from 0.20 to 0.23 in the Poisson specifications, corresponding to roughly an 11 percent increase and a 22 to 25 percent increase in patenting, respectively. The additional controls enter with sensible signs: manufacturing capital and steamboat frequency are positively associated with patenting, whereas farm value is not. Importantly, including these controls does not attenuate the estimated railroad effect. These findings suggest that the positive relationship between railroad access and innovation is not driven by industrialization, agricultural wealth, or prior access to water transportation, but instead reflects the railroad’s own contribution to local innovative activity.

Taken together, the results support the interpretation that the expansion of the railroad network fostered innovation by facilitating greater patenting in newly connected counties.

5.2 The effect of information flow on patent quality

I next examine the effect of railroad access on patent quality at the county level using the same stacked difference-in-differences framework. I consider three dimensions of patent quality constructed from the textual similarity measures of Kelly et al. (2021). *Importance* is measured as the ratio of forward to backward similarity. *Novelty* is measured as the

inverse of backward similarity, so that patents that are less similar to prior art receive higher scores. *Impact* is measured by forward similarity, so that patents that are more closely reflected in subsequent inventions receive higher scores. Because these similarity measures are first available in 1840, I restrict the sample to cohorts with treatment years from 1845 onward and include the same controls used in the patent-quantity analysis, together with manufacturing capital. For each measure, I report both OLS estimates using the logarithm of the quality measure and Poisson estimates using the level of the quality measure. If improved access to external knowledge enhanced patent quality, I would expect a positive coefficient on $Treat \times Post$ across these dimensions.

Table 3 reports the results from the saturated specification that includes county, year, cohort, cohort-by-year, and cohort-by-county fixed effects. Columns (1) through (4) examine patent importance. Columns (1) and (2) use the ratio of five-year forward similarity to five-year backward similarity, while Columns (3) and (4) use the ratio of ten-year forward similarity to five-year backward similarity. Across all four specifications, the coefficient on $Treat \times Post$ is small, negative, and statistically insignificant. The estimated effects range from -0.009 to -0.012 , with absolute t -statistics below 1.6. These results provide no evidence that railroad access increased the importance of patents produced in treated counties.

Columns (5) and (6) examine patent novelty, measured as the inverse of five-year backward similarity. The estimated coefficient on $Treat \times Post$ is negative in both specifications, equal to -0.033 in the OLS model and -0.043 in the Poisson model, but neither estimate is statistically significant. While the negative point estimates suggest that patents may have become somewhat more similar to existing knowledge following railroad access, the estimates are too imprecise to support a meaningful change in novelty.

Columns (7) and (8) examine patent impact, measured by five-year forward similarity. The coefficient on $Treat \times Post$ is positive in both specifications, equal to 0.022 in the OLS model and 0.005 in the Poisson model. Although both estimates are statistically insignificant, the positive signs are consistent with the hypothesis that improved information

access increased the influence of treated counties' patents on subsequent inventions. Among the three quality dimensions, impact is the only one whose estimated effect moves in the predicted direction.

A statistically insignificant result raises the question of whether the absence of a quality effect reflects omitted local economic conditions. Railroads were often extended toward economically dynamic counties, where manufacturing activity, agricultural development, and water transportation could independently influence patent quality. To address this concern, I augment the quality regressions by adding two additional controls, farm value and steamboat frequency, while retaining manufacturing capital as a control variable. Because farm value is first available in 1850, the analysis is restricted to cohorts with treatment years from 1855 onward, while continuing to control for manufacturing capital. Table A.5 reports estimates from the saturated specification. In this later and more tightly controlled sample, the quality estimates become more negative. The coefficient on $Treat \times Post$ is negative and statistically significant for both the five-year and ten-year importance measures under both estimation methods, ranging from -0.012 to -0.019 and significant at the 5 or 10 percent level. Novelty is also negative and statistically significant in the Poisson specification (-0.182 , $t = -2.19$), while impact remains positive but statistically insignificant (0.038 in OLS and 0.008 in Poisson). Overall, these results reinforce the conclusion that railroad access did not improve patent quality. The post-1855 sample suggests a modest decline in both the importance and novelty of patents produced in counties connected to the railroad network.

The county-level quality measures average patent characteristics within each county-year and do not account for differences in the technological composition of patenting activity. Because textual similarity measures vary systematically across technology fields, a railroad-induced shift in the distribution of patent classes could affect county-level averages even if patent quality within each field remained unchanged. To address these concerns, I re-estimate the quality regressions at the patent level using cohorts with treatment years from 1845 onward. Each patent is assigned to a county and grant year based on the location of

its first-listed inventor and inherits the treatment status of its county-year. The dependent variable is the patent’s own log similarity measure rather than a county-year average. I further include CPC technology-class fixed effects so that the coefficient on $Treat \times Post$ is identified from within-field variation in patent quality. Table A.6 reports the results. The odd-numbered columns include county, year, cohort, and technology-class fixed effects, while the even-numbered columns additionally include cohort-by-year and cohort-by-county fixed effects.

The patent-level estimates reinforce the conclusion that railroad access did not improve patent quality. The coefficient on $Treat \times Post$ for importance remains negative across all specifications. It is small and statistically insignificant when importance is measured using the five-year ratio, but becomes marginally significant at the 10 percent level, with estimates ranging from approximately -0.012 to -0.013, when the ten-year ratio is used. The estimated effect on novelty is close to zero and statistically insignificant throughout. The weak positive effect on impact observed in the county-level analysis disappears once technological composition is held constant.

Overall, these results suggest that the railroad’s primary effect on innovation operated through an increase in the quantity of patenting rather than through changes in patent quality. The estimated effects on importance, novelty, and impact are small relative to those on patent counts.

5.3 Treatment intensity analysis

In the previous sections, I focused on the average treatment effect of first railroad access. I now examine treatment intensity by asking not only whether a county received a railroad, but also how intensively it was served by the railroad mail network. I use two measures of treatment intensity constructed from the railroad mail-route data. The first is *Trips*, defined as the total number of weekly railroad mail trips passing through a county, which captures the frequency of mail service and thus the intensity of information exchange. The second

is *Routes*, defined as the number of distinct railroad mail routes serving a county, which captures the breadth of a county’s connections within the communication network.

To assess whether more intensive mail service amplified the innovation effects of railroad access, I estimate a treatment-intensity specification by interacting $Treat \times Post$ with *Trips* and *Routes*, following the specification in Model (6). Because railroad mail-route data are available only from 1843 to 1880, I restrict the sample to cohorts with treatment years between 1848 and 1875, ensuring that the full eleven-year event window surrounding each treatment cohort is observed within the mail-route data period.

Columns (1) and (2) of Panel A of Table 4 report the results using *Trips* as a measure of service intensity. The baseline effect of railroad access, captured by $Treat \times Post$, is positive and highly significant. The OLS estimate implies an increase of approximately 10 percent in patenting, while the Poisson estimate suggests an increase of roughly 22 percent in patent counts. The coefficient on $Treat \times Post \times Trips$ is positive and statistically significant across all specifications (approximately 0.004), indicating that, among treated counties, the innovation gains were larger where rail mail service was more intensive. These patterns are robust across estimation methods and fixed-effect specifications, including county, year, and cohort fixed effects as well as cohort-by-year and cohort-by-county fixed effects.

The magnitude of the effect is economically significant. The Poisson coefficient of 0.004 implies that each additional weekly mail trip increases the post-railroad patent gain by approximately 0.4 percent. As a result, a treated county at the 90th percentile of mail service intensity among treated counties, roughly 20 weekly trips, experiences an increase in patenting of about 33 percent, compared with approximately 22 percent for an otherwise similar treated county with no rail mail service. This dose-response relationship is consistent with the view that the innovation benefits of railroad access increase with the intensity of information flows transmitted through the rail mail network.

The same pattern holds on the extensive margin. Columns (3) and (4) of Panel A of Table 4 report the results using *Routes* as a measure of treatment intensity. Replacing *Trips*

with *Routes*, I find that the coefficient on $Treat \times Post \times Routes$ remains positive and statistically significant under both OLS and Poisson estimation (0.048 and 0.050, respectively). These results indicate that the innovation gains from railroad access increase as a county becomes more deeply integrated into the mail network. The magnitudes are economically meaningful. An additional mail route increases the post-access patent response by approximately 5 percent. A county at the 90th percentile of route coverage, which is served by one mail route, experiences about a 26 percent increase in patenting, compared with roughly 20 percent for an otherwise similar county that gained railroad access but received no mail service. The baseline effect of railroad access remains positive and statistically significant across all specifications (0.086 under OLS and 0.178 under Poisson), and the estimates are robust to the inclusion of county, year, cohort, cohort-by-year, and cohort-by-county fixed effects. Taken together, the results in Panel A suggest that the innovation benefits of railroad access were greatest in counties receiving the most intensive mail service, whether measured by the frequency of mail deliveries or the number of mail routes.

The same logic extends from the intensity of mail service a county received to its position within the railroad network. Panel B of Table 4 replaces the mail-service measures with two measures of network centrality. The first is *Degree*, defined as the number of direct rail connections. The second is *Betweenness*, which measures how frequently a county lies on the shortest paths connecting other counties. Both measures are entered as within-year ranks and are fixed at the year in which the county first became connected to the railroad network. The coefficient on $Treat \times Post \times Centrality$ is positive and statistically significant for both measures under OLS and Poisson (0.067 and 0.385 for degree, 0.081 and 0.299 for betweenness). Because the dose is fixed at the moment of connection, this gradient cannot reflect a county's network position responding to its own subsequent patenting. The magnitudes are large. For the most central connected counties, combining the baseline effect with the centrality gradient implies that patenting rose by about 15 percent in the OLS specifications and by more than 40 percent in the Poisson specifications, several times

the effect for a minimally connected county at the bottom of the network, given by the $Treat \times Post$ row. Read alongside Panel A, these estimates say that the innovation gains from railroad access were not a fixed premium for being connected; they scaled with how intensively and how centrally a county was tied into the network through which information circulated.

Having shown that treatment intensity raises the quantity of patenting, I next ask whether it also affects patent importance. Table A.7 re-estimates the intensity specifications at the county level with patent importance as the outcome, $\ln(FS_5/BS_5)$ in Panel A and $\ln(FS_{10}/BS_5)$ in Panel B, interacting $Treat \times Post$ across columns with each of the four connection measures of Table 4: mail service (*Trips* and *Routes*) and network centrality (degree and betweenness, frozen at the connection year). In contrast to the quantity results, none of the four gradients moves patent importance: all are economically small and statistically insignificant, with the centrality point estimates ranging from -0.022 to 0.001 (t -statistics below 1.5). The same frozen network position that predicts a steep gradient in patent quantity thus predicts none in its importance—even the most centrally connected counties, whose patenting rose by more than 40 percent, produced no more consequential patents. Information intensity governed how much invention occurred, not how important it was.

I next re-estimate the intensity tests at the patent level, assigning each patent the stacked treatment status of its first inventor’s county; this uses all patents rather than county-year averages and adds technology-class (CPC) fixed effects. Table A.8 reports importance, novelty, and impact, interacting $Treat \times Post$ with *Trips* (Panel A) and *Routes* (Panel B). Delivery frequency is unrelated to every quality measure, and neither margin moves patent importance. The lone exception is novelty on the extensive margin: $Treat \times Post \times Routes$ for novelty ($1/BS_5$) is positive and significant (0.015 , $t = 1.89$; 0.020 , $t = 2.34$ with the saturated fixed effects), implying patents about 2 percent less similar to prior art, while impact (FS_5) turns marginally negative in the most saturated column. More mail routes

thus yielded somewhat more novel but no more important patents.

Taken together, the evidence from both the intensive margin, *Trips*, and the extensive margin, *Routes*, points to the same conclusion: the innovation gains from railroad access are concentrated in counties that received more extensive and more intensive railroad mail service. These gains extend beyond the quantity of innovation. Along the extensive margin, counties served by a larger number of mail routes also produced more novel patents. Thus, greater railroad mail service increased not only the amount of innovation but also the originality of the innovations produced.

5.4 Heterogeneity by innovation hub status and water access

A natural question is whether the innovation gains from railroad access accrued disproportionately to counties that were already centers of invention. If access to knowledge mainly benefits places with the absorptive capacity to exploit it, the effect should concentrate in established innovation hubs, and railroad expansion would have reinforced the existing geography of invention rather than broadening it. The opposite prediction is equally plausible: established hubs were already information-rich, so the information constraint may have bound hardest in peripheral counties, in which case the gains should concentrate outside the hubs and railroad expansion would have broadened the geography of invention instead.

To test this, I split the treatment effect by a predetermined measure of a county’s inventive standing. For each cohort, I sum a county’s patents from 1836 through the year preceding the cohort’s treatment year and rank that stock nationally against all populated U.S. counties as of the same year. A county is classified as a hub if this rank lies at or above the 75th percentile, the top quartile. Because the measure is fixed before treatment, it cannot be contaminated by the railroad’s own effect on patenting. I then replace the baseline $Treat \times Post$ term with two triple interactions, $Treat \times Post \times Hub$ and $Treat \times Post \times NonHub$, so that each coefficient is read directly as the difference-in-differences effect within its group and their difference is the formal test of heterogeneity.

I run the same test along a second dimension, access to water transportation. *Water* equals one for matched pairs whose treated county intersects a navigable river, a canal, the Atlantic coast, or the Great Lakes, with each control county inheriting the classification of its treated county. Unlike the hub split, this split carries a directional prior. Mail already traveled along canals, rivers, and coastal routes, so a county with a waterway connection had a channel for information exchange before the railroad arrived; if railroad access operates through information, its incremental effect should accordingly be weaker in water counties than in counties for which the railroad was the first low-cost link.

Table 5 reports the results for patent quantity, and the headline is a positive, highly significant effect in both groups. The baseline OLS estimate is 0.135 for hub counties and 0.099 for non-hub counties. The hub point estimates are somewhat larger, though the two groups are not statistically distinguishable, so railroad access raised patenting substantially in established innovation hubs and in less inventive counties alike. A stricter top-decile definition of hub status yields the same conclusion (Appendix Table A.3).

Panel B examines the same question for water access. Counties with and without water transportation both gained significantly, and by nearly identical amounts in the OLS columns (0.107 and 0.106 under the baseline). The Poisson point estimates are somewhat lower for water counties, in the direction the substitution hypothesis anticipates: counties that already had a waterway, and thus a prior channel for obtaining information, appear to have relied somewhat less on the railroad, though the gap is not statistically significant. On the quantity margin the railroad's effect was nonetheless broadly shared, reaching counties that already had a waterway about as strongly as those for which the railroad was the first low-cost link.

Table 6 repeats the exercise for patent importance, measured by the natural logarithm of the ratio of ten-year forward to five-year backward similarity, and here it is the water split that responds sharply, in the direction opposite to the substitution prediction. Rather than being weaker where a waterway already existed, the importance effect is positive in counties with water access and negative in those without, and the difference between them

is economically large and highly significant. Railroad access raised the importance of locally produced invention where the county also had access to water transportation, but lowered it where water access was absent. The hub split, by contrast, shows little importance heterogeneity: the effect is small and statistically insignificant for hub counties and negative and marginally significant for non-hub counties, with a positive but not significant difference. The hub pattern is unchanged under a stricter top-decile definition (Appendix Table A.4).

Together, these tests reveal a nuanced pattern. Patent quantity increased by a similar magnitude across established innovation hubs and peripheral counties, as well as counties with and without water transportation. In contrast, the quality of innovation was heterogeneous. Patent importance increased only in counties that combined railroad access with water transportation, while it declined in counties without water access. Appendix Table A.9 further shows that, among counties with water transportation, the increase in patent importance was driven by innovation hubs, where the estimated effect is positive and statistically significant. By comparison, the effect is economically and statistically insignificant in non-hub counties with water access.

These findings directly address the concern that pre-existing waterways rendered railroad-based information transmission redundant. If waterways had fully substituted for railroad mail, counties with water access should have benefited less from railroad expansion. Instead, these counties experienced similar gains in patent quantity and larger gains in patent importance. The results also clarify the aggregate null effect on patent quality documented above. The average masks substantial heterogeneity, with patent importance rising in water-access innovation hubs while falling in counties lacking both water transportation and an established innovative base. Overall, the evidence suggests that railroad mail broadly relaxed constraints on the quantity of inventive activity by improving access to information. However, transforming that information into high-impact innovation still depended on complementary local conditions, particularly the combination of transportation infrastructure and an existing stock of innovative capabilities.

6 Mechanism: County-Pair Similarity Analysis

In this section, I explore the potential channels underlying the observed effects. I begin with describing county-pair sample construction. Then I describe regression specification and provide empirical evidence for the underlying channel.

6.1 County-pair sample construction and empirical methodology

If information and knowledge diffuse along railroads, then counties connected by the same line should share more similar knowledge, and consequently the patents generated in those counties should become more alike after the connection. As in my baseline analysis, testing this underlying channel requires addressing the concern of the heterogeneity of the treatment effect and the heterogeneity between treatment and control counties.

To address the concern of the heterogeneity of the treatment effect, I construct county-pair level stacked DID sample. Specifically, I start by identifying the treatment years in which a county pair was first connected by a common mail route. For each of these treatment years, I create a cohort defined by a five-year window before and after the connection. County pairs that became connected for the first time in the treatment year are classified as treated pairs, while pairs that remained unconnected during the 11-year window—i.e., not-yet-treated or never-treated pairs—form the pool of potential controls within that cohort.³⁶ I restrict attention to cohorts whose treatment year falls between 1848 and 1875, so that the full eleven-year window of every cohort lies within the 1843–1880 period spanned by the mail-route data.

To address heterogeneity between treated and control counties, within the stacked sample I perform matching at the directed-pair level, anchoring on the focal county. Specifically, for each focal county, I match a treated partner county to a control partner county selected

³⁶The design is directional throughout. Each connected county pair enters the sample as two ordered observations: one with county (A) as the focal county and county (B) as the partner county, and another with the roles reversed. This structure is necessary because the outcome itself is directional, comparing newly generated patents in the focal county with prior patents in the partner county and thereby capturing the direction of knowledge diffusion.

from the same focal county’s unconnected counterparts. As a result, each treated pair and its matched control share the same focal county and differ only in the partner county.

To select a control partner with characteristics that most closely resemble the treated partner, I estimate propensity scores separately for each cohort using a logit model of treatment status on the log of partner population, the partner’s urban and white population shares, the log distance between the focal and partner counties, and the focal–partner patent similarity one and four years before connection. Conditioning on these pre-connection similarities helps ensure that treated and control pairs exhibit comparable pre-treatment trends. I then perform one-to-one nearest-neighbor matching on the estimated propensity score without replacement. This procedure assigns each treated focal-partner pair a single control focal-partner pair from the same focal county, balancing both pre-connection county characteristics and pre-connection patent similarity trajectories.

6.2 Regression specification and results

Using the stacked and matched DID sample, I examine how mail-route connections affect text-based patent similarities between county pairs. Specifically, I estimate the following specification:

$$\begin{aligned} Similarity_{f,p,t,g} = & \beta Connected_{f,p,g} \times Post_{t,g} + \alpha_1 Connected_{f,p,g} + \alpha_2 Post_{t,g} \\ & + \mathbf{X}_{f,p,t} + \gamma_{f,t} + \gamma_{p,t} + \gamma_g + \epsilon_{f,p,t,g}, \end{aligned} \quad (7)$$

where $Similarity_{f,p,t,g}$ is the within-year percentile rank of the average text similarity between focal county f ’s patents in year t and partner county p ’s patents in years $[t - 10, t - 3]$, within cohort g . I start the window at $t - 10$ to include more patents into the stock of prior knowledge, and I end the window at $t - 3$ to ensure that it captures the partner’s established knowledge stock while avoiding the immediate period preceding the connection, when anticipation effects may arise.

$Connected_{f,p,g}$ is an indicator equal to one if focal county f and partner county p in

the treated pair become connected by a common mail route in cohort g . $Post_{t,g}$ equals one in years on or after the connection year of cohort g . The control vector $\mathbf{X}_{f,p,t}$ contains the distance between counties f and p and the log number of cross-county patent pairs underlying the similarity measure. In the baseline specification I include focal-county $\times$ year $\gamma_{f,t}$, partner-county $\times$ year $\gamma_{p,t}$, and cohort γ_g fixed effects. In the saturated specifications I replace the cohort fixed effects with directed pair $\times$ cohort fixed effects; the final specification further excludes geographically nearby pairs. The coefficient β on $Connected_{f,p,g} \times Post_{t,g}$ measures the effect of a mail-route connection on between-county patent similarity, and I expect β to be positive and significant.

Table 7 reports the estimates. Across all three specifications, gaining a mail-route connection significantly increases the similarity between the patents produced by a county pair. In the baseline specification shown in column (1), the coefficient on $Connected \times Post$ is 0.266 ($t = 2.26$). This estimate implies that, following a new connection, the focal county's patents move approximately 0.27 percentile points higher in the within-year distribution of similarity to the partner county's prior patents. The estimate becomes larger and more precisely estimated in column (2), where the cohort fixed effects are replaced by directed-pair by cohort fixed effects. Under this more demanding specification, the coefficient increases to 0.365 ($t = 3.11$), indicating that the result is identified from within-pair changes before and after the connection. The increase in magnitude suggests that the findings are not driven by connected county pairs being inherently more similar prior to the railroad connection.

The remaining coefficients support this interpretation. In column (1), the coefficient on $Connected$ is positive and highly significant, indicating that connected county pairs are substantially more similar than unconnected pairs even before the connection occurs. This pattern likely reflects underlying geographic and economic proximity that makes two counties more likely to be linked by a mail route. Because this pre-existing similarity represents selection rather than treatment, it is absorbed by the directed-pair by cohort fixed effects in columns (2) and (3). The estimated post-connection effect of 0.365 is roughly 30 percent

of the pre-existing similarity gap of 1.222, suggesting that the mail-route connection itself accounts for a meaningful share of the observed convergence in patent content. Consistent with expectations, the included controls enter with the anticipated signs: the pair distance is negative and the logarithm of the number of cross-county patent pairs used to construct the similarity measure is positive. Finally, the results remain robust when nearby county pairs, those whose centroids lie within 35 miles of each other, are excluded from the sample. As shown in column (3), the coefficient remains virtually unchanged at 0.354 ($t = 2.95$), indicating that the findings are not driven by local spillovers or the mechanical sharing of inventors across nearby counties.

To examine the timing of the effect and assess the parallel-trends assumption underlying the research design, I estimate a dynamic version of Equation (7) that replaces $Connected_{f,p,g} \times Post_{t,g}$ with a full set of interactions between $Connected_{f,p,g}$ and indicators for each year relative to the connection, omitting the year immediately preceding the connection as the reference period. Figure 5 plots the resulting coefficients for the three specifications and reveals two key findings. First, the pre-connection coefficients are small, statistically insignificant, and exhibit no systematic trend across any of the three panels. This indicates that treated and control pairs followed similar trajectories prior to the connection. Consistent with this interpretation, the lead coefficients are jointly indistinguishable from zero across all specifications. The result remains robust under the baseline fixed-effects specification in Panel A, the directed-pair-by-cohort fixed-effects specification in Panel B, and the specification excluding nearby pairs in Panel C, indicating that the estimated effects are not driven by pre-existing differences or divergent pre-treatment trends between treated and control counties.

Second, the effect emerges at the time of connection and grows steadily thereafter. The estimated coefficients become positive in the year of connection and increase to approximately 0.6 percentile points by the fifth year. This pattern, characterized by flat pre-treatment coefficients followed by a persistent increase in similarity, is consistent with counties drawing on the existing stock of knowledge produced by their newly connected partners. The ef-

fect accumulates over time as additional post-connection patents are generated, rather than reflecting a contemporaneous shock affecting both counties at the moment of connection.

Finally, I ask whether this convergence reflects diffusion from the established innovation centers or circulation among peripheral counties. The research design disciplines this question. Because the focal-county-by-year fixed effects and the matched control pairs absorb any change in a county’s overall access to the national stock of knowledge, the estimates capture information flows that depend on the specific link between the two counties. Splitting the connection effect by whether the partner county is an innovation hub then yields opposing predictions. If the marginal value of a new link lay in reaching the technological frontier, the convergence should be concentrated among hub partners: newly connected counties would absorb the leading centers’ superior technology. If instead the leading centers’ knowledge already traveled widely while the knowledge of ordinary counties circulated narrowly, then a direct link supplies new information precisely when the partner is a peer, and the convergence should be concentrated among non-hub partners.

Panel B of Table 7 implements this split, interacting $Connected \times Post$ separately with hub-partner and non-hub-partner status, defining a hub as a partner county whose pre-treatment national patent-stock rank lies in the top quartile. The convergence is driven entirely by non-hub partners: the coefficient on $Connected \times Post \times Non\text{-}hub\ partner$ is positive and highly significant across all three specifications—about 1.1 to 1.3 percentile points—whereas $Connected \times Post \times Hub\ partner$ is small and never statistically distinguishable from zero, so that the two groups differ sharply (the Hub – Non-hub gap is roughly -1.1 , $t \approx -3.2$). This pattern is robust to defining hubs at the top decile (90th percentile) as well (Table A.12).

The evidence in Panel B is consistent with the second hypothesis: the information introduced by a new mail-route connection primarily originated from ordinary counties rather than from established innovation centers. The estimates identify the direction of these marginal information flows, indicating that knowledge circulated from ordinary counties to

newly connected locations. This pattern is consistent with the county-level results presented earlier: the benefits of improved mail access are concentrated on the extensive margin of innovation and among ordinary patents.

7 Additional Analyses

In this section, I conduct four additional analyses to probe the information channel and address identification concerns. First, I exploit the mid-nineteenth-century postal reforms as an exogenous reduction in communication costs, isolating mail-based information from passenger travel. Second, I use the Civil War’s disruption of Confederate postal service as a reverse shock to the information network. Third, I restrict attention to counties incidentally connected along mail routes, whose access was a byproduct of linking demand-selected endpoints. Finally, I test for knowledge spillovers to neighboring counties, excluding adjacent controls and estimating the spillover directly.

7.1 Postal reforms

Railroads were more likely to expand into populous and economically active areas, where demand for transportation was highest and where inventors were more likely to reside and innovate. This raises an important question: is the observed effect on innovation truly driven by information flows, and if so, what form does that information take? In principle, the effect could arise from the diffusion of tacit knowledge through face-to-face interactions facilitated by railroad travel, or from the dissemination of codified knowledge through letters, newspapers, and other printed materials that inventors could access, study, and recombine at scale. Distinguishing between these channels requires a shock that changes the cost of exchanging written information while leaving the movement of people largely unaffected.

The mid-nineteenth-century postal reforms provide a useful setting for examining the role of information flows in innovation for three reasons. First, a sequence of postal reforms

progressively reduced and simplified the cost of sending letters, triggering a dramatic expansion in mail volume. Before 1845, postage rates varied across five distance categories, making long-distance correspondence particularly expensive. For example, mailing a half-ounce letter from Baltimore to New York in 1840 cost 18.75 cents. The Postal Act of 1845 simplified this schedule into two distance zones and sharply reduced rates, lowering the cost of the same letter to 5 cents, a reduction of 73 percent. The Cheap Postage Act of 1851 further reduced communication costs by extending the lowest rate to nearly the entire country and encouraging prepayment, thereby greatly diminishing the importance of distance in determining postage. The Act of 1863 completed this process by eliminating distance-based pricing altogether, establishing a uniform rate based solely on weight and introducing free city delivery in larger cities.³⁷ Across these reforms, total mail volume roughly doubled within a few years of the 1845 Act as lower postage rates stimulated correspondence and expanded newspaper circulation through greater postal traffic and improved system efficiency (John, 1995; Kielbowicz, 2002; John, 2015).

Second, railroad mail service operated under exclusive postal contracts that specified fixed delivery schedules and were separate from commercial passenger operations. As a result, the postal reforms increased mail volume and demand for mail transportation without directly affecting passenger capacity. This institutional feature helps isolate the diffusion of written information from the movement of people and goods. Third, the reforms were motivated primarily by political and fiscal considerations rather than local innovation activity, making them well suited for studying how reductions in communication costs influence innovation.³⁸

³⁷Under the schedule in force before 1845, a single-sheet letter cost between roughly 6 and 25 cents depending on distance, ranging from about 6 cents for distances under 30 miles to 25 cents for distances exceeding 400 miles. Each additional sheet was charged as a separate letter. The Act of March 3, 1845 (effective July 1, 1845) replaced this system with two weight-based rates: 5 cents per half-ounce for distances under 300 miles and 10 cents for longer distances. The Act also restricted private letter carriers. The Cheap Postage Act of March 3, 1851 (effective July 1, 1851) lowered the rate to 3 cents per half-ounce for prepaid letters traveling up to 3,000 miles, while unpaid letters were charged 5 cents. Letters traveling farther were charged double rates. The Act of March 3, 1863 (effective July 1, 1863) eliminated distance from letter pricing, established a uniform rate based on weight, classified mail into first-class letters, second-class periodicals, and third-class printed matter, and authorized free city carrier delivery, initially implemented in the largest cities.

³⁸The postal administration explicitly framed each reform as a response to political and fiscal pressures

Building on this setting, I examine how the effect of railroad access on innovation evolved across successive postal regimes. My hypothesis is that railroad access has a larger effect on patenting as postal reforms progressively reduce the cost of exchanging knowledge and disseminating ideas. As communication becomes cheaper, the innovation benefits of railroad access should increase across successive reform regimes. To test this hypothesis, I trace how the railroad’s effect on innovation evolved across the successive postal regimes. Specifically, I construct two indicator variables: *1851Reform*, which equals one for the years 1851 through 1862, and *1863Reform*, which equals one for years from 1863 onward. I pool the pre-1845 and 1845–1850 years into a common pre-1851 period and let the pre-1851 period serve as the omitted baseline category.³⁹ I interact each reform indicator with the key variables from the baseline specification (Model 4), yielding a triple-difference framework in which $Treat_c \times Post \times 1851Reform$ and $Treat_c \times Post \times 1863Reform$ measure the incremental effect of railroad access on patenting during each reform regime relative to the pre-1851 period.

I restrict the estimation sample to railroad cohorts with treatment years between 1840 and 1868. This restriction ensures that, after incorporating five-year pre- and post-treatment event windows, each cohort spans the period covering the 1845–1863 postal reforms. Con-

originating outside the postal economy. The Act of March 3, 1845 was prompted by competition from private mail-express companies that had emerged along the new railroad routes, threatening the postal monopoly’s revenue base; the Postmaster General warned in 1843 that "the revenue of the department will in time be so far affected by the inroads of private expresses, that the service will either have to be reduced below the just wants of the public, or appropriations from the General Treasury will be required" (*Annual Report of the Postmaster General* 1843, p. 7). The Act of March 3, 1851 responded to a populist cheap-postage movement: "A considerable reduction of letter postage. . . is desired by a large portion of the people of the United States" (*Annual Report* 1850, p. 5). The Act of March 3, 1863 was driven by Postmaster General Montgomery Blair’s push for administrative uniformity — abolishing distance-based rates and authorizing free city delivery — rather than rate reduction per se (*Annual Report* 1862, pp. 27–31). None of the three reforms responded to local innovation conditions, and contemporaneous documentation invokes neither inventive activity nor patent statistics as a rationale.

³⁹I use the pre-1851 period, rather than the pre-1845 period, as the baseline because of the cohort structure of the stacked design. Among the railroad cohorts included in the estimation sample, only the 1840 cohort is observed in its post-treatment window before 1845. Consequently, a separately identified effect of the 1845 reform would rely on a single cohort. Pooling the pre-1845 and 1845–1850 years into a common pre-1851 baseline instead allows the reference period to draw on the 1840, 1845, and 1850 cohorts, yielding a more stable benchmark. The effect of the 1845 reform is therefore absorbed into the baseline period, and the coefficients for 1851 and 1863 should be interpreted as the effect of railroad access under those postal regimes relative to the already reduced post-1845 communication-cost environment.

sequently, the pre-1851 baseline as well as the 1851 and 1863 postal regimes are observed within every cohort’s event window.

Table 8 reports the results on patent quantity. The triple-interaction coefficients are positive and statistically significant in the OLS specifications, columns (1) and (3), and increase across reform regimes. Relative to the pre-1851 baseline, railroad access raises patenting by approximately 7 to 8 percent more during the 1851 regime and by 19 to 20 percent more during the 1863 regime. The Poisson estimate for the 1851 regime in column (2) is also positive and statistically significant. Combined with the baseline $Treat_c \times Post$ coefficient, these estimates imply that the effect of railroad access on patenting increases monotonically over time, rising from roughly 5 percent before 1851 to about 14 percent under the 1851 reform and nearly 28 percent under the 1863 reform. This pattern is consistent with the hypothesis that lower communication costs magnify the innovation benefits of railroad access. As successive postal reforms reduced the cost of written communication, counties connected to the railroad network experienced increasingly larger gains in patenting relative to their matched controls. The Poisson estimates for the 1863 regime in columns (2) and (4) are not statistically significant. Because Poisson estimation places greater weight on the small number of counties with exceptionally high patent counts, it is more sensitive to outliers, whereas the OLS estimates capture the broader extensive-margin response that underlies the observed gradient across reform regimes.

The quantity gains documented above do not necessarily translate into higher-quality innovation. To examine this possibility, I re-estimate the triple-difference specification using three measures of patent quality as dependent variables: *Novelty*, *Impact*, and *Importance*. The results are reported in Appendix Table A.10. In contrast to the quantity results, I find no robust evidence that railroad access affected patent quality during the postal reform era. The OLS estimates are statistically insignificant across all quality measures and reform regimes.

Together, these findings help answer the question raised at the outset: did railroads foster

invention by transporting people and facilitating the exchange of tacit knowledge through face-to-face interactions, or by transmitting written information and thereby diffusing codified knowledge? The postal reforms provide a useful test of this distinction. By reducing the cost of exchanging written information while leaving passenger travel largely unaffected, the reforms increased communication intensity without directly changing opportunities for in-person interaction, since railroad mail operated under contracts separate from commercial passenger service. The fact that the patenting advantage of railroad-connected counties increased as successive reforms reduced communication costs points to the diffusion of codified knowledge, transmitted through letters, newspapers, and other printed materials, as an important mechanism through which railroads stimulated invention.

The amplification, however, is concentrated on the quantity margin. Cheaper communication combined with railroad access increased the volume of patenting without producing detectable changes in patent novelty, impact, or importance. This pattern suggests that access to codified knowledge raised the volume of patenting by lowering the cost of obtaining knowledge rather than by improving the average quality of inventions. More broadly, the evidence identifies the circulation of codified knowledge as one channel through which railroads fostered innovation. At the same time, because the analysis isolates variation in communication costs while holding passenger travel largely fixed, it cannot rule out a complementary role for tacit knowledge in explaining the overall effect of railroads on innovation. Nevertheless, the results establish that the diffusion of written information was an important mechanism through which railroad access translated into new ideas.

7.2 Civil war railroad mail suspension

The preceding results show that railroads, together with the rail mail services they carried, increased patenting by improving the flow of information. However, the Post Office Department, much like the railroad companies with which it contracted, was likely to expand service toward counties already experiencing population growth and commercial develop-

ment. Consequently, the observed association between communication access and patenting may partly reflect reverse causality, whereby mail service followed economic opportunity rather than generated it.

The Civil War provides a useful setting for addressing this concern. Following secession, the U.S. Post Office Department terminated mail service throughout the Confederacy in May 1861.⁴⁰ The termination of railroad mail contracts disrupted a major channel of information transmission in the South, cutting off Confederate counties from the national circulation of correspondence, newspapers, and printed materials for the duration of the Civil War.

Figure 4 illustrates the magnitude of this disruption by plotting the number of railroad mail routes and the total mileage of those routes in Union and Confederate states from 1843 to 1880. Both measures reveal a striking disruption caused by the Civil War. In the Union, railroad mail service remained largely intact throughout the conflict, with route counts and mileage remaining stable before rising rapidly after the war. In contrast, Confederate railroad mail service collapsed. After expanding steadily through 1860, both the number of routes and total route mileage fell abruptly to zero during the war years and remained there until the Confederacy's defeat. Recovery did not begin until after 1865, highlighting the severe wartime breakdown of communication infrastructure in the South.

This abrupt loss of communication infrastructure provides a powerful test of the information channel for two reasons. First, it is plausibly exogenous to local economic conditions, as the suspension resulted from secession and military conflict rather than from the economic or innovative trajectory of any individual county. Second, it represents the opposite of the postal reforms examined earlier. Whereas the Postal Act reduced the cost of written communication and expanded information flows, the Civil War effectively eliminated access to those flows, allowing me to examine whether the removal of information channels reduced innovative activity.

⁴⁰Mail service in the eleven Confederate states—Alabama, Arkansas, Florida, Georgia, Louisiana, Mississippi, North Carolina, South Carolina, Tennessee, Texas, and Virginia—was terminated on May 31, 1861 under the Act of February 28, 1861 (excepting West Virginia), and remained suspended until 1866. See *Annual Report of the Postmaster General*, 1861, p. 10; 1862, p. 5.

If access to written knowledge fostered innovation, counties that lost mail service should have experienced a disproportionate decline in inventive activity. I test this prediction by comparing Confederate counties whose railroad mail routes were discontinued with otherwise similar counties whose service remained intact, before the war (1856–1860) and during the conflict (1861–1865). Because the war affected Southern counties through many channels beyond the postal system, I complement the baseline difference-in-differences analysis with a triple-difference specification that differences out the average wartime shock to Southern counties, thereby isolating the effect attributable to the loss of mail service.

To implement this comparison, I use the full county-year panel of all U.S. counties from 1856 to 1865, covering the five pre-war years and the five years of the Civil War, rather than the matched stacked sample employed in the main analysis. *War* is an indicator equal to one for the war years 1861 to 1865 and zero for the prewar period 1856 to 1860. The treatment indicator, *SouthLostRoute*, equals one for counties located in Confederate states that had at least one railroad mail route before the war but none during the war. All other counties constitute the comparison group. In a more restrictive specification, I limit the sample to counties that operated at least one railroad mail route before the war, thereby comparing Southern counties that lost mail-route access with Northern counties that remained connected throughout the period. My baseline specification is the difference-in-differences regression,

$$Patent_{c,t} = \beta (SouthLostRoute_c \times War_t) + \mathbf{X}'_{c,t}\gamma + \alpha_c + \tau_t + \varepsilon_{c,t}, \quad (8)$$

where α_c and τ_t are county and year fixed effects and $\mathbf{X}_{c,t}$ is a vector of county controls comprising log population, the urban and white population shares, log manufacturing capital, log farm value, log boat frequency, and the log distances to the nearest river and canal. I estimate the equation by OLS, with the outcome measured as $\ln(1 + Patent_{c,t})$, and by Poisson on the patent count.⁴¹ The coefficient of interest, β , measures the change in patenting in

⁴¹The county fixed effects absorb the time-invariant *SouthLostRoute* indicator and all permanent differences

counties that lost their mail, relative to other counties, from before to during the war. Standard errors are clustered by county.

To distinguish the effect of the mail loss from the broader collapse of the wartime South, I also estimate the triple-difference specification,

$$\begin{aligned} Patent_{c,t} = & \beta (PreWarRoute_c \times South_c \times War_t) \\ & + \gamma_1(PreWarRoute_c \times War_t) + \gamma_2(South_c \times War_t) + \mathbf{X}'_{c,t}\theta + \alpha_c + \tau_t + \varepsilon_{c,t}, \end{aligned} \quad (9)$$

where *PreWarRoute_c* indicates a county with a pre-war mail route and *South_c* a Confederate county. The interaction *PreWarRoute* $\times$ *War* controls for the common wartime trajectory of route-bearing counties and *South* $\times$ *War* for the average wartime shock to all Southern counties, so that the triple coefficient β isolates the additional decline borne by Southern counties whose mail was discontinued.⁴²

Table 9 reports the results. Columns (1) and (2) present the baseline difference-in-differences estimates using the full sample. In both specifications, the coefficient on *SouthLostRoute* $\times$ *War* is large, negative, and highly significant. The OLS estimate in column (1) implies that counties whose mail service was discontinued generated approximately 24 percent fewer patents than other counties during the war. The Poisson estimate in column (2) implies a decline of roughly 80 percent in patent counts. The difference in magnitude reflects the distinct weighting schemes of the two estimators. The Poisson specification places greater weight on high-patenting counties whose innovative activity nearly disappeared during the war, whereas the OLS estimate captures the broader extensive-margin decline across all counties. Column (3) restricts the sample to counties that had railroad mail service before the war, comparing Southern counties that lost mail service with Northern counties that retained it. The estimated decline increases to approximately 30 percent, confirming that the result reflects the loss of mail service rather than differences between connected Southern

across counties; the year fixed effects absorb *War* and any nationwide shocks.

⁴²The remaining lower-order terms—*PreWarRoute*, *South*, their interaction, and *War*—are absorbed by the county and year fixed effects.

counties and otherwise unconnected rural areas.

A natural concern is that the decline simply reflects the broader economic collapse of the wartime, including the effects of blockades, physical destruction, and military conscription, all of which affected Confederate counties regardless of their access to mail. Column (4) addresses this concern directly. The triple-difference specification absorbs both the average Southern wartime shock $South \times War$ and the general wartime trajectory of counties with prewar mail routes $PreWarRoute \times War$. The triple interaction $PreWarRoute \times South \times War$ therefore captures the additional decline experienced specifically by counties that lost mail service. The estimated coefficient of -0.262 is nearly identical to the baseline estimate and remains highly significant. Even relative to other Southern counties exposed to the same wartime conditions, counties whose mail service was suspended experienced substantially lower patenting activity. The evidence therefore points to the disruption of information flow, rather than a general Southern downturn, as the primary driver of the decline. One concern is that route-bearing counties were more economically developed and strategically important prior to the war, making them more vulnerable to wartime disruptions unrelated to mail service. As a result, I cannot completely rule out alternative explanations, and the estimated effect should be interpreted as an upper bound on the mail-specific effect. Even under this conservative interpretation, the magnitude of the effect suggests that losing access to mail had important consequences for innovation.

Taken together, these results provide complementary evidence for the information channel. Whereas lower postage rates strengthened the relationship between railroad access and innovation, the wartime disruption of mail service sharply weakened it. Because the suspension of mail routes was driven by secession and military conflict rather than local economic conditions, the estimated decline is plausibly causal and less susceptible to concerns about reverse causality. Counties that were cut off from the national circulation of written information generated substantially fewer patents, even after accounting for the broader effects of the Civil War. These findings are consistent with access to written information transmitted

through the mail being an important input into nineteenth-century innovation.

7.3 Incidental connections

A natural concern with the baseline estimates is that railroads and the mail routes they carried were constructed where demand for transportation and communication was greatest. Railroads were designed to connect commercially important locations, implying that a county's connection may reflect its economic prospects as much as influence them. In particular, rapidly growing counties may have been more likely to attract both railroad connections and inventive activity. A common strategy in the infrastructure literature addresses this concern by distinguishing between locations that a route was intended to serve and those it merely traverses. A route's origin and destination were selected for their economic importance, whereas the intermediate counties were connected as a byproduct of linking those endpoints so that their access to mail service was less likely to reflect their own demand for mail or economic prospects.⁴³

I apply this logic to the railroad mail routes and re-estimate the baseline difference-in-differences specification after excluding treated counties that served as route endpoints, either origins or destinations, during the event window. The remaining treated counties consist exclusively of pass-through counties. If the estimated effect of railroad access reflects the endogenous selection of economically promising endpoints, once those endpoints are excluded from the sample the effect should weaken or disappear. If, instead, railroad access causally promotes innovation, the effect should remain among intermediate counties whose connection was a byproduct of linking the endpoints rather than a response to their own economic potential.

I restrict the analysis to railroad cohorts whose event windows fall within the period for which mail-route data are available, corresponding to treatment years between 1848 and

⁴³See [Banerjee, Duflo, and Qian \(2020\)](#), [Chandra and Thompson \(2000\)](#), [Michaels \(2008\)](#), [Redding and Turner \(2015\)](#), and [Faber \(2014\)](#) among others.

1875.⁴⁴ I then re-estimate the baseline difference-in-differences specification first using the restricted matched sample and subsequently after excluding treated counties that served as route endpoints. Table 10 reports the results. Consistent with the main analysis, railroad access significantly increases patenting in the full sample (columns 1 and 2). Excluding demand-selected endpoints leaves the results largely unchanged. Among pass-through counties, the OLS estimate declines by roughly one-third, from 0.133 to 0.092, but remains economically meaningful and highly significant ($t = 6.5$). The Poisson estimate is, if anything, larger, increasing from 0.238 to 0.273. The persistence of the treatment effect among intermediate counties is difficult to reconcile with the possibility that the estimated effect merely reflects endogenous route placement.

The results are also robust to alternative definitions of route endpoints. Appendix Table A.11 reports three alternative specifications: excluding the matched control county alongside each excluded endpoint, defining endpoints using the matched railroad network rather than the mail route endpoints on the Annual Reports, and excluding any county that ever served as an endpoint. Across all specifications, the estimated effect for pass-through counties remains remarkably stable, at approximately 0.09 in the OLS regressions and 0.26 in the Poisson regressions, with statistical significance at the one-percent level throughout. Consistent with the main findings, the effect is concentrated on the quantity margin. The corresponding estimates for patent novelty, impact, and importance are small and statistically insignificant in both the full and pass-through samples. Together with the wartime mail suspension test, which exploits an exogenous removal of communication services, this analysis exploits plausibly exogenous route placement and addresses endogeneity concerns from a complementary direction. Taken together, the evidence helps mitigate the concern that the estimated relationship reflects endogenous network siting rather than a causal effect

⁴⁴This test cannot be conducted from the railroad network alone. Because the railroad data are recorded at the segment level rather than as complete lines, they do not identify each route’s origin and destination, making it infeasible to distinguish endpoints from the intermediate counties a line passes through. Postal routes, by contrast, have explicitly defined termini, which makes the endpoint–pass-through distinction observable.

of information access.

7.4 Spatial spillovers to adjacent counties

In this section, I examine whether knowledge spillovers from treated counties to their immediate neighbors contaminate the control group and bias the estimated treatment effect. If information and knowledge diffuse from rail-connected counties into adjacent counties, then control counties bordering treated counties may themselves experience elevated patenting, thereby attenuating the measured contrast between treated and control units. I assess this concern in two complementary ways. A control county is defined as adjacent if it shares a border with any treated county in the same adoption cohort based on 1890 county boundaries. First, I exclude all adjacent control counties from the sample and re-estimate the model using only treated counties and non-adjacent controls. If adjacent controls benefit from spillovers, removing them should strengthen the treated-control contrast and increase the estimated treatment effect. Second, using the full sample, I include the interaction term $AdjControl \times Post$, whose coefficient directly captures spillovers onto adjacent control counties.

Columns (1) and (2) of Table 11 report the exclusion test. Excluding adjacent control counties leaves the coefficient on $Treat \times Post$ at 0.114 in the OLS specification and 0.232 in the Poisson specification, both positive and highly significant. These estimates exceed the corresponding full-sample baseline values of 0.110 and 0.219 (Table 2, columns (3) and (4)) by roughly 0.004 and 0.013. The shift is in the direction predicted by the spillover hypothesis, since dropping potentially contaminated neighbors should sharpen the treated-control contrast, but its magnitude is economically small: railroad access remains associated with approximately an 11 percent increase in patenting whether or not adjacent counties are excluded.

Columns (3) and (4) report the direct-measurement test, estimated on the full sample. Including the spillover term $AdjControl \times Post$ has little effect on the estimated treatment

effect. The coefficient on $Treat \times Post$ remains stable at 0.114 in the OLS specification and 0.230 in the Poisson specification, compared with baseline estimates of 0.110 and 0.219 on Table 2. The spillover coefficient is positive but economically small and statistically insignificant in both specifications. Thus, although the positive coefficients are consistent with the direction predicted by knowledge diffusion, control counties bordering treated counties do not exhibit a detectable larger increase in patenting than more distant control counties.

Taken together, the two tests lead to the same conclusion. I find no statistically significant evidence that innovative activity spilled over from treated counties to their immediate neighbors, and the estimated effect of railroad access remains stable whether adjacent control counties are excluded from the sample or their potential spillovers are explicitly accounted for. While the consistently positive coefficients leave open the possibility of information diffusion across county boundaries, the estimated treatment effect is not driven by spillover contamination of the control group.

8 Conclusion

This paper studies how information flows shape innovation by exploiting the staggered expansion of the U.S. railroad network between 1840 and 1900, a setting in which communication costs declined sharply at different times across counties. Using a stacked difference-in-differences design on a propensity score matched sample, I find that counties receiving their first railroad experienced a sustained increase in patenting activity, approximately 11 to 23 percent, relative to otherwise similar counties. The effect is concentrated on the quantity of innovation. Railroad access increased the number of patents without measurably improving their novelty, impact, or importance, suggesting that lower communication costs reduced barriers to inventive activity rather than improving the quality of the average invention. Consistent with an information-based mechanism, the gains are larger in counties more intensively served by the railroad mail network, whether measured by the frequency of mail

trips or the number of distinct mail routes.

A series of tests is consistent with the diffusion of written information through the mail as an important channel. A county-pair analysis shows that patents produced in two counties become more similar after the counties are connected by a common mail route, consistent with knowledge traveling through the communication network. Two opposing shocks to the cost of written communication further support this interpretation. The patenting advantage of rail-connected counties increased following the postal reforms of 1845 to 1863, which progressively reduced postage rates, and declined sharply when Confederate mail service was disrupted during the Civil War. Because these shocks moved the cost of written communication while leaving passenger travel largely unaffected, these patterns are consistent with the circulation of written information through letters, newspapers, and other printed materials, though they cannot rule out a complementary role for in-person contact. Additional tests address concerns about endogenous route placement. The effect remains among counties through which railroads merely passed en route to other destinations, and it is unchanged when counties adjacent to treated counties are excluded from the control group, helping to mitigate the concern that the results are driven by route placement or spillovers to neighboring controls.

Beyond these findings, the paper introduces large language model embeddings as a tool for measuring the semantic similarity, and thus the diffusion, of historical patents in settings where citation data are unavailable. More broadly, the results demonstrate that reducing the cost of transmitting information can stimulate inventive activity, a mechanism whose relevance extends well beyond the nineteenth century.

References

- Acemoglu, D. (2008). *Introduction to modern economic growth*. Princeton university press.
- Acemoglu, D., Moscona, J., & Robinson, J. A. (2016). State capacity and American technology: Evidence from the nineteenth century. *American Economic Review*, 106(5), 61–67.
- Aghion, P., Akcigit, U., & Howitt, P. (2014). What do we learn from Schumpeterian growth theory? In *Handbook of economic growth* (Vol. 2, pp. 515–563). Elsevier.
- Aghion, P., Akcigit, U., & Howitt, P. (2015). The Schumpeterian growth paradigm. *Annual Review of Economics*, 7(1), 557–575.
- Aghion, P., Antonin, C., & Bunel, S. (2021). *The power of creative destruction: Economic upheaval and the wealth of nations*. Harvard University Press.
- Aghion, P., Bloom, N., Blundell, R., Griffith, R., & Howitt, P. (2005). Competition and innovation: An inverted-U relationship. *The Quarterly Journal of Economics*, 120(2), 701–728.
- Aghion, P., & Howitt, P. (1992). A model of growth through creative destruction. *Econometrica: Journal of the Econometric Society*, 323–351.
- Aghion, P., Howitt, P., Brant-Collett, M., & García-Peñalosa, C. (1998). *Endogenous growth theory*. MIT press.
- Aghion, P., & Howitt, P. W. (2008). *The economics of growth*. MIT press.
- Ahern, K. R., & Harford, J. (2014). The importance of industry links in merger waves. *The Journal of Finance*, 69(2), 527–576.
- Akcigit, U., Grigsby, J., & Nicholas, T. (2017). *The rise of American ingenuity: Innovation and inventors of the golden age* (Tech. Rep.). National Bureau of Economic Research.
- Akcigit, U., & Kerr, W. R. (2018). Growth through heterogeneous innovations. *Journal of Political Economy*, 126(4), 1374–1443.
- Aneja, A., & Xu, G. (2024). Strengthening state capacity: Civil service reform and public sector performance during the Gilded Age. *American Economic Review*, 114(8), 2352–

2387. doi: 10.1257/aer.20230019

- Atack, J. (2015, August). *Historical geographic information systems (GIS) database of U.S. railroads for selected years*. (Digitized and assembled by Jeremy Atack, Vanderbilt University and National Bureau of Economic Research. Data may be freely used with proper attribution.)
- Atack, J. (2018). Creating historical transportation shapefiles of navigable rivers, canals, and railroads for the United States before World War I. In *The routledge companion to spatial history* (pp. 169–184). Routledge.
- 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, 1850–1860. *Social Science History*, 34(2), 171–197.
- Atkin, D., Chen, M. K., & Popov, A. (2022). *The returns to face-to-face interactions: Knowledge spillovers in Silicon Valley* (Tech. Rep.). National Bureau of Economic Research.
- Bahar, D., Choudhury, P., Kim, D. Y., & Koo, W. W. (2023). Innovation on wings: Nonstop flights and firm innovation in the global context. *Management Science*, 69(10), 6202–6223.
- Bai, J., Jin, W., & Zhou, S. (2024). Proximity and knowledge spillovers: Evidence from the introduction of new airline routes. *Management Science*, 70(11), 7464–7485. doi: 10.1287/mnsc.2021.01717
- Baker, A., Callaway, B., Cunningham, S., Goodman-Bacon, A., & Sant’Anna, P. H. (2026). Difference-in-differences designs: A practitioner’s guide. *Journal of Economic Literature*, 64(2), 498–557. doi: 10.1257/jel.20251650
- Baker, A. C., Larcker, D. F., & Wang, C. C. (2022). How much should we trust staggered difference-in-differences estimates? *Journal of Financial Economics*, 144(2), 370–395.
- Banerjee, A., Duflo, E., & Qian, N. (2020). On the road: Access to transportation infrastructure and economic growth in China. *Journal of Development Economics*, 145,

102442.

- Bena, J., & Li, K. (2014). Corporate innovations and mergers and acquisitions. *The Journal of Finance*, 69(5), 1923–1960.
- Bergeaud, A., & Verluise, C. (2024). A new dataset to study a century of innovation in Europe and in the US. *Research Policy*, 53(1), 104903.
- Bernstein, S., Diamond, R., Jiranaphawiboon, A., McQuade, T., & Pousada, B. (2022). *The contribution of high-skilled immigrants to innovation in the United States* (Tech. Rep.). National Bureau of Economic Research.
- Bernstein, S., Giroud, X., & Townsend, R. R. (2016). The impact of venture capital monitoring. *The Journal of Finance*, 71(4), 1591–1622.
- Bernstein, S., McQuade, T., & Townsend, R. R. (2021). Do household wealth shocks affect productivity? evidence from innovative workers during the Great Recession. *The Journal of Finance*, 76(1), 57–111.
- Blevins, C. (2021). *Paper trails: The US post and the making of the American West*. New York, NY: Oxford University Press.
- Blondheim, M. (1994). *News over the wires: The telegraph and the flow of public information in America, 1844–1897*. Cambridge, MA: Harvard University Press.
- Bloom, N., Schankerman, M., & Van Reenen, J. (2013). Identifying technology spillovers and product market rivalry. *Econometrica*, 81(4), 1347–1393.
- Bryan, K. A., & Ozcan, Y. (2021). The impact of open access mandates on invention. *Review of Economics and Statistics*, 103(5), 954–967.
- Callaway, B., & Sant’Anna, P. H. (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics*, 225(2), 200–230.
- Carlino, G., & Kerr, W. R. (2015). Agglomeration and innovation. *Handbook of Regional and Urban Economics*, 5, 349–404.
- Cengiz, D., Dube, A., Lindner, A., & Zipperer, B. (2019). The effect of minimum wages on low-wage jobs. *The Quarterly Journal of Economics*, 134(3), 1405–1454.

- Chandra, A., & Thompson, E. (2000). Does public infrastructure affect economic activity?: Evidence from the rural interstate highway system. *Regional Science and Urban Economics*, 30(4), 457–490.
- Chikis, C. A., Kleinman, B., & Prato, M. (2025). *The geography of innovative firms* (Tech. Rep.). National Bureau of Economic Research.
- Cohn, J. B., Liu, Z., & Wardlaw, M. I. (2022). Count (and count-like) data in finance. *Journal of Financial Economics*, 146(2), 529–551.
- Dai, L., Monroe, G. S., Naiker, V., & Qin, L. (2024). *Innovation spillover through information acquisition of SEC filings*. (Working paper, available at SSRN 5020404)
- Dai, R., Donohue, L., Drechsler, Q., & Jiang, W. (2023). Dissemination, publication, and impact of finance research: When novelty meets conventionality. *Review of Finance*, 27(1), 79–141.
- De Rassenfosse, G., Jensen, P. H., Julius, T., Palangkaraya, A., & Webster, E. (2023). Is the patent system an even playing field? the effect of patent attorney firms. *The Journal of Industrial Economics*, 71(1), 124–142.
- Deshpande, M., & Li, Y. (2019). Who is screened out? application costs and the targeting of disability programs. *American Economic Journal: Economic Policy*, 11(4), 213–248.
- Donaldson, D., & Hornbeck, R. (2016). Railroads and American economic growth: A “market access” approach. *The Quarterly Journal of Economics*, 131(2), 799–858.
- Du Boff, R. B. (1980). Business demand and the development of the telegraph in the United States, 1844–1860. *Business History Review*, 54(4), 459–479.
- Du Boff, R. B. (1983). The telegraph and the structure of markets in the United States, 1845–1890. *Research in Economic History*, 8, 253–277.
- Ellison, G., Glaeser, E. L., & Kerr, W. R. (2010). What causes industry agglomeration? evidence from coagglomeration patterns. *American Economic Review*, 100(3), 1195–1213.
- Faber, B. (2014). Trade integration, market size, and industrialization: evidence from

- China's National Trunk Highway System. *Review of Economic Studies*, 81(3), 1046–1070.
- Fishlow, A. (1965). *American railroads and the transformation of the ante-bellum economy*. Cambridge, MA: Harvard University Press.
- Fogel, R. W. (1964). *Railroads and American economic growth*. Johns Hopkins Press Baltimore.
- Gardner, J. (2022). Two-stage differences in differences. *arXiv preprint arXiv:2207.05943*.
- Giroud, X., Liu, E., & Mueller, H. M. (2026). Innovation spillovers across US tech clusters. *Journal of Financial Economics*, 179, 104264. doi: 10.1016/j.jfineco.2026.104264
- Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics*, 225(2), 254–277.
- Gormley, T. A., Jha, M., & Wang, M. (2026). The politicization of social responsibility. *Journal of Finance*. (Forthcoming)
- Gormley, T. A., & Matsa, D. A. (2011). Growing out of trouble? corporate responses to liability risk. *The Review of Financial Studies*, 24(8), 2781–2821.
- Gormley, T. A., & Matsa, D. A. (2016). Playing it safe? managerial preferences, risk, and agency conflicts. *Journal of Financial Economics*, 122(3), 431–455.
- Gross, D. P. (2020). Collusive investments in technological compatibility: lessons from US railroads in the late 19th century. *Management Science*, 66(12), 5683–5700.
- Grossman, G. M., & Helpman, E. (1993). *Innovation and growth in the global economy*. MIT press.
- Hegde, D., Herkenhoff, K., & Zhu, C. (2023). Patent publication and innovation. *Journal of Political Economy*, 131(7), 1845–1903.
- Helpman, E. (1998). *General purpose technologies and economic growth*. MIT press.
- Higgs, R. (1971). American inventiveness, 1870-1920. *Journal of Political Economy*, 79(3), 661–667.
- Hirshleifer, D., Low, A., & Teoh, S. H. (2012). Are overconfident CEOs better innovators?

- The Journal of Finance*, 67(4), 1457–1498.
- Hoberg, G., & Phillips, G. (2016). Text-based network industries and endogenous product differentiation. *Journal of Political Economy*, 124(5), 1423–1465.
- Hombert, J., & Matray, A. (2017). The real effects of lending relationships on innovative firms and inventor mobility. *The Review of Financial Studies*, 30(7), 2413–2445.
- Hornbeck, R. (2010). Barbed wire: Property rights and agricultural development. *The Quarterly Journal of Economics*, 125(2), 767–810.
- Hornbeck, R., & Rotemberg, M. (2024). Growth off the rails: Aggregate productivity growth in distorted economies. *Journal of Political Economy*, 132(11), 3547–3602.
- Hunt, J., Cockburn, I. M., & Bessen, J. (2024). *Is distance from innovation a barrier to the adoption of artificial intelligence?* (Tech. Rep.). National Bureau of Economic Research.
- 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.
- John, R. R. (1995). *Spreading the news: The American postal system from Franklin to Morse*. Harvard University Press.
- John, R. R. (2010). *Network nation: Inventing American telecommunications*. Cambridge, MA: Harvard University Press.
- John, R. R. (2015). *The postal age: The emergence of modern communications in nineteenth-century America*. Chicago: University of Chicago Press.
- Jones, C. I., & Vollrath, D. (2013). *Introduction to economic growth* (3rd ed.). New York: W. W. Norton & Company.
- Kelly, B., Papanikolaou, D., Seru, A., & Taddy, M. (2021). Measuring technological innovation over the long run. *American Economic Review: Insights*, 3(3), 303–320.
- Kennedy, J. (1957). Development of postal rates: 1845–1955. *Land Economics*, 33(2), 93–112.

- Kerr, W. R. (2008). Ethnic scientific communities and international technology diffusion. *The Review of Economics and Statistics*, 90(3), 518–537.
- Khan, B. Z. (2005). *The democratization of invention: patents and copyrights in American economic development, 1790-1920*. Cambridge University Press.
- Kielbowicz, R. B. (2002). *Universal postal service: A policy history, 1790-1970* (Tech. Rep.). Postal Rate Commission. Retrieved from https://prc.gov/sites/default/files/papers/paper_1.pdf
- Lamoreaux, N. R., & Sokoloff, K. L. (2001). Market trade in patents and the rise of a class of specialized inventors in the 19th century United States. *American Economic Review*, 91(2), 39–44.
- Manson, S., Schroeder, J., Van Riper, D., Knowles, K., Kugler, T., Roberts, F., & Ruggles, S. (2024). *IPUMS National Historical Geographic Information System: Version 19.0*. Minneapolis, MN: IPUMS. (Dataset) doi: 10.18128/D050.V19.0
- Matray, A. (2021). The local innovation spillovers of listed firms. *Journal of Financial Economics*, 141(2), 395–412.
- Michaels, G. (2008). The effect of trade on the demand for skill: Evidence from the interstate highway system. *The Review of Economics and Statistics*, 90(4), 683–701.
- Mokyr, J. (2002). *The gifts of Athena: Historical origins of the knowledge economy*. Princeton, NJ: Princeton University Press.
- Nicholas, T. (2010). The role of independent invention in US technological development, 1880–1930. *The Journal of Economic History*, 70(1), 57–82.
- Nonnenmacher, T. (2001). *History of the U.S. telegraph industry*. EH.Net Encyclopedia, edited by Robert Whaples. Retrieved from <https://eh.net/encyclopedia/history-of-the-u-s-telegraph-industry/>
- Perlman, E. R. (2015). *Dense enough to be brilliant: Patents, urbanization, and transportation in nineteenth century America* (CEH Discussion Paper No. 2015-036). Centre for Economic History, Australian National University.

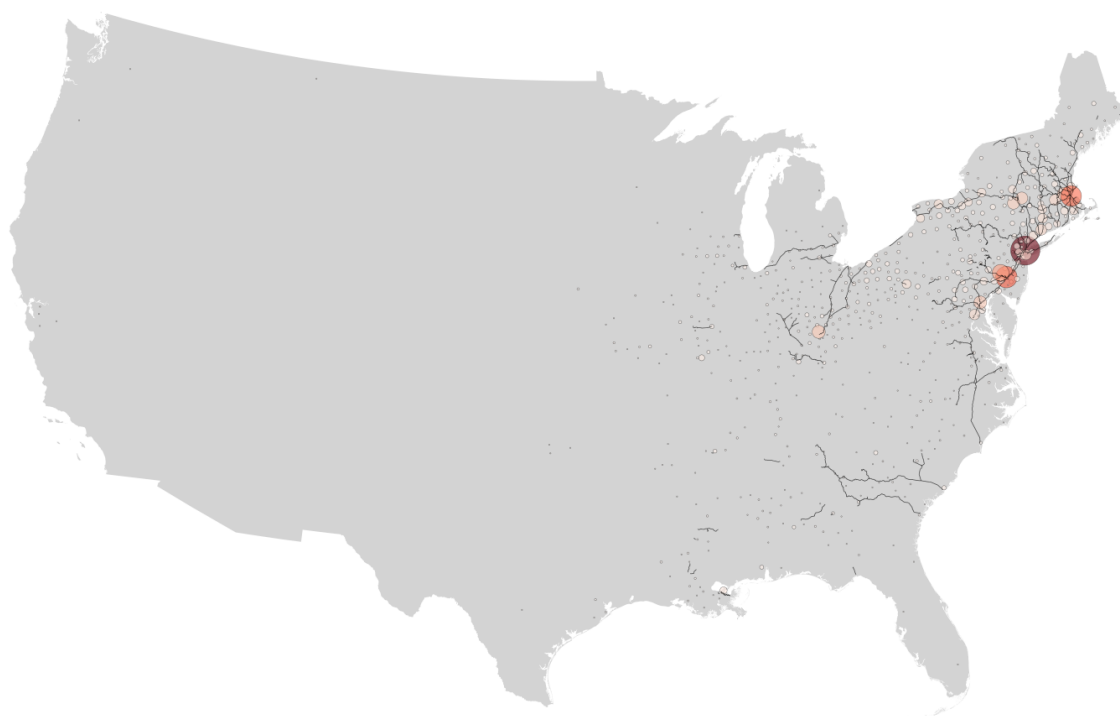
- Puffert, D. J. (2002). Path dependence in spatial networks: the standardization of railway track gauge. *Explorations in Economic History*, 39(3), 282–314.
- Puffert, D. J. (2009). *Tracks across continents, paths through history: the economic dynamics of standardization in railway gauge*. University of Chicago Press.
- Redding, S. J., & Turner, M. A. (2015). Transportation costs and the spatial organization of economic activity. *Handbook of Regional and Urban Economics*, 5, 1339–1398.
- Romer, P. M. (1990). Endogenous technological change. *Journal of Political Economy*, 98(5, Part 2), S71–S102.
- Sandvik, J. J., Saouma, R. E., Seegert, N. T., & Stanton, C. T. (2020). Workplace knowledge flows. *The Quarterly Journal of Economics*, 135(3), 1635–1680.
- Schumpeter, J. A. (1950). *Capitalism, socialism and democracy* (3rd ed.). New York: Harper Collins. (Reprinted 2008)
- Seru, A. (2014). Firm boundaries matter: Evidence from conglomerates and R&D activity. *Journal of Financial Economics*, 111(2), 381–405.
- Sokoloff, K. L., & Engerman, S. L. (2000). History lessons: institutions, factor endowments, and paths of development in the New World. *Journal of Economic Perspectives*, 14(3), 217–232.
- Sun, L., & Abraham, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*, 225(2), 175–199.
- Taylor, G. R., & Neu, I. D. (1956). *The American railroad network, 1861–1890*. Cambridge, MA: Harvard University Press.
- Wang, T. (2025). The electric telegraph, news coverage, and political participation. *The Journal of Economic History*, 85(1), 1–32. doi: 10.1017/S0022050725000026
- Weitzman, M. L. (1998). Recombinant growth. *The Quarterly Journal of Economics*, 113(2), 331–360.
- Zacchia, P. (2020). Knowledge spillovers through networks of scientists. *The Review of Economic Studies*, 87(4), 1989–2018.

Zhang, D., Li, J., Zeng, Z., & Wang, F. (2025). Jasper and Stella: distillation of SOTA embedding models. *arXiv preprint arXiv:2412.19048*.

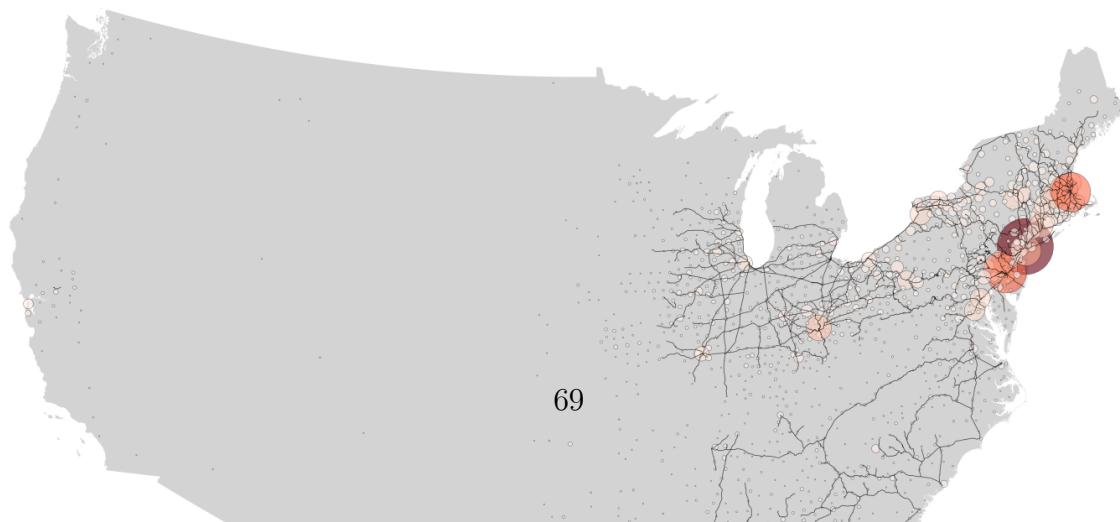
Figure 1: **Geographic Distribution of Patent and Railroads, by Decade**

These maps present the spatial distribution of patents published in each county and the railroad lines from 1840-1900, by decades. The maps include the 2,797 counties, and county boundaries correspond to county boundaries in 1890. In each Panel, the size of the circles represents the number of patents published by each county during the period of 1841-1850 (Panel A), 1851-1860 (Panel B), 1861-1870 (Panel C), 1871-1880 (Panel D), 1881-1890 (Panel E), and 1891-1900 (Panel F). Darker shades denote greater number of patents. The black lines represent constructed railroads completed by 1850 (Panel A), 1860 (Panel B), 1870 (Panel C), 1880 (Panel D), 1890 (Panel E), 1900 (Panel F). The size of the circles represents the number of patents published during each decade by county.

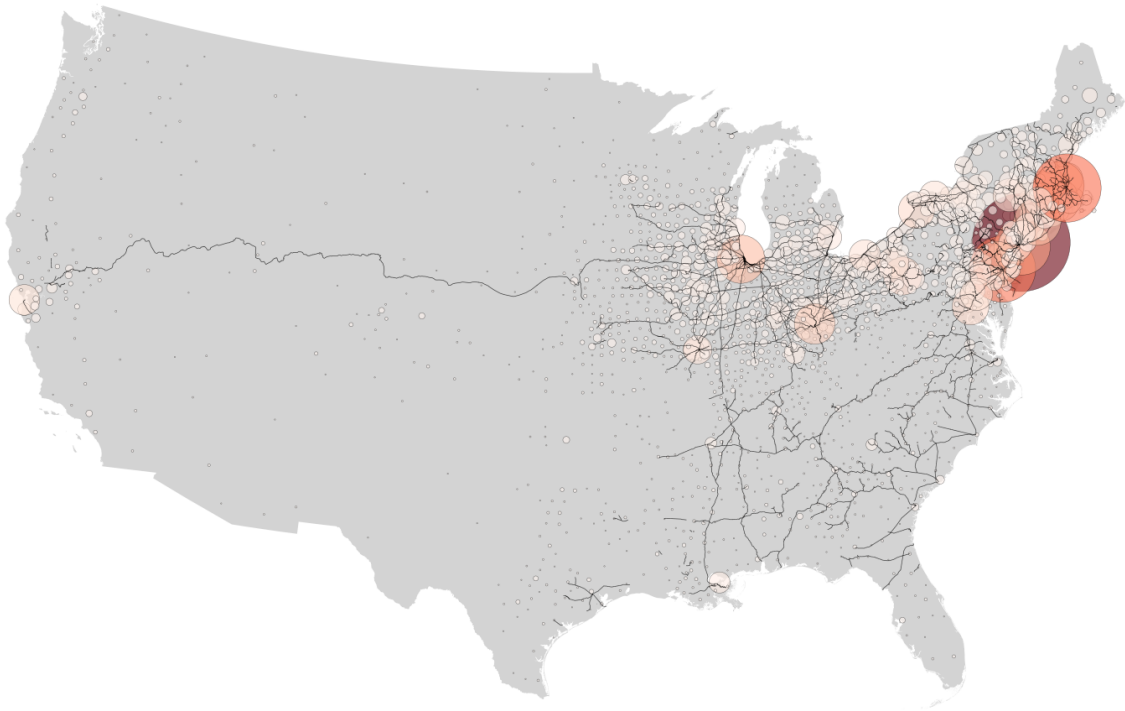
Panel A: 1841–1850



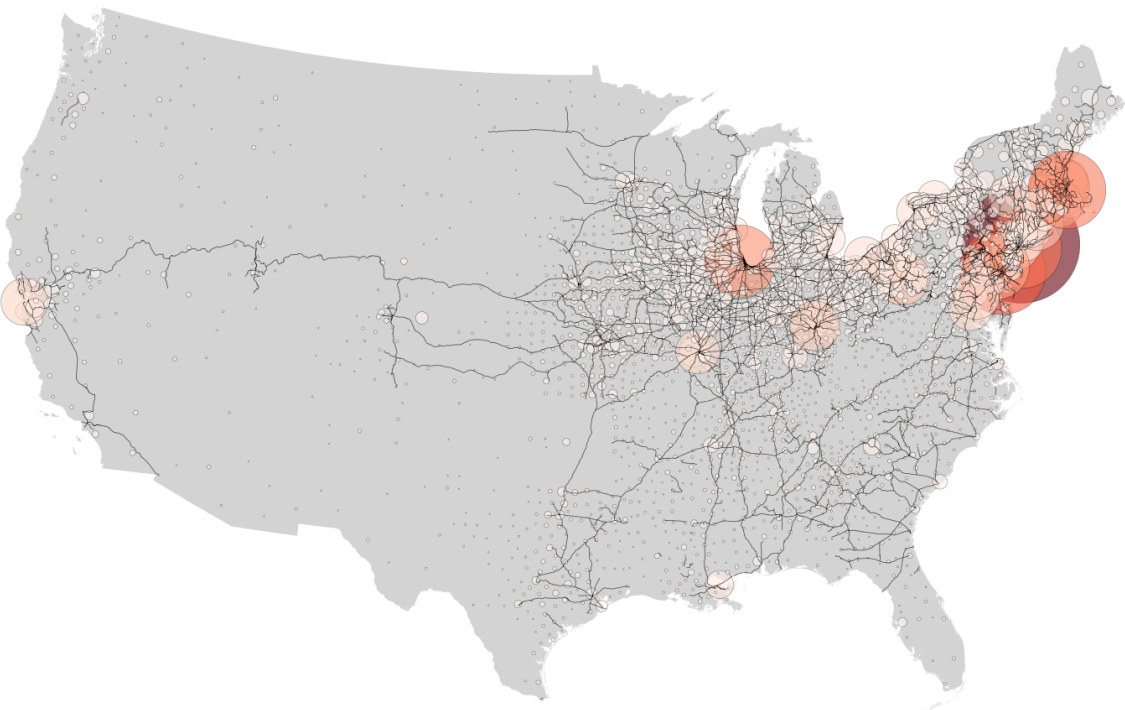
Panel B: 1851–1860



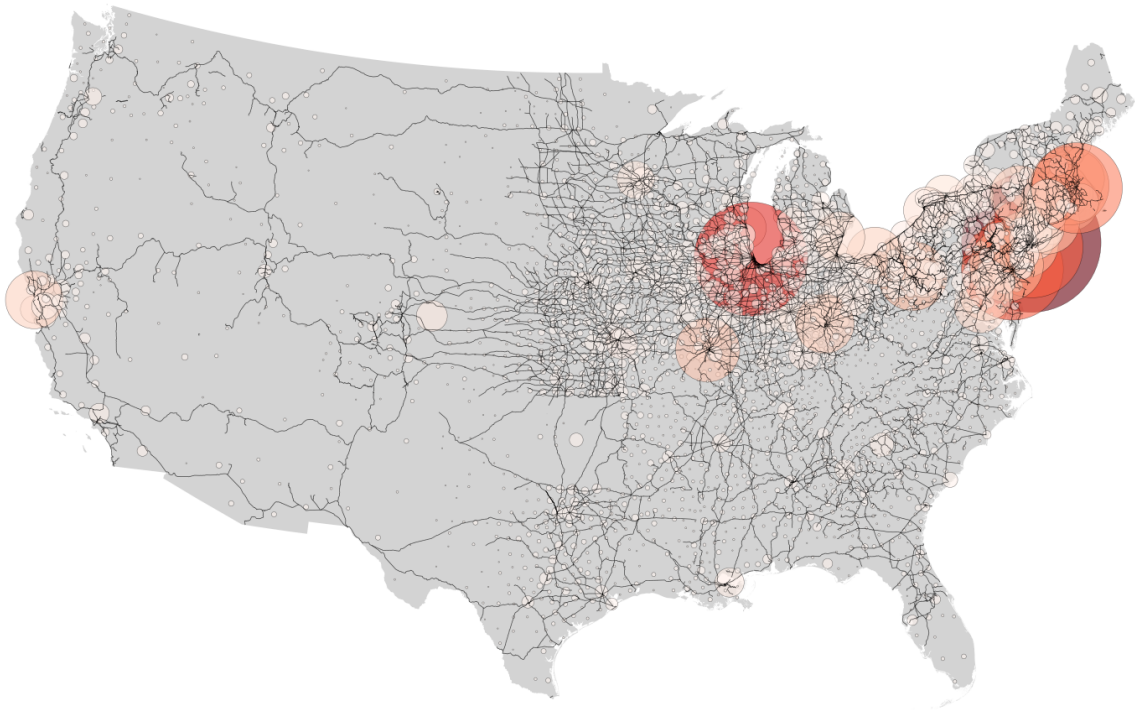
Panel C: 1861–1870



Panel D: 1871–1880



Panel E: 1881–1890



Panel F: 1891–1900

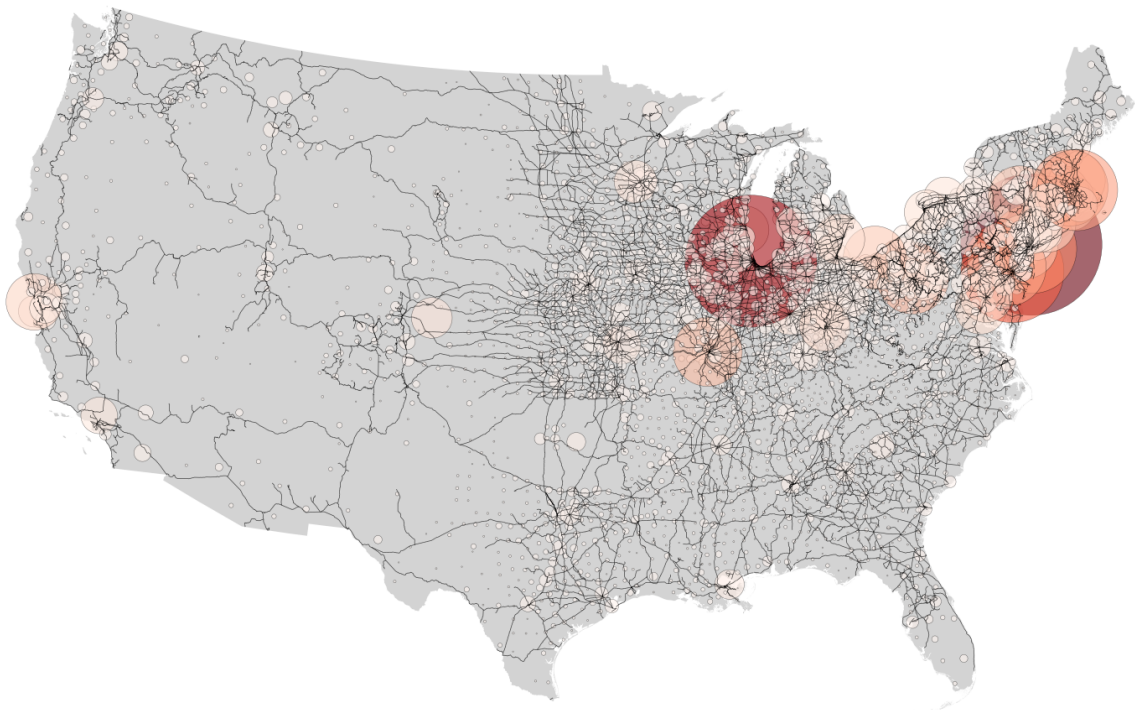


Figure 2: **Treatment Status of Counties by Railroad-Arrival Cohort**

This figure shows the composition of the stacked difference-in-differences sample by treatment cohort prior to propensity score matching. The y -axis reports the treatment cohort, where a cohort is defined by its treatment year, the calendar year in which a county first receives railroad access. Each horizontal bar partitions all counties observed within the cohort's eleven-year event window into four mutually exclusive groups based on their treatment status relative to the cohort year t : counties that already had an operating railroad before $t - 5$, counties first treated in year t , counties that are not yet treated and first receive a railroad after $t + 5$, and never-treated counties that do not receive a railroad during the sample period. The length of each segment represents the number of distinct counties in the corresponding group, while the value displayed at the right end of each bar indicates the total number of counties observed in the cohort's event window. The clean control group is drawn from never-treated counties and counties that remain untreated throughout the entire event window ending in year $t + 5$.

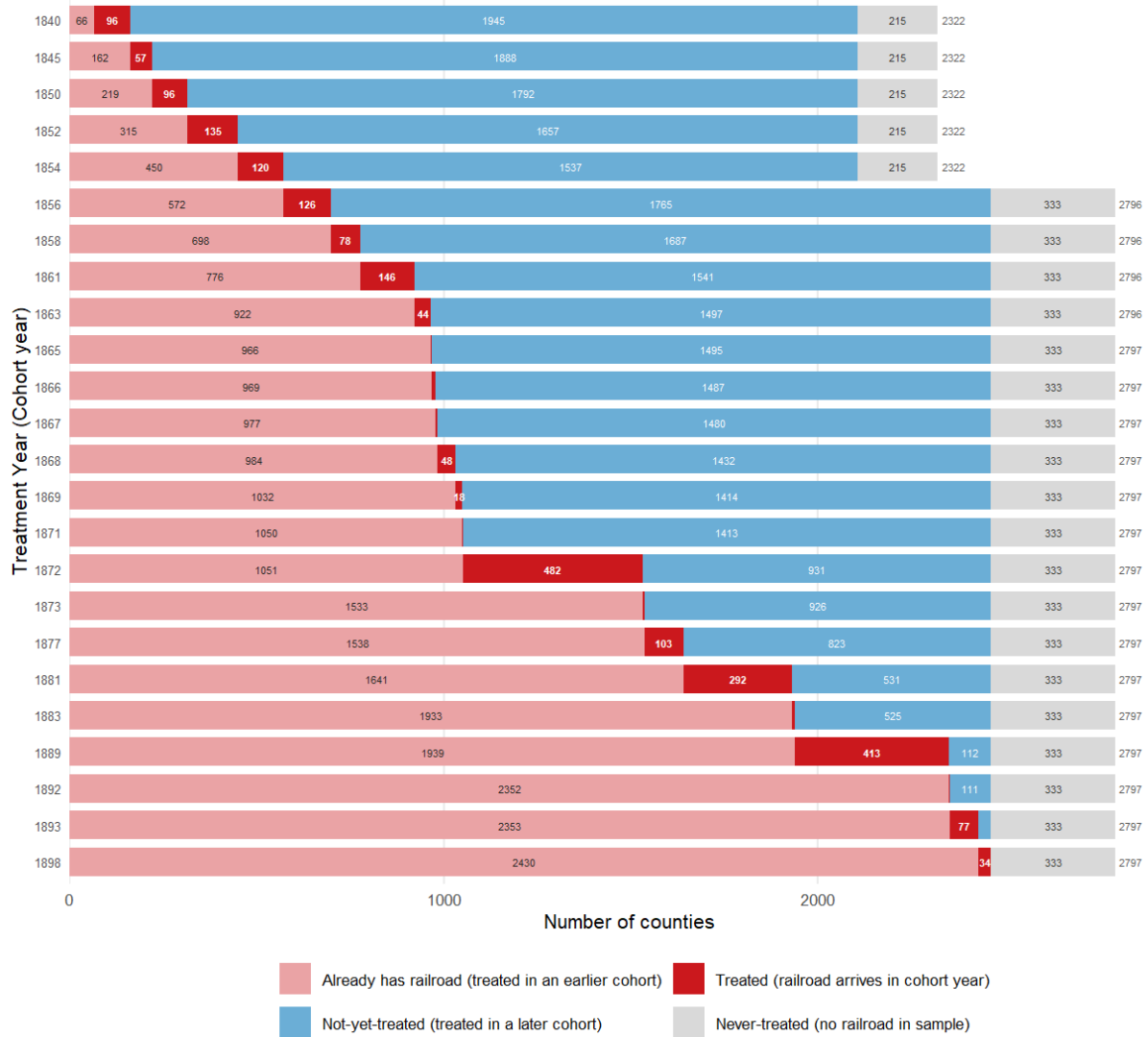
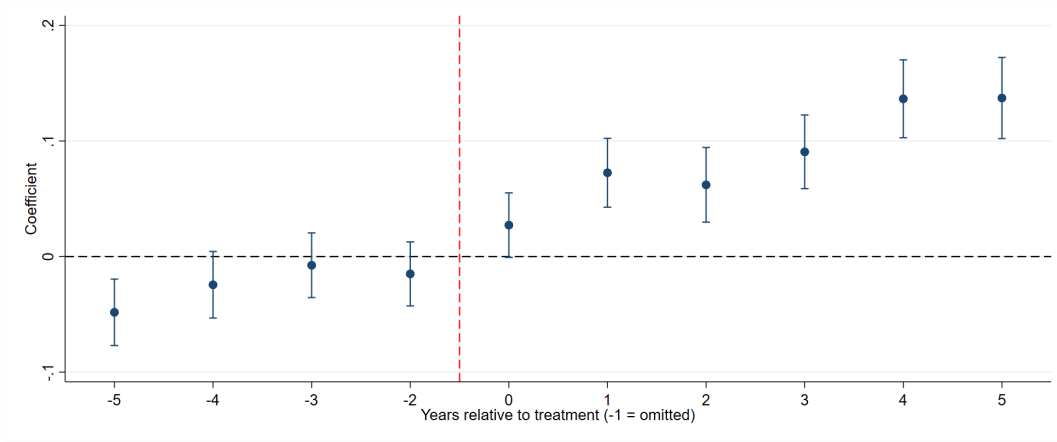


Figure 3: **Dynamics Treatment effect on Innovation Quantity**

This figure plots the dynamic (event-study) effect of first railroad completion on county patenting using the stacked difference-in-differences matched sample, with treatment years spanning 1840 to 1898. The treatment year is defined as the first year in which a county has a railroad in operation. For each treatment cohort, I construct an event window covering five years before and five years after treatment and stack these cohorts into a panel aligned by relative treatment time. Within each cohort, treated counties are those whose first railroad began operating in the treatment year, while matched control counties are selected from counties with no railroad construction during the cohort's eleven-year event window. Matching is conducted using one-to-one nearest-neighbor matching without replacement based on propensity scores estimated from a logit model of treatment status on log population, urban share, and white share measured in the year prior to treatment. Each point represents the coefficient on the interaction between the treatment indicator and an event-time indicator, estimated by OLS with $\ln(1+Patents)$, defined as the natural logarithm of one plus the county-year patent count, as the dependent variable. The specification controls for event-time indicators and the covariates included in the baseline model. The omitted reference period is the year immediately preceding treatment $t - 1$. Panel A includes county, year, and cohort fixed effects. Panel B additionally includes cohort-by-year and cohort-by-county fixed effects. Standard errors are clustered at the county level, and the vertical bars denote 95% confidence intervals.

Panel A: County, year, and cohort FE



Panel B: Adds cohort $\times$ year and cohort $\times$ county FE

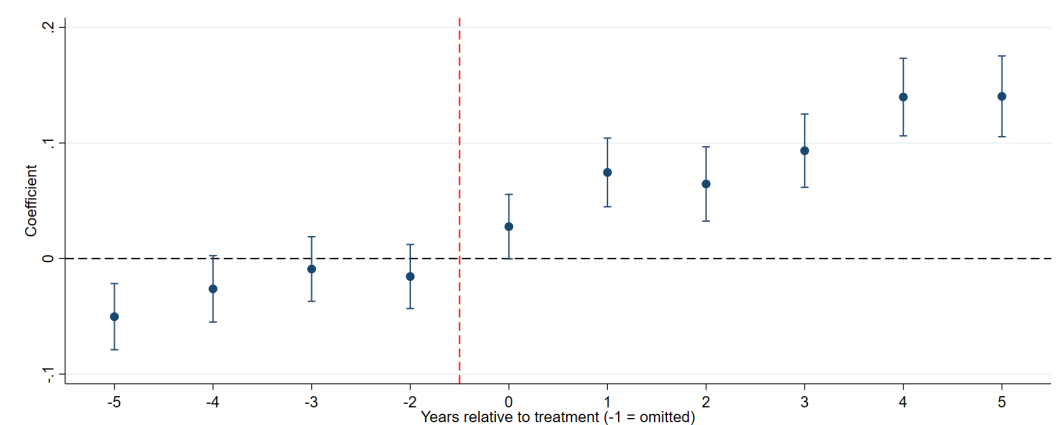
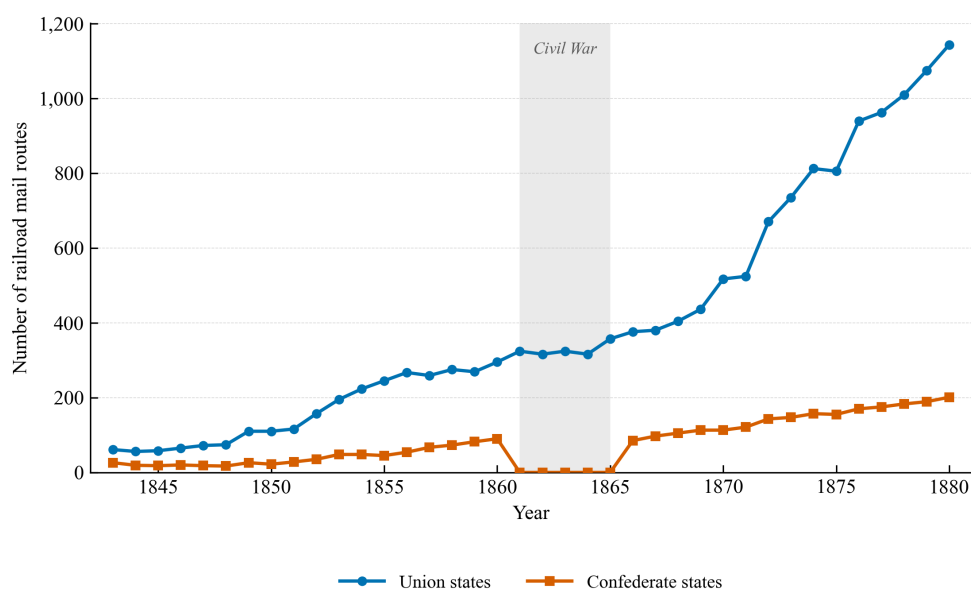


Figure 4: **Railroad mail routes in Union and Confederate states, 1843–1880**

This figure illustrates the evolution of U.S. railroad mail service separately for the eleven Confederate states (Alabama, Arkansas, Florida, Georgia, Louisiana, Mississippi, North Carolina, South Carolina, Tennessee, Texas, and Virginia) and for all remaining states and territories (hereafter, “Union states”). The upper panel reports the number of railroad mail routes in operation each year, while the lower panel reports the corresponding total route mileage. The shaded area denotes the Civil War period from 1861 to 1865. Railroad mail service in the Confederate states falls to zero during the war because the U.S. Post Office Department suspended operations in the seceded states; service resumed in 1866. West Virginia and the border states of Delaware, Kentucky, Maryland, and Missouri are classified as Union states throughout the sample period. The unit of observation is a route-year. Source: *Annual Report of the Postmaster General*, 1843–1880.

Panel A: Number of railroad mail routes



Panel B: Total railroad mail route mileage

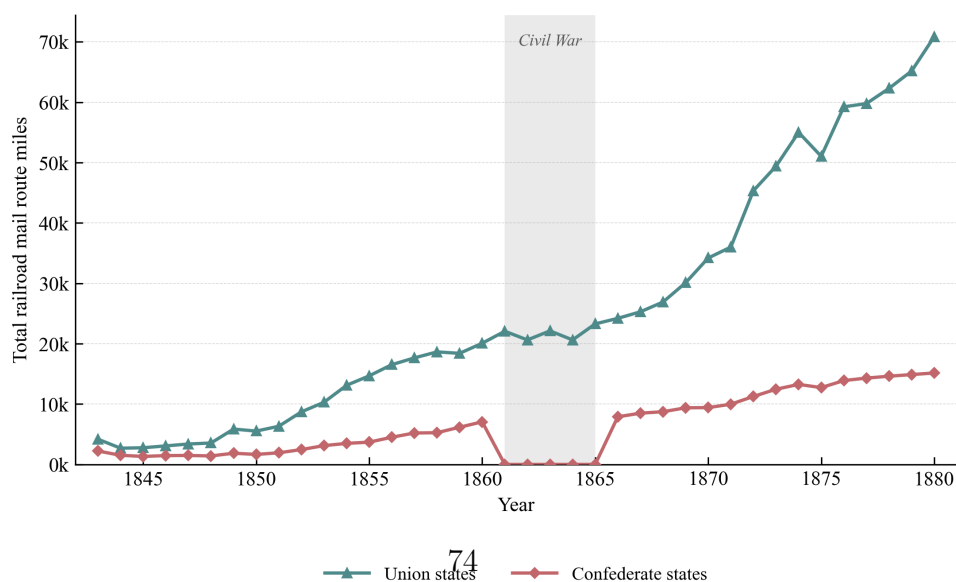
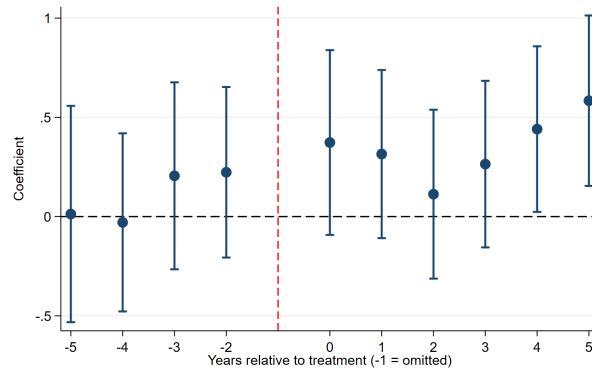


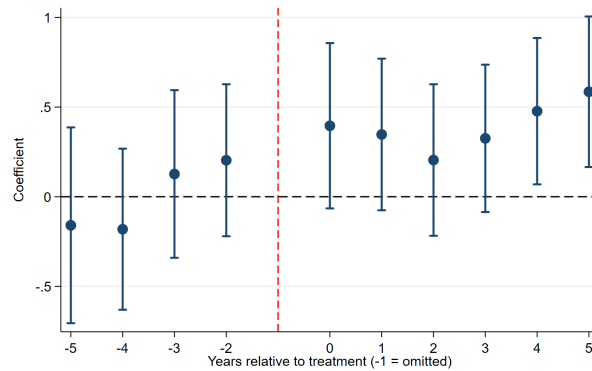
Figure 5: **Dynamic effects of mail-route connections on county-pair patent similarity**

This figure plots estimates of the dynamic effect of a mail-route connection on the patent similarity between a county pair. The plotted coefficients are the interactions of $Connected_{f,p,g}$ with indicators for each year relative to the connection year $\tau = 0$, estimated from a dynamic version of Equation (7); the year before connection ($\tau = -1$, marked by the red dashed line) is the omitted reference. The dependent variable is the within-year percentile rank of $Similarity[t - 10, t - 3]$, the similarity between focal county f 's patents in year t and partner county p 's patents in years $[t - 10, t - 3]$, so the vertical axis is in percentile points. Markers are point estimates and capped bars are 95% confidence intervals based on standard errors clustered by focal county. The sample consists of matched directed county pairs in cohorts with connection years between 1848 and 1875. Panel A includes focal $\times$ year, partner $\times$ year, and cohort fixed effects; panels B and C replace the cohort fixed effects with directed-pair $\times$ cohort fixed effects, and panel C further excludes nearby pairs whose centroids lie within 35 miles of each other. Across all three specifications the pre-connection coefficients are individually insignificant and jointly indistinguishable from zero.

Panel A: Baseline fixed effects



Panel B: Directed-pair $\times$ cohort fixed effects



Panel C: Nearby pairs excluded

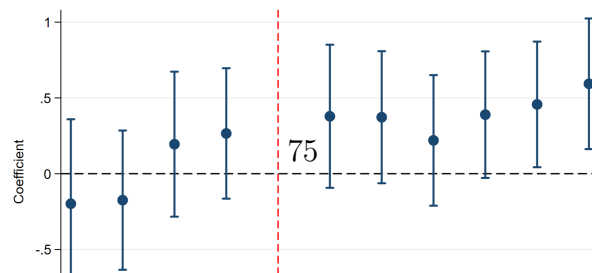


Table 1: **Summary Statistics and Covariate Balance**

This table reports covariate balance and summary statistics for the matched stacked difference-in-differences estimation sample with treatment years span 1840–1898. Treated counties are those whose first railroad began operating in the treatment year; potential controls are counties with no railroad built during that cohort’s eleven-year event window. Within each cohort I match one-to-one and without replacement on the propensity score, estimated from a logit of treatment status on log population, urban share, and white share. Census variables are interpolated and area-weighted to constant 1890 county boundaries. Panel A reports covariate balance between treated and control counties before and after matching, measured in the baseline year of each cohort (the year before treatment, $t - 1$). Asterisks denote a two-sample t -test of equal means: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$. Panel B reports summary statistics for the matched stacked difference-in-differences estimation sample, deduplicated to unique county-years. P10, Median, and P90 are the 10th, 50th, and 90th percentiles. *Trips* is the total weekly trips on the rail mail routes serving the county, and *Routes* is the number of distinct rail mail routes; both are observed over 1843–1880 (the mail-route data period). The reported percentiles are unconditional across the estimation sample; among treated county-years in the post-treatment period, the 90th percentile of *Trips* is about 20 weekly trips. *Degree centrality* and *Betweenness centrality* measure the county’s position in the railroad network, which spans the full 1840–1898 sample.

Panel A: Covariate balance

	Before matching			After matching		
	Treated	Control	Difference	Treated	Control	Difference
Ln(Population)	8.705	6.507	2.197***	8.748	8.698	0.050
Urban share	0.030	0.016	0.015***	0.020	0.017	0.004
White share	0.855	0.844	0.011**	0.850	0.852	−0.002
Counties (treated / control)	2,394 / 31,075			2,122 / 2,122		

Panel B: Summary statistics

	Mean	SD	P10	Median	P90	N
Patents	0.596	1.448	0.000	0.000	2.000	41,826
Population	9,976	8,924	1,312	8,201	19,899	41,826
Urban share	0.022	0.208	0.000	0.000	0.000	41,600
White share	0.846	0.203	0.507	0.950	0.999	41,600
Distance to river (km)	101.9	140.0	8.2	50.6	272.5	41,826
Distance to canal (km)	526.7	503.8	43.5	384.8	1,214.4	41,826
Trips	2.162	6.551	0.000	0.000	7.000	28,550
Routes	0.288	0.764	0.000	0.000	1.000	28,550
Degree centrality ($\times 10^3$)	0.278	0.674	0.000	0.000	0.715	41,826
Betweenness centrality ($\times 10^3$)	0.498	2.783	0.000	0.000	0.592	41,826

Table 2: **Treatment Effect on Innovation Quantity**

This table reports the effect of first railroad completion on patent quantity using the stacked difference-in-differences matched sample. I define the first year in which a county has a railroad in operation as its treatment year, with treatment years spanning 1845–1898. Around each treatment year I construct a cohort with a time window of five years before through five years after treatment, and I stack these cohorts into a single panel aligned by relative treatment timing. Within each cohort, treated counties are those that received their first railroad in the treatment year, while matched control counties are selected from those with no railroad built during the 11-year event window using one-to-one nearest-neighbor matching. The matching variables include population, urban share, and white share. The dependent variable, *Patents*, is the total number of the patents generated within each county-year, entered either in levels or as $\ln(1 + \textit{Patents})$, the natural logarithm of one plus this count. The independent variable *Treat* is an indicator equal to one if the county had its first railroad built in that cohort. *Post* is an indicator equal to one for the years after the event year. Standard errors are clustered at the county level, and *t*-statistics are reported in parentheses below the coefficients. Coefficients marked with ***, **, and * are significant at the 1%, 5%, and 10% level, respectively.

	(1) $\ln(1 + \textit{Patents})$	(2) <i>Patents</i>	(3) $\ln(1 + \textit{Patents})$	(4) <i>Patents</i>
Treat $\times$ Post	0.107*** (11.00)	0.207*** (4.80)	0.110*** (11.46)	0.219*** (4.98)
Treat	0.097*** (6.98)	0.027 (0.38)		
Post	-0.058*** (-6.25)	-0.085* (-1.94)		
Ln(Population)	0.069*** (11.61)	0.379*** (6.56)	0.052*** (9.22)	0.295*** (6.60)
Urban share	0.052 (0.88)	0.163 (1.57)	0.026 (0.68)	0.149** (2.00)
White share	-0.438*** (-8.93)	-3.139*** (-5.24)	-0.365*** (-7.48)	-3.195*** (-5.97)
Ln(Distance to river)	-0.006 (-0.28)	-0.243* (-1.78)	-0.032* (-1.85)	-0.244** (-2.48)
Ln(Distance to canal)	-0.002 (-0.20)	0.071** (2.00)	0.012 (0.81)	0.063** (2.13)
County, Year, Cohort FE	Y	Y	Y	Y
Cohort-Year FE			Y	Y
Cohort-County FE			Y	Y
Estimation Method	OLS	Poisson	OLS	Poisson
Observations	46,684	42,294	46,684	38,631
<i>R</i> -squared	0.452		0.511	

Table 3: **Treatment Effect on Innovation Quality**

This table reports the effect of first railroad completion on patent quality using the stacked difference-in-differences matched sample over the period 1840–1900. I define the year in which a county first receives a railroad as its treatment year. Around each treatment year I construct a cohort spanning the five years before and the five years after treatment, and these cohorts are stacked to form a panel aligned by relative treatment timing. Within each cohort, treated counties are those that received their first railroad in the treatment year, while matched control counties are selected, by one-to-one nearest-neighbor matching, from counties with no railroad built during the 11-year event window. The matching variables are population, urban share, and white share. The dependent variables are patent textual-similarity measures, averaged across the patents granted in each county-year, and the table reports three quality concepts: importance, novelty, and impact. In columns (1) and (2), $FS5/BS5$ is the ratio of the five-year forward similarity $FS5$ to the five-year backward similarity $BS5$. In columns (3) and (4), $FS10/BS5$ is the ratio of the ten-year forward similarity to the five-year backward similarity. In columns (5) and (6), *Novelty* is the inverse of the five-year backward similarity, $1/BS5$, so that higher values indicate patents less similar to prior art (i.e., more novel). In columns (7) and (8), *Impact* is the five-year forward similarity, $FS5$, capturing the extent to which subsequent patents resemble the focal county-year’s patents. The independent variable *Treat* is an indicator equal to one if the county had its first railroad built in that cohort, and *Post* is an indicator equal to one for the years on or after the event year. All specifications control for log population, the urban and white population shares, the log distances to the nearest river and canal, and log manufacturing capital. Standard errors are clustered at the county level, and *t*-statistics are reported in parentheses below the coefficients. Coefficients marked with ***, **, and * are significant at the 1%, 5%, and 10% level, respectively.

	Importance		Importance		Novelty		Impact	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	$\ln(\frac{FS5}{BS5})$	$FS5/BS5$	$\ln(\frac{FS10}{BS5})$	$FS10/BS5$	$\ln(1/BS5)$	$1/BS5$	$\ln(FS5)$	$FS5$
Treat×Post	-0.010 (-1.54)	-0.010 (-1.13)	-0.012 (-1.51)	-0.009 (-0.86)	-0.033 (-1.10)	-0.043 (-0.50)	0.022 (0.76)	0.005 (0.16)
Controls	Y	Y	Y	Y	Y	Y	Y	Y
County, Year, Cohort FE	Y	Y	Y	Y	Y	Y	Y	Y
Cohort-Year FE	Y	Y	Y	Y	Y	Y	Y	Y
Cohort-County FE	Y	Y	Y	Y	Y	Y	Y	Y
Estimation Method	OLS	Poisson	OLS	Poisson	OLS	Poisson	OLS	Poisson
Observations	11,754	11,754	11,754	11,754	11,754	11,754	11,754	11,754
R-squared	0.847		0.752		0.814		0.752	

Table 4: **Treatment Intensity: Mail Service and Railroad-Network Position**

This table reports treatment-intensity tests for the effect of first railroad access on patent quantity, using the stacked difference-in-differences matched sample. Both panels ask whether the innovation gains were larger for counties that were more intensively or more centrally connected, by interacting $Treat \times Post$ with a continuous connection measure. Panel A measures the intensity of rail *mail service* over the 1843–1880 period spanned by the mail-route data: *Trips* is the total weekly mail trips on the routes serving a county (the intensive margin) and *Routes* is the number of distinct mail routes (the extensive margin). Panel B measures a county’s *position in the railroad network*, which is defined for the full 1840–1898 sample: *Degree* is the number of direct railroad connections and *Betweenness* is the county’s betweenness centrality, each entered as its within-year percentile rank among connected counties and *frozen at the connection year*, so that the dose is fixed before the post-treatment period and cannot respond to subsequent network growth. All specifications control for log population, the urban and white population shares, the log distances to the nearest river and canal, log manufacturing capital, and log steamboat mail frequency. Standard errors are clustered at the county level, and *t*-statistics are reported in parentheses below the coefficients. Coefficients marked with ***, **, and * are significant at the 1%, 5%, and 10% level, respectively.

	(1) ln (1 + Patents)	(2) Patents	(3) ln (1 + Patents)	(4) Patents
Panel A: Mail service (1843–1880)				
	Trips		Routes	
Treat $\times$ Post $\times$ Service	0.004*** (5.04)	0.004** (2.16)	0.048*** (5.60)	0.050** (2.25)
Treat $\times$ Post	0.100*** (7.43)	0.203*** (3.04)	0.086*** (6.13)	0.178*** (2.64)
Observations	25,960	21,458	25,960	21,458
<i>R</i> -squared	0.537		0.538	
Panel B: Network position (1840–1898)				
	Degree		Betweenness	
Treat $\times$ Post $\times$ Centrality	0.067** (2.29)	0.385*** (3.89)	0.081*** (3.35)	0.299*** (3.48)
Treat $\times$ Post	0.080*** (4.82)	0.056 (0.93)	0.089*** (7.78)	0.144*** (2.96)
Observations	46,684	38,631	46,684	38,631
<i>R</i> -squared	0.511		0.511	
Controls	Y	Y	Y	Y
County, Year, Cohort FE	Y	Y	Y	Y
Cohort-Year FE	Y	Y	Y	Y
Cohort-County FE	Y	Y	Y	Y
Estimation Method	OLS	Poisson	OLS	Poisson

Table 5: **Heterogeneous Treatment Effects on Patent Quantity**

This table examines whether the effect of railroad access on patent quantity differs across counties along two dimensions, innovation-hub status (Panel A) and access to water (Panel B), using the stacked difference-in-differences matched sample over cohorts with treatment years from 1840 to 1898. In Panel A, *Hub* is a predetermined measure of a county's pre-treatment inventive stock. For each cohort, I sum the county's patents from 1836 through the year preceding the cohort's treatment year and rank that stock nationally against all populated U.S. counties, those with positive population, as of the same year; a county's rank is the fraction of that national universe whose stock it strictly exceeds. A county is classified as a hub if this pre-treatment rank lies at or above the 75th percentile, the top quartile; Appendix Table A.3 reports the same split under a stricter top-decile (90th percentile) cutoff. In Panel B, *Water* is an indicator equal to one for matched pairs whose treated county has access to water transportation, defined as intersecting a navigable river, a canal, the Atlantic coast, or the Great Lakes; each matched control county inherits the classification of its treated county. In each panel, the two triple interactions are the difference-in-differences effect within the indicated group and its complement, and *Difference* reports the gap between the two, together with its *t*-statistic in the OLS columns or *z*-statistic in the Poisson columns. Each specification also includes the corresponding lower-order interaction and level terms, whose coefficients are omitted for brevity. All specifications control for log population, the urban and white population shares, and the log distances to the nearest river and canal. Standard errors are clustered at the county level, and *t*-statistics are reported in parentheses below the coefficients. Coefficients marked with ***, **, and * are significant at the 1%, 5%, and 10% levels, respectively.

	(1) ln (1 + Patents)	(2) Patents	(3) ln (1 + Patents)	(4) Patents
Panel A: Innovation hub				
Treat×Post×Hub	0.135*** (4.23)	0.192*** (2.73)	0.137*** (4.28)	0.173** (2.28)
Treat×Post×NonHub	0.099*** (9.98)	0.236*** (4.56)	0.103*** (10.49)	0.233*** (4.46)
<i>Difference (Hub – NonHub)</i>	0.036 (1.09)	-0.044 (-0.51)	0.034 (1.02)	-0.060 (-0.65)
<i>R-squared</i>	0.457		0.511	
Panel B: Waterway access				
Treat×Post×Water	0.107*** (4.01)	0.117 (1.16)	0.108*** (4.07)	0.103 (1.02)
Treat×Post×No-water	0.106*** (10.33)	0.229*** (4.84)	0.110*** (10.81)	0.246*** (5.10)
<i>Difference (Water – No-water)</i>	0.000 (0.01)	-0.112 (-1.00)	-0.003 (-0.09)	-0.143 (-1.29)
<i>R-squared</i>	0.452		0.511	
Controls	Y	Y	Y	Y
County, Year, Cohort FE	Y	Y	Y	Y
Cohort-Year FE	80		Y	Y
Cohort-County FE			Y	Y
Estimation Method	OLS	Poisson	OLS	Poisson
Observations	46,684	42,294	46,684	38,631

Table 6: **Heterogeneous Treatment Effects on Patent Quality**

This table examines whether the effect of railroad access on patent importance differs by innovation-hub status (columns 1–2) and access to water transportation (columns 3–4), using the stacked difference-in-differences matched sample over cohorts with treatment years from 1850 to 1890. The dependent variable is $\ln(FS10/BS5)$, the natural logarithm of the ratio of ten-year forward to five-year backward patent similarity, averaged over the patents produced in each county-year. *Hub* equals one for counties whose pre-treatment national patent-stock rank lies at or above the 75th percentile, defined as in Table 5; *Water* equals one for matched pairs whose treated county intersects a navigable river, a canal, the Atlantic coast, or the Great Lakes. The two triple interactions give the difference-in-differences effect within the indicated group and its complement, and the *Difference* rows report the gap between them together with its *t*-statistic. Each specification also includes the corresponding lower-order interaction and level terms, whose coefficients are omitted for brevity. All specifications control for log population, the urban and white population shares, the log distances to the nearest river and canal, log manufacturing capital, and log steamboat mail frequency. Appendix Table A.4 reports the hub split under a stricter top-decile (90th percentile) cutoff. Standard errors are clustered at the county level, and *t*-statistics are reported in parentheses below the coefficients. Coefficients marked with ***, **, and * are significant at the 1%, 5%, and 10% levels, respectively.

	$\ln(FS10/BS5)$			
	Hub		Waterway access	
	(1)	(2)	(3)	(4)
Treat×Post×Hub	0.000 (0.01)	0.006 (0.34)		
Treat×Post×NonHub	-0.016* (-1.86)	-0.016* (-1.70)		
Treat×Post×Water			0.033* (1.66)	0.044** (2.03)
Treat×Post×No-water			-0.019** (-2.25)	-0.017* (-1.93)
<i>Difference (Hub – NonHub)</i>	0.017 (0.89)	0.021 (1.11)		
<i>Difference (Water – No-water)</i>			0.052** (2.46)	0.061*** (2.66)
Controls	Y	Y	Y	Y
County, Year, Cohort FE	Y	Y	Y	Y
Cohort-Year FE		Y		Y
Cohort-County FE		Y		Y
Estimation Method	OLS	OLS	OLS	OLS
Observations	11,271	10,969	11,271	10,969

Table 7: **The Effect of Mail-Route Connections on County-Pair Patent Similarity**

The table reports stacked difference-in-differences estimates of the effect of a mail-route connection between two counties on the text similarity of their patents. An observation is a directed county pair, consisting of a focal county and a partner county, observed in a cohort-by-year cell. The dependent variable is the within-year percentile rank (0–100) of $Similarity[t - 10, t - 3]$, defined as the count-weighted mean cosine similarity between the text embeddings of patents filed in focal county in year t and patents filed in partner county during years $t - 10$ through $t - 3$. *Connected* equals one for treated pairs, where the focal county and partner county first share a common mail route in cohort year, and zero for control pairs. *Post* equals one for years after the connected year. Within each focal county, treated partners are matched one-to-one with never-connected control partners using nearest-neighbor matching without replacement. Matching variables include logarithm of partner population, partner urban and white population shares, the logarithm of distance, and the focal–partner patent similarity one and four years before connection. The sample is restricted to cohorts 1848–1875 with 11-year window. The directed-pair-by-cohort fixed effects in columns (2) and (3) absorb *Connected*, *Distance*, and all other time-invariant pair characteristics. Column (3) further excludes nearby county pairs whose centroids are located within 35 miles of one another. Controls comprise the pair distance and $Log(\# \text{ Patent Pairs})$, the logarithm of the number of cross-county patent pairs used to construct each similarity measure. Panel B splits the effect by whether the partner county is an innovation hub—a partner in the top quartile (75th percentile) of the pre-treatment national patent-stock distribution—reporting the connection effect within each group and their *Difference (Hub – Non-hub)*, whose t -statistic is the square root of the F -statistic testing equality of the two triples; Panel B also includes the lower-order interaction terms, omitted for brevity, and the split holds at the top-decile cutoff as well (Appendix Table A.12). Standard errors are clustered by focal county, and t -statistics are reported below the coefficients in the parentheses. Coefficients marked with ***, **, and * are significant at the 1%, 5%, and 10% level, respectively.

	(1) Similarity	(2) Similarity	(3) Similarity
Panel A: Pooled			
Connected×Post	0.266** (2.26)	0.365*** (3.11)	0.354*** (2.95)
Connected	1.222*** (9.02)		
Post	-0.106 (-0.80)	-0.142 (-1.07)	-0.105 (-0.78)
Panel B: By partner hub status			
Connected×Post×Hub partner	0.092 (0.81)	0.162 (1.46)	0.142 (1.25)
Connected×Post×Non-hub partner	1.119*** (3.34)	1.259*** (3.66)	1.298*** (3.71)
<i>Difference (Hub – Non-hub)</i>	-1.027*** (-3.02)	-1.097*** (-3.18)	-1.156*** (-3.30)
Controls	Y	Y	Y
Focal × Year FE	Y	Y	Y
Partner × Year FE	Y	Y	Y
Cohort FE	82 Y		
Directed-pair × Cohort FE		Y	Y
Observations	403,849	403,819	392,868
<i>R</i> -squared	0.834	0.874	0.874

Table 8: **The Effect of the Postal Reforms on Innovation Output**

This table reports the effect of railroad access on patent quantity across the nineteenth-century postal-reform regimes, using the stacked DID matched sample. Treated counties are those that received their first railroad in the treatment year, while matched control counties are selected from those with no railroad built during the 11-year event window using one-to-one nearest-neighbor matching. The sample is restricted to cohorts with treatment years between 1840 and 1868 so that the pre-1851, 1851, and 1863 regimes are each observed within the event windows. *1851 Reform* is an indicator equal to one for the years 1851–1862, the regime established by the Postal Act of 1851, and *1863 Reform* is an indicator equal to one for the years from 1863 onward, the regime established by the Postal Act of 1863; the omitted reference is the pre-1851 regime (the years through 1850, which subsume the 1845 Act). The dependent variables are $\ln(1 + \textit{Patents})$, the natural logarithm of one plus the total number of patents generated within each county-year, and *Patents*, that raw count. The independent variable *Treat* is an indicator equal to one if the county had its first railroad built in that cohort. *Post* is an indicator equal to one for the years on or after the treatment year. Across all columns I control for *Population*, the urban share, the white share, and the log distances (in kilometers) from the county to its nearest river and nearest canal. Standard errors are clustered at the county level, and *t*-statistics are reported below the coefficients in the parentheses. Coefficients marked with ***, **, and * are significant at the 1%, 5%, and 10% level, respectively.

	(1) $\ln(1 + \textit{Patents})$	(2) <i>Patents</i>	(3) $\ln(1 + \textit{Patents})$	(4) <i>Patents</i>
Treat $\times$ Post $\times$ 1851 Reform	0.079** (2.53)	0.327* (1.92)	0.069** (2.20)	0.210 (1.24)
Treat $\times$ Post $\times$ 1863 Reform	0.200*** (2.85)	-0.051 (-0.21)	0.192*** (2.62)	-0.235 (-0.98)
Treat $\times$ Post	0.052** (2.08)	0.039 (0.28)	0.052** (2.11)	0.065 (0.47)
Treat $\times$ 1851 Reform	0.005 (0.23)	0.097 (0.68)	0.026 (1.07)	0.252* (1.68)
Treat $\times$ 1863 Reform	-0.155** (-2.17)	0.190 (0.67)	-0.119 (-1.59)	0.673** (2.45)
Post $\times$ 1851 Reform	-0.028 (-1.08)	0.011 (0.07)		
Post $\times$ 1863 Reform	-0.083 (-1.48)	0.193 (0.71)		
Treat	0.071* (1.96)	-0.181 (-1.12)		
Post	-0.033 (-1.41)	-0.105 (-0.79)		
Controls	Y	Y	Y	Y
County, Year, Cohort FE	Y	Y	Y	Y
Cohort-Year FE			Y	Y
Cohort-County FE			Y	Y
Estimation Method	OLS	Poisson	OLS	Poisson
Observations	43,550	39,555	43,550	36,074
<i>R</i> -squared	0.497	83	0.553	

Table 9: **The Effect of the Civil War Mail-Route Discontinuation on Patent Quantity**

This table reports the effect of the wartime suspension of Southern mail routes on patenting activity using the full county-year panel from 1856 to 1865. *South* is an indicator equal to one for counties located in the eleven Confederate states. *War* is an indicator equal to one for the wartime period, 1861–1865, and zero for the prewar period, 1856–1860. *PreWarRoute* is an indicator equal to one if a county was served by at least one railroad mail route before the war. *SouthLostRoute* is an indicator equal to one if a county is located in the Confederacy, had a railroad mail route before the war, and had no railroad mail route during the war. $\ln(1 + \textit{Patents})$ is the natural logarithm of one plus the total number of patents generated in a county-year. Columns (1) and (2) use the full sample of counties, whereas column (3) restricts the sample to counties that had railroad mail service before the war. This restriction allows the comparison to be made between Southern counties that lost mail service and Northern counties that retained it. All specifications control for $\ln(\textit{Population})$, the urban share, the white share, $\ln(\textit{ManufacturingCapital})$, $\ln(\textit{FarmValue})$, $\ln(\textit{SteamboatMailFrequency})$, and the logarithmic distances to the nearest river and canal. The time-invariant indicators *South*, *PreWarRoute*, and *SouthLostRoute* are absorbed by the county fixed effects, while *War* is absorbed by the year fixed effects. Standard errors are clustered at the county level. *t*-statistics, or *z*-statistics in the Poisson specification, are reported in parentheses. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

	(1) $\ln(1 + \textit{Patents})$	(2) <i>Patents</i>	(3) $\ln(1 + \textit{Patents})$	(4) $\ln(1 + \textit{Patents})$
SouthLostRoute×War	-0.271*** (-11.64)	-1.711*** (-8.62)	-0.352*** (-10.59)	
PreWarRoute×South×War				-0.262*** (-8.44)
PreWarRoute×War				0.072*** (3.56)
South×War				-0.119*** (-11.18)
Controls	Y	Y	Y	Y
County FE	Y	Y	Y	Y
Year FE	Y	Y	Y	Y
Estimation Method	OLS	Poisson	OLS	OLS
Observations	24,232	16,760	7,254	24,232
<i>R</i> -squared	0.815		0.841	0.817

Table 10: **The Effect of Railroad Access on Patent Quantity: Incidental Connection**

This table examines whether the estimated effect of railroad access on patent quantity is robust to the potentially endogenous placement of railroad mail routes, using the stacked difference-in-differences matched sample over the mail-route period (railroad cohorts with treatment years from 1848 to 1875). I re-estimate the baseline difference-in-differences specification after excluding treated counties that served as mail-route endpoints. A county is classified as an endpoint if it was the origin or destination of a mail route in any year within its eleven-year event window. Columns (1) and (2) use the full matched sample, while columns (3) and (4) exclude the treated endpoint counties, leaving only pass-through counties among the treated observations. The dependent variable is $\ln(1 + \textit{Patents})$, the natural logarithm of one plus the total number of patents generated within a county-year, estimated by OLS in columns (1) and (3), and the patent count, estimated by Poisson, in columns (2) and (4). *Treat* is an indicator equal to one for counties that gained railroad access in a given cohort and zero for their matched controls, while *Post* equals one for event years on or after the treatment year. All specifications control for $\ln(\text{Population})$, the urban share, the white share, $\ln(\text{Manufacturing Capital})$, $\ln(\text{Steamboat Mail Frequency})$, and the logarithmic distances to the nearest river and canal. Standard errors are clustered at the county level, and *t*-statistics are reported in parentheses. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

	(1) $\ln(1 + \textit{Patents})$	(2) <i>Patents</i>	(3) $\ln(1 + \textit{Patents})$	(4) <i>Patents</i>
<i>Treat</i> × <i>Post</i>	0.133*** (10.34)	0.238*** (3.62)	0.092*** (6.54)	0.273*** (3.62)
Controls	Y	Y	Y	Y
Year FE	Y	Y	Y	Y
Cohort FE	Y	Y	Y	Y
Cohort-Year FE	Y	Y	Y	Y
Cohort-County FE	Y	Y	Y	Y
Estimation Method	OLS	Poisson	OLS	Poisson
Observations	25,960	21,458	21,241	17,003
<i>R</i> -squared	0.536		0.500	

Table 11: **Spatial Knowledge Spillovers to Neighboring Control Counties**

This table tests whether knowledge spillovers from treated counties to bordering control counties contaminate the control group and bias the baseline difference-in-differences estimates. Adjacency is defined by a shared border, computed from 1890 county boundary polygons. The sample comprises railroad-adoption cohorts from 1840 to 1898, each observed within a ± 5 -year window around its first year of railroad operation. Columns (1)–(2) report an exclusion test: every adjacent control county is dropped from the sample, and the model is re-estimated on the remaining treated and non-adjacent control counties. Columns (3)–(4) report a direct-measurement test on the full sample, where *AdjControl* equals one for a control county that borders any treated county of the same cohort. The dependent variable is $\ln(1 + \textit{Patents})$, estimated by OLS, or the patent count, estimated by Poisson. All specifications control for log population, the urban and white population shares, and the log distances to the nearest river and canal. Standard errors are clustered at the county level, and *t*-statistics are reported in parentheses below the coefficients. Coefficients marked with ***, **, and * are significant at the 1%, 5%, and 10% level, respectively.

	(1) $\ln(1 + \textit{Patents})$	(2) <i>Patents</i>	(3) $\ln(1 + \textit{Patents})$	(4) <i>Patents</i>
Treat $\times$ Post	0.114*** (11.17)	0.232*** (4.78)	0.114*** (11.19)	0.230*** (4.75)
AdjControl $\times$ Post			0.014 (1.12)	0.031 (0.40)
Controls	Y	Y	Y	Y
County, Year, Cohort FE	Y	Y	Y	Y
Cohort-Year FE	Y	Y	Y	Y
Cohort-County FE	Y	Y	Y	Y
Estimation Method	OLS	Poisson	OLS	Poisson
Observations	39,314	32,391	46,684	38,631
<i>R</i> -squared	0.520		0.511	

Appendix A

Table A.1: **Variable Definitions**

Variable	Definition	Source
Panel A: Innovation outcomes		
Patents	Number of patents (unique publication numbers) with at least one inventor geolocated to county c in year t .	PatentCity
$\ln(1 + \text{Patents})$	Natural logarithm of one plus the county-year patent count.	PatentCity
Importance	County-year average of the ratio of forward to backward textual similarity (FS_5/BS_5 and $\text{FS}_{10}/\text{BS}_5$ variants); Kelly et al. (2021) “breakthrough” index, high for patents that are both distinct from prior art and echoed by later work.	Kelly et al. (2021)
Novelty	County-year average of the inverse of five-year backward similarity, $1/\text{BS}_5$, where BS_5 is the similarity of a patent to those in the prior five years; higher values denote less overlap with prior art. Reported as $10^6/\text{BS}_5$ in the regression tables for legibility—the 10^6 scaling is immaterial to the estimates.	Kelly et al. (2021)
Impact	County-year average of five-year forward similarity, FS_5 , the similarity of a patent to those filed in the subsequent five years; higher values denote greater influence.	Kelly et al. (2021)
Panel B: Treatment		
Treat	Indicator equal to one if county c is a treated unit—it gained railroad (rail-mail) access in its cohort year—and zero for its matched control county.	Atack (2015)
Post	Indicator equal to one in event years on or after the cohort’s treatment year ($t \geq 0$).	
Panel C: County controls		
$\ln(\text{Population})$	Natural logarithm of one plus total county population.	NHGIS
Urban share	Urban population divided by total county population.	NHGIS
White share	White population divided by total county population.	NHGIS
$\ln(\text{Distance to river})$	Natural logarithm of one plus the distance from the county to the nearest navigable river.	
$\ln(\text{Distance to canal})$	Natural logarithm of one plus the distance from the county to the nearest canal.	

Continued on next page

Table A.1 (continued)

Variable	Definition	Source
Ln(Manufacturing capital)	Natural logarithm of one plus the capital invested in manufacturing establishments in county c ; available from 1840 and used as an additional control in robustness tests.	Census of Manufactures
Ln(Farm value)	Natural logarithm of one plus the cash value of farms (land and buildings) in county c ; available from 1850 and used as an additional control in robustness tests.	Census of Agriculture
Ln(Steamboat mail frequency)	Natural logarithm of one plus the number of steamboat mail trips per week serving county c ; available from 1843 and used as an additional control in robustness tests.	<i>Annual Report of the Postmaster General</i>
Panel D: Railroad mail service and network position		
Trips	Total number of weekly trips on the rail mail routes serving county c ; available 1843–1880.	<i>Annual Report of the Postmaster General</i>
Routes	Number of distinct rail mail routes serving county c ; available 1843–1880.	<i>Annual Report of the Postmaster General</i>
Degree centrality	Number of counties to which county c is directly connected in the railroad network in year t , entered as a within-year percentile rank among connected counties and frozen at the connection year; available 1840–1898.	Atack (2015)
Betweenness centrality	Extent to which county c lies on the shortest paths linking other counties in the railroad network in year t , entered as a within-year percentile rank and frozen at the connection year; available 1840–1898.	Atack (2015)
Panel E: Mechanism and additional-analysis variables		
Patent similarity	Within-year percentile rank of the cosine similarity between a focal county’s patents in year t and a partner county’s patents in years $[t - 10, t - 3]$, computed from patent text embeddings using the Stella model; the outcome in the county-pair analysis.	
Connected	Indicator equal to one for a directed county pair joined by a common rail mail route in its cohort year, and zero for matched control pairs.	<i>Annual Report of the Postmaster General</i>

Continued on next page

Table A.1 (continued)

Variable	Definition	Source
1851 Reform	Indicator equal to one for the years 1851–1862, the regime of the Cheap Postage Act of 1851; the pre-1851 period is the omitted baseline.	
1863 Reform	Indicator equal to one for the years 1863 onward, the regime of the Postal Act of 1863.	
AdjControl	Indicator equal to one for a control county that shares a border with a treated county of the same adoption cohort (1890 county boundaries); used in the spatial-spillover test.	
Panel F: Civil War mail-route discontinuation		
South	Indicator equal to one if county c lies in one of the eleven Confederate states (Alabama, Arkansas, Florida, Georgia, Louisiana, Mississippi, North Carolina, South Carolina, Tennessee, Texas, Virginia).	NHGIS
War	Indicator equal to one in the war years 1861–1865 and zero in the pre-war baseline years 1856–1860.	
PreWarRoute	Indicator equal to one if county c had at least one rail mail route in any pre-war year (1856–1860).	<i>Annual Report of the Postmaster General</i>
SouthLostRoute	Indicator equal to one if county c is Confederate, had a pre-war mail route, and had no mail route during the war (1861–1865)—a county whose mail service was discontinued after secession.	<i>Annual Report of the Postmaster General</i>

Table A.2: **Treatment Effect on Innovation Quantity with Additional Controls**

This table assesses the robustness of the patent-quantity effect to controls for local economic conditions, using the stacked difference-in-differences matched sample. Columns (1) and (2) add log manufacturing capital to the baseline controls and restrict the sample to cohorts with treatment years from 1845 onward, the period over which manufacturing capital is observed throughout the event window; columns (3) and (4) further add log farm value and log steamboat mail frequency and restrict the sample to cohorts from 1855 onward, over which these additional controls are observed throughout the event window. The dependent variable is $\ln(1 + \text{Patents})$ or the patent count, estimated by OLS and Poisson. The baseline controls are population, the urban and white shares, and the log distances to the nearest river and canal. The independent variable *Treat* is an indicator equal to one if the county had its first railroad built in that cohort, and *Post* is an indicator equal to one for the years on or after the event year. Standard errors are clustered at the county level, and *t*-statistics are reported in parentheses below the coefficients. Coefficients marked with ***, **, and * are significant at the 1%, 5%, and 10% level, respectively.

	(1) ln (1 + Patents)	(2) Patents	(3) ln (1 + Patents)	(4) Patents
Treat×Post	0.112*** (11.28)	0.226*** (5.03)	0.108*** (9.57)	0.198*** (4.05)
Ln(Population)	0.050*** (8.14)	0.260*** (5.68)	0.055*** (7.48)	0.268*** (5.55)
Urban share	0.023 (0.64)	0.145* (1.93)	0.018 (0.57)	0.136* (1.91)
White share	-0.371*** (-7.42)	-3.110*** (-5.77)	-0.387*** (-7.15)	-3.039*** (-5.68)
Ln(Distance to river)	-0.044*** (-2.97)	-0.238** (-2.55)	-0.044 (-0.92)	-0.977** (-2.47)
Ln(Distance to canal)	0.011 (0.73)	0.063** (2.16)	0.004 (0.20)	0.063** (2.13)
Ln(Manufacturing capital)	0.004*** (2.67)	0.023** (2.18)	0.004** (2.40)	0.024** (2.19)
Ln(Farm value)			-0.004 (-1.39)	-0.012 (-0.79)
Ln(Steamboat mail frequency)			0.019** (2.44)	0.054** (2.20)
County, Year, Cohort FE	Y	Y	Y	Y
Cohort-Year FE	Y	Y	Y	Y
Cohort-County FE	Y	Y	Y	Y
Estimation Method	OLS	Poisson	OLS	Poisson
Observations	44,124	36,720	35,861	30,247
R-squared	0.511		0.506	

Table A.3: **Heterogeneous Treatment Effects on Patent Quantity: Top-Decile Hub Definition**

This table reports the innovation-hub heterogeneity in the patent-quantity effect of railroad access under a stricter, top-decile definition of hub status, as a robustness check on Panel A of Table 5, which uses the top-quartile cutoff. A county is classified as a hub if its pre-treatment national patent-stock rank lies at or above the 90th percentile; *Hub* and *NonHub* are defined exactly as in Table 5, and the sample, specification, and controls are identical. The two triple interactions report the difference-in-differences effect within the hub and non-hub groups, and *Difference (Hub – NonHub)* reports the gap between them, with its *t*-statistic in the OLS columns and *z*-statistic in the Poisson columns. All specifications control for log population, the urban and white population shares, and the log distances to the nearest river and canal. Standard errors are clustered at the county level, and *t*-statistics are reported in parentheses below the coefficients. Coefficients marked with ***, **, and * are significant at the 1%, 5%, and 10% levels, respectively.

	(1) ln (1 + Patents)	(2) Patents	(3) ln (1 + Patents)	(4) Patents
Treat×Post×Hub	0.197*** (3.45)	0.281** (2.50)	0.195*** (3.39)	0.303*** (2.69)
Treat×Post×NonHub	0.099*** (10.17)	0.209*** (4.44)	0.103*** (10.66)	0.217*** (4.57)
<i>Difference (Hub – NonHub)</i>	0.098* (1.69)	0.072 (0.59)	0.092 (1.58)	0.087 (0.71)
<i>R</i> -squared	0.454		0.512	
Controls	Y	Y	Y	Y
County, Year, Cohort FE	Y	Y	Y	Y
Cohort-Year FE			Y	Y
Cohort-County FE			Y	Y
Estimation Method	OLS	Poisson	OLS	Poisson
Observations	46,684	42,294	46,684	38,631

Table A.4: **Heterogeneous Treatment Effects on Patent Importance: Top-Decile Hub Definition**

This table reports the innovation-hub heterogeneity in patent importance under a stricter, top-decile definition of hub status, as a robustness check on the hub columns of Table 6, which use the top-quartile cutoff. A county is a hub if its pre-treatment national patent-stock rank is at or above the 90th percentile. The dependent variable is $\ln(FS10/BS5)$, the log ratio of ten-year forward to five-year backward similarity, averaged over the patents produced in each county-year. All specifications control for log population, the urban and white population shares, the log distances to the nearest river and canal, log manufacturing capital, and log steamboat mail frequency. Standard errors are clustered at the county level, and t -statistics are in parentheses. ***, **, and * denote significance at the 1%, 5%, and 10% levels.

	$\ln(FS10/BS5)$	
	(1)	(2)
Treat $\times$ Post $\times$ Hub	0.031 (0.95)	0.034 (0.98)
Treat $\times$ Post $\times$ NonHub	-0.015** (-1.97)	-0.014* (-1.66)
<i>Difference (Hub – NonHub)</i>	0.046 (1.39)	0.047 (1.34)
Controls	Y	Y
County, Year, Cohort FE	Y	Y
Cohort-Year FE		Y
Cohort-County FE		Y
Estimation Method	OLS	OLS
Observations	11,271	10,969
R -squared	0.846	0.858

Table A.5: **Treatment Effect on Innovation Quality with Additional Controls**

This table assesses the robustness of the patent-quality effect to controls for local economic conditions, using the stacked difference-in-differences matched sample. I restrict the sample to cohorts with treatment years from 1855 onward, the period over which additional controls is observed throughout the event window. The dependent variables are patent textual-similarity measures, averaged across the patents granted in each county-year, estimated by OLS in logs and by Poisson in levels. *Importance* is the ratio of the five-year forward similarity *FS5* to the five-year backward similarity *BS5*, or of the ten-year forward similarity to *BS5*. *Novelty* is the inverse of the five-year backward similarity, $1/BS5$, so that higher values indicate patents less similar to prior art (i.e., more novel). *Impact* is the five-year forward similarity, *FS5*. The independent variable *Treat* is an indicator equal to one if the county had its first railroad built in that cohort, and *Post* is an indicator equal to one for the years on or after the event year. All specifications include the baseline controls: population, the urban and white shares, and the log distances to the nearest river and canal. Standard errors are clustered at the county level, and *t*-statistics are reported in parentheses below the coefficients. Coefficients marked with ***, **, and * are significant at the 1%, 5%, and 10% level, respectively.

	Importance		Importance		Novelty		Impact	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	$\ln(\frac{FS5}{BS5})$	FS5/BS5	$\ln(\frac{FS10}{BS5})$	FS10/BS5	$\ln(1/BS5)$	1/BS5	$\ln(FS5)$	FS5
Treat×Post	-0.012*	-0.017**	-0.013*	-0.019**	-0.050	-0.182**	0.038	0.008
	(-1.82)	(-2.13)	(-1.67)	(-1.98)	(-1.55)	(-2.19)	(1.20)	(0.26)
Ln(Farm value)	-0.001	-0.001	-0.000	0.000	-0.015*	-0.005	0.015*	0.012
	(-0.40)	(-0.38)	(-0.06)	(0.10)	(-1.85)	(-0.37)	(1.94)	(1.60)
Ln(Steamboat mail frequency)	-0.007	-0.007	-0.008	-0.010*	0.001	0.177*	-0.008	-0.010
	(-1.48)	(-1.42)	(-1.56)	(-1.66)	(0.04)	(1.95)	(-0.42)	(-0.60)
Ln(Manufacturing capital)	-0.001	0.000	-0.001	-0.000	-0.002	0.003	0.001	-0.002
	(-0.67)	(0.05)	(-1.01)	(-0.24)	(-0.35)	(0.26)	(0.14)	(-0.39)
Controls	Y	Y	Y	Y	Y	Y	Y	Y
County, Year, Cohort FE	Y	Y	Y	Y	Y	Y	Y	Y
Cohort-Year FE	Y	Y	Y	Y	Y	Y	Y	Y
Cohort-County FE	Y	Y	Y	Y	Y	Y	Y	Y
Estimation Method	OLS	Poisson	OLS	Poisson	OLS	Poisson	OLS	Poisson
Observations	10,022	10,022	10,022	10,022	10,022	10,022	10,022	10,022
R-squared	0.807		0.808		0.698		0.612	

Table A.6: **Treatment Effect on Innovation Quality at Patent Level**

This table reports patent-level estimates of the effect of first railroad access on patent quality, using the stacked difference-in-differences matched sample, restricted to cohorts with treatment years from 1845 onward. Each observation is an individual patent, assigned to a county and grant year by the location of its first-listed inventor and inheriting the treatment status of that county-year; the dependent variable is the natural logarithm of the patent's own similarity measure. Importance is the ratio of five-year forward to five-year backward similarity, $FS5/BS5$, or of ten-year forward to five-year backward similarity, $FS10/BS5$; Novelty is the inverse of five-year backward similarity, $1/BS5$, so that higher values indicate patents less similar to prior art; and Impact is the five-year forward similarity, $FS5$. All columns are estimated by OLS and control for population, the urban and white shares, the log distances to the nearest river and canal, and log manufacturing capital. The fixed effects included in each column, reported at the foot of the table, comprise county, year, cohort, and CPC technology-class fixed effects, with cohort-by-year and cohort-by-county fixed effects in the extended specifications. The independent variable *Treat* is an indicator equal to one if the county had its first railroad built in that cohort, and *Post* is an indicator equal to one for the years on or after the event year. Standard errors are clustered at the county level, and *t*-statistics are reported in parentheses below the coefficients. Coefficients marked with ***, **, and * are significant at the 1%, 5%, and 10% level, respectively.

	Importance $\ln(FS5/BS5)$		Importance $\ln(FS10/BS5)$		Novelty $\ln(1/BS5)$		Impact $\ln(FS5)$	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treat×Post	-0.009 (-1.43)	-0.007 (-1.19)	-0.013* (-1.87)	-0.012* (-1.72)	-0.001 (-0.05)	-0.009 (-0.32)	-0.007 (-0.30)	0.001 (0.05)
Treat	0.005 (0.78)		0.006 (0.86)		-0.028 (-1.05)		0.034 (1.19)	
Post	0.015** (2.45)		0.016** (2.52)		0.026 (1.03)		-0.011 (-0.46)	
Controls	Y	Y	Y	Y	Y	Y	Y	Y
County, Year, Cohort FE	Y	Y	Y	Y	Y	Y	Y	Y
CPC-class FE	Y	Y	Y	Y	Y	Y	Y	Y
Cohort-Year FE		Y		Y		Y		Y
Cohort-County FE		Y		Y		Y		Y
Observations	23,762	23,453	23,762	23,453	23,762	23,453	23,762	23,453
<i>R</i> -squared	0.767	0.775	0.787	0.795	0.734	0.741	0.662	0.671

Table A.7: **Treatment Intensity and Patent Importance: Mail Service and Network Position**

This table reports treatment-intensity tests for the effect of first railroad access on patent importance, using the county-level stacked difference-in-differences matched sample. The dependent variable is the natural logarithm of patent importance, the ratio of forward to backward textual similarity averaged over the patents produced in each county-year; Panel A uses the five-year forward ratio $\ln(FS5/BS5)$ and Panel B the ten-year forward ratio $\ln(FS10/BS5)$. Each column interacts $Treat \times Post$ with a different connection measure, denoted *Measure* in the coefficient labels. In the mail-service columns, *Trips* is the weekly mail trips serving a county and *Routes* is the number of distinct mail routes, observed over 1843–1880. In the network-centrality columns, degree and betweenness centrality are each entered as a within-year percentile rank among connected counties, frozen at the connection year, and defined over the full 1840–1898 sample. The mail-service and network columns use different samples because the mail-route measures are observed only for 1843–1880, whereas railroad-network centrality is defined throughout; this accounts for the difference in observation counts. Each specification also includes the lower-order *Treat* and *Post* terms and controls for log population, the urban and white population shares, and the log distances to the nearest river and canal, whose coefficients are omitted for brevity. Standard errors are clustered at the county level, and *t*-statistics are reported in parentheses below the coefficients. Coefficients marked with ***, **, and * are significant at the 1%, 5%, and 10% levels, respectively.

	Mail service		Network centrality	
	Trips	Routes	Degree	Betweenness
	(1)	(2)	(3)	(4)
Panel A: Importance, $\ln(FS5/BS5)$				
$Treat \times Post \times Measure$	0.000	0.004	0.001	-0.015
	(1.36)	(1.19)	(0.07)	(-1.21)
$Treat \times Post$	-0.007	-0.008	-0.009	-0.005
	(-0.62)	(-0.69)	(-1.00)	(-0.61)
Observations	6,246	6,246	11,534	11,534
<i>R</i> -squared	0.853	0.853	0.848	0.848
Panel B: Importance, $\ln(FS10/BS5)$				
$Treat \times Post \times Measure$	0.000	0.004	-0.004	-0.022
	(0.86)	(1.26)	(-0.27)	(-1.45)
$Treat \times Post$	-0.003	-0.006	-0.007	-0.002
	(-0.27)	(-0.47)	(-0.61)	(-0.27)
Observations	6,246	6,246	11,039	11,039
<i>R</i> -squared	0.851	0.851	0.858	0.858
Controls	Y	Y	Y	Y
County, Year, Cohort FE	Y	Y	Y	Y
Cohort-Year FE	Y	Y	Y	Y
Cohort-County FE	Y	Y	Y	Y
Estimation Method	OLS	OLS	OLS	OLS

Table A.8: **Treatment Intensity and Patent Quality at the Patent Level**

This table reports patent-level treatment-intensity tests of the effect of first railroad access on patent quality using the stacked DID matched sample over the 1843 to 1880 period, for which railroad mail-route data are available. The sample is constructed using an 11-year event window around each county's first railroad connection. Each observation corresponds to a patent and is assigned the stacked treatment status of the county associated with its first inventor. *Trips* (Panel A) is the number of weekly railroad mail trips passing through a county and captures delivery frequency, while *Routes* (Panel B) is the number of distinct railroad mail routes serving a county and captures the breadth of communication links. FS_5/BS_5 is the ratio of five-year forward similarity to five-year backward similarity, FS_{10}/BS_5 the ratio of ten-year forward similarity to five-year backward similarity, *Novelty* is defined as $1/BS_5$, so that larger values indicate patents that are less similar to prior art and therefore more novel, and *Impact* is measured by FS_5 ; all outcomes are estimated by OLS. The independent variable *Treat* is an indicator equal to one for patents whose county received its first railroad in that cohort, and *Post* is an indicator equal to one for patents granted in years on or after the treatment year. Each specification also includes the lower-order *Treat* and *Post* terms, whose coefficients are omitted for brevity, and the fixed effects reported at the foot of the table: county, year, cohort, and CPC technology-class fixed effects, with cohort-by-year and cohort-by-county fixed effects in the extended columns. Standard errors are clustered at the county level, and *t*-statistics are reported in parentheses below the coefficients. Coefficients marked with ***, **, and * are significant at the 1%, 5%, and 10% levels, respectively.

	Importance $\ln(FS_5/BS_5)$		Importance $\ln(FS_{10}/BS_5)$		Novelty $\ln(1/BS_5)$		Impact $\ln(FS_5)$	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Delivery Frequency								
Treat×Post×Trips	0.000 (0.44)	0.000 (1.19)	0.000 (0.14)	0.000 (0.81)	0.001 (0.77)	0.001 (1.09)	-0.001 (-0.65)	-0.001 (-0.75)
Treat×Post	-0.008 (-0.79)	-0.009 (-0.87)	-0.007 (-0.69)	-0.010 (-0.88)	0.027 (0.78)	0.008 (0.21)	-0.035 (-1.02)	-0.017 (-0.48)
R-squared	0.774	0.782	0.773	0.780	0.709	0.714	0.616	0.622
Panel B: Number of distinct routes								
Treat×Post×Routes	0.002 (0.72)	0.004 (1.50)	0.002 (0.63)	0.004 (1.38)	0.015* (1.89)	0.020** (2.34)	-0.013 (-1.60)	-0.016* (-1.79)
Treat×Post	-0.010 (-1.01)	-0.012 (-1.14)	-0.011 (-0.96)	-0.014 (-1.17)	0.014 (0.40)	-0.008 (-0.21)	-0.024 (-0.70)	-0.004 (-0.13)
R-squared	0.773	0.780	0.774	0.782	0.709	0.714	0.616	0.622
Controls	Y	Y	Y	Y	Y	Y	Y	Y
County, Year, Cohort FE	Y	Y	Y	Y	Y	Y	Y	Y
CPC-class FE	Y	Y	Y	Y	Y	Y	Y	Y
Cohort-Year FE		Y		Y		Y		Y
Cohort-County FE		Y		Y		Y		Y
Observations	12,787	12,661	12,787	12,661	12,787	12,661	12,787	12,661

Table A.9: **Innovation Hub Heterogeneity in Patent Importance, by Water Access**

This table examines whether the innovation-hub premium on patent importance differs by a county's access to water. Estimating separately within the subsamples of counties with and without water access, I run the innovation-hub two-triple difference-in-differences: $Treat \times Post \times Hub$ and $Treat \times Post \times NonHub$ give the treatment effect for hub and non-hub counties within the subsample, and $Difference (Hub - NonHub)$ is their gap, computed from a linear combination of the two coefficients, with its t -statistic reported below. Hub is the predetermined innovation-stock measure defined in Table 5, here using the top-quartile threshold (pre-treatment rank at or above the 75th percentile). $Water$ equals one for matched pairs whose treated county intersects a navigable river, a canal, the Atlantic coast, or the Great Lakes, and zero otherwise. The importance measures $\ln(FS5/BS5)$ and $\ln(FS10/BS5)$ are estimated over cohorts with treatment years from 1850 to 1895 and 1850 to 1890, respectively. All specifications control for log population, the urban and white population shares, the log distances to the nearest river and canal, log manufacturing capital, and log steamboat mail frequency. Standard errors are clustered at the county level. Coefficients marked with ***, **, and * are significant at the 1%, 5%, and 10% levels, respectively.

	$\ln(FS5/BS5)$				$\ln(FS10/BS5)$			
	Water		No-water		Water		No-water	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$Treat \times Post \times Hub$	0.052** (2.03)	0.062** (2.36)	-0.014 (-0.84)	-0.011 (-0.63)	0.060** (2.24)	0.070** (2.50)	-0.017 (-0.88)	-0.014 (-0.72)
$Treat \times Post \times NonHub$	-0.004 (-0.15)	-0.007 (-0.27)	-0.020*** (-2.66)	-0.019** (-2.41)	0.008 (0.24)	0.001 (0.02)	-0.021** (-2.25)	-0.019** (-1.97)
$Difference (Hub - NonHub)$	0.055 (1.55)	0.069* (1.86)	0.006 (0.31)	0.008 (0.44)	0.052 (1.26)	0.069 (1.57)	0.004 (0.17)	0.005 (0.21)
Controls	Y	Y	Y	Y	Y	Y	Y	Y
County, Year, Cohort FE	Y	Y	Y	Y	Y	Y	Y	Y
Cohort-Year FE		Y		Y		Y		Y
Cohort-County FE		Y		Y		Y		Y
Estimation Method	OLS	OLS	OLS	OLS	OLS	OLS	OLS	OLS
Observations	1,531	1,521	10,180	9,903	1,472	1,462	9,764	9,503

Table A.10: The Effect of the Postal Reforms on Patent Quality

This table reports the effect of railroad access on patent quality across the nineteenth-century postal-reform regimes, using the same stacked difference-in-differences matched sample restricted to cohorts with treatment years from 1840-1868, and triple-difference specification as Table 8. The pre-1851 period is the omitted baseline. 1851*Reform* equals one for 1851–1862 and 1863*Reform* equals one for 1863 onward. Importance is the ratio of forward to backward similarity, $FS5/BS5$ or $FS10/BS5$; Novelty is the inverse of five-year backward similarity, $1/BS5$; and Impact is the five-year forward similarity, $FS5$. All quality measures are county-year averages, and all specifications control for log population, the urban and white shares, and the log distances to the nearest river and canal. Standard errors are clustered at the county level. Coefficients marked with ***, **, and * are significant at the 1%, 5%, and 10% level, respectively.

	Importance		Importance		Novelty		Impact	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	$\ln(\frac{FS5}{BS5})$	$FS5/BS5$	$\ln(\frac{FS10}{BS5})$	$FS10/BS5$	$\ln(1/BS5)$	$1/BS5$	$\ln(FS5)$	$FS5$
Treat×Post×1851Reform	0.011 (0.22)	-0.021 (-0.43)	0.019 (0.37)	-0.048 (-0.88)	0.197 (1.15)	-0.174 (-0.62)	-0.195 (-1.16)	-0.296** (-2.13)
Treat×Post×1863Reform	-0.019 (-0.37)	-0.057 (-1.05)	0.021 (0.37)	-0.046 (-0.76)	0.118 (0.60)	-0.024 (-0.07)	-0.128 (-0.66)	-0.273* (-1.68)
Treat×Post	-0.017 (-0.44)	0.016 (0.39)	-0.037 (-0.83)	0.025 (0.54)	-0.116 (-0.72)	0.087 (0.33)	0.107 (0.69)	0.212 (1.62)
Treat×1851Reform	-0.003 (-0.07)	0.010 (0.29)	0.022 (0.47)	0.051 (1.13)	-0.001 (-0.01)	0.132 (0.68)	-0.009 (-0.08)	-0.035 (-0.34)
Treat×1863Reform	0.051 (0.90)	0.064 (1.10)	0.036 (0.56)	0.069 (1.02)	-0.068 (-0.37)	-0.351 (-1.13)	0.090 (0.53)	-0.012 (-0.08)
Controls	Y	Y	Y	Y	Y	Y	Y	Y
County, Year, Cohort FE	Y	Y	Y	Y	Y	Y	Y	Y
Cohort-Year FE	Y	Y	Y	Y	Y	Y	Y	Y
Cohort-County FE	Y	Y	Y	Y	Y	Y	Y	Y
Estimation Method	OLS	Poisson	OLS	Poisson	OLS	Poisson	OLS	Poisson
Observations	4,070	4,070	4,070	4,070	4,070	4,070	4,070	4,070
R-squared	0.827		0.790		0.727		0.738	

Table A.11: **Incidental connection: alternative definitions of route endpoints**

This appendix table verifies that the incidental-connection result in Table 10 is not sensitive to how route endpoints are defined or excluded. In the baseline test I drop treated counties that were the origin or destination of a mail route in any year of their event window; here I report three alternatives. In columns (1) and (2), I drop not only each treated endpoint county but also its matched control, removing the entire matched pair so that the comparison is not affected by orphaned controls ("whole-pair exclusion"). In columns (3) and (4), I define endpoints from the railroad route network rather than the mail-route network, dropping treated counties that were a railroad-route origin or destination in their event window. In columns (5) and (6), I broaden the criterion to drop any treated county that served as a mail-route endpoint in any year, not only within its cohort window ("ever-endpoint"). In every case the surviving treated units are pass-through, incidentally connected counties. The odd-numbered columns report OLS estimates with the dependent variable $\ln(1 + \textit{Patents})$, the natural logarithm of one plus the total number of patents generated within each county-year, and the even-numbered columns Poisson estimates on the patent count. *Treat* is an indicator equal to one for the county that gained railroad access in its cohort and zero for its matched control; *Post* equals one in event years on or after the treatment year; and the coefficient of interest is the difference-in-differences term $\textit{Treat} \times \textit{Post}$. The sample comprises railroad cohorts with treatment years 1848–1875. Across all columns I control for $\ln(\text{Population})$, the urban share, the white share, $\ln(\text{Manufacturing capital})$, $\ln(\text{Steamboat mail frequency})$, and the log distances to the nearest river and canal. Standard errors are clustered at the county level, and *t*-statistics are reported in parentheses below the coefficients. Coefficients marked with ***, **, and * are significant at the 1%, 5%, and 10% level, respectively.

	(1) $\ln(1 + \textit{Patents})$	(2) <i>Patents</i>	(3) $\ln(1 + \textit{Patents})$	(4) <i>Patents</i>	(5) $\ln(1 + \textit{Patents})$	(6) <i>Patents</i>
<i>Treat</i> × <i>Post</i>	0.100*** (6.90)	0.272*** (2.79)	0.092*** (6.79)	0.255*** (3.38)	0.091*** (6.50)	0.268*** (3.54)
Controls	Y	Y	Y	Y	Y	Y
Year FE	Y	Y	Y	Y	Y	Y
Cohort FE	Y	Y	Y	Y	Y	Y
Cohort-Year FE	Y	Y	Y	Y	Y	Y
Cohort-County FE	Y	Y	Y	Y	Y	Y
Estimation Method	OLS	Poisson	OLS	Poisson	OLS	Poisson
Observations	16,522	13,181	21,967	17,696	21,230	16,992
<i>R</i> -squared	0.487		0.497		0.501	

Table A.12: **County-Pair Similarity by Partner Hub Status, Across Hub Cutoffs**

This table reports the partner-hub split of the county-pair similarity effect (Panel B of Table 7) across two thresholds for defining a hub partner: the 75th percentile (the top quartile) and the 90th of the partner county's pre-treatment national patent-stock distribution. The dependent variable, sample, matching, and controls are as in Table 7. *Connected*×*Post*×*Hub partner* and *Connected*×*Post*×*Non-hub partner* are the connection effects within the hub-partner and non-hub-partner groups, and *Difference (Hub – Non-hub)* is the gap between them, whose *t*-statistic is the square root of the *F*-statistic testing equality of the two triples. The convergence toward non-hub partners holds at both the 75th and 90th percentiles. Each specification also includes the lower-order interaction terms, the pair distance, and $\ln(\# \text{ patent pairs})$, whose coefficients are omitted for brevity. Standard errors are clustered by focal county, and *t*-statistics are reported in parentheses below the coefficients. Coefficients marked with ***, **, and * are significant at the 1%, 5%, and 10% levels, respectively.

	Similarity					
	Hub top 25%			Hub top 10%		
	(1)	(2)	(3)	(4)	(5)	(6)
Connected×Post×Hub partner	0.092 (0.81)	0.162 (1.46)	0.142 (1.25)	-0.000 (-0.00)	0.114 (0.86)	0.080 (0.58)
Connected×Post×Non-hub partner	1.119*** (3.34)	1.259*** (3.66)	1.298*** (3.71)	0.450*** (2.83)	0.508*** (3.25)	0.512*** (3.21)
<i>Difference (Hub – Non-hub)</i>	-1.027*** (-3.02)	-1.097*** (-3.18)	-1.156*** (-3.30)	-0.450** (-2.26)	-0.394** (-2.11)	-0.432** (-2.26)
Controls	Y	Y	Y	Y	Y	Y
Focal × Year FE	Y	Y	Y	Y	Y	Y
Partner × Year FE	Y	Y	Y	Y	Y	Y
Cohort FE	Y			Y		
Directed-pair × Cohort FE		Y	Y		Y	Y
Excludes pairs < 35 mi			Y			Y
Observations	403,849	403,819	392,868	403,849	403,819	392,868
<i>R</i> -squared	0.834	0.874	0.874	0.834	0.874	0.874

A.1 Patent Similarity

I calculate pairwise similarities of patent pair (i, j) using patent descriptions. First, for each term w in patent i I calculate the term-frequency-backward-inverse-document-frequency ($TFBIDF$) as

$$TFBIDF_{w,i,t} = TF_{w,i} \times BIDF_{w,t}, \quad (10)$$

The first component of the weight, term frequency $TF_{w,i}$, is defined as the times term w appears in patent i adjusted for the patent's length. The second component, backward inverse document frequency $BIDF$ of term w , is defined as

$$BIDF_{w,i,t} = \log \left(\frac{\# \text{ patents prior to } i}{1 + \# \text{ documents prior to } p \text{ that include term } w} \right) \quad (11)$$

Second, for patent j in patent pair (i, j) , I repeat the process and calculate $TFBIDF_{w,j,t}$. Then I normalize each $TFBIDF$ vector $TFBIDF_{i,t}$ and $TFBIDF_{j,t}$ and calculate the cosine similarity between the two normalized vectors:

$$\rho_{i,j} = \frac{TFBIDF_{i,t}}{\|TFBIDF_{i,t}\|} \cdot \frac{TFBIDF_{j,t}}{\|TFBIDF_{j,t}\|}, \quad (12)$$

where the cosine similarity $\rho_{i,j}$ takes value between 0 and 1. A similarity of one indicates that patent i and patent j use the exact same set of words in the same proportion, whereas a similarity of zero indicates that these two patents have no overlapping terms.

Appendix B

Construction of the Mail-Route–Railroad Matched Data

This section describes the construction of the historical mail-route panel used in the analysis and reports summary statistics for the resulting dataset. The constructed sample spans 38 annual cross-sections from 1843 to 1880 and comprises 18,220 route-year records.

B.0.1 Data collection

The data combine two primary sources. First, the route records are transcribed from the per-route tables of railroad mail service published annually in the *Annual Report of the Postmaster General*.⁴⁵ Each record reports a numbered postal route within a state, the route’s termini (as free text), the scheduled route distance in miles, the contracted service frequency (weekly trips), and the contracting carrier’s corporate title. Second, railroad geometries are taken from the digitized historical railroad GIS of [Atack \(2015\)](#), in which the network is represented as line segments, each carrying a unique segment identifier and a “first year in operation”; for every route-year I restrict the candidate network to segments in operation by that year.

With the mail-route records, I next perform two steps to prepare the data for matching: route-table extraction and endpoint geocoding. To extract route records from the digitized reports, I use the vision-capable large language model Google Gemini Pro 3, prompted with year-specific instructions and a predefined JSON schema.⁴⁶ (See prompt [B.1](#)) This approach substantially improves extraction accuracy relative to manual transcription and conventional OCR methods, both of which are prone to errors when processing dense nineteenth-century tabular records.

⁴⁵Digital scans were obtained from the HathiTrust Digital Library. The relevant tables appear in the *Second Assistant Postmaster-General’s* section of each report.

⁴⁶Because the column layout and headings vary across reports, the prompt is tailored to each year’s document. For example, antebellum tables report only an ‘Annual Cost’ column, whereas the 1870 report additionally includes an ‘Annual Cost per Mile’ column, and some years contain separate steamboat-route tables.

Route origin and destination place names are geocoded using a five-source gazetteer cascade consisting of NHGIS 1900 incorporated-place points, the GNIS Populated Places file, the GNIS Domestic Names file, the GNIS All Names file containing variant and historical spellings, and, as a final step, an edit-distance-one fuzzy-matching procedure.⁴⁷

B.0.2 Matching routes to the railroad network

Matching postal mail routes to the railroad network is challenging because the two sources share no common identifier and are organized at fundamentally different levels of aggregation. The historical railroad GIS represents the system as tens of thousands of short geometric segments rather than as named, end-to-end routes; a mail route, by contrast, is described in the Post Office tables by its terminal place names, its scheduled distance, and the corporate title of the contracting carrier.⁴⁸ A single route therefore corresponds not to one record in the railroad data but to an ordered set of segments that must be reconstructed, while any given segment may be traversed by several routes, so the mapping is inherently many-to-many.⁴⁹ I therefore match routes to railroad segments in two regimes, through 1870 and 1871–1880, through a layered procedure to to maximize precision and to record the provenance of every match.

⁴⁷Before matching, place names are standardized by normalizing abbreviations, directional prefixes, and railroad-specific suffixes such as *Junction*, *Depot*, and *Landing*. Ambiguous place names are further resolved using state information implied by the route or embedded in the place name itself. Each successfully geocoded endpoint is then assigned to the county in which it is located.

⁴⁸although each segment carries attributes—a segment identifier, the operating railroad’s name, its track gauge, and the year it entered operation—these are an imperfect basis for matching. The company name attached to a segment frequently differs from the corporate title of the contracting carrier recorded in the Post Office tables, because the era’s pervasive mergers, leases, and renamings separate the entity that holds the mail contract from the road that physically owns the track, and the GIS names are themselves often missing or inconsistently recorded.

⁴⁹In addition to that, it is the absence of a shared key, noisy place and carrier names, a time-varying network, and path ambiguity—that renders any single matching rule insufficient and motivates the layered, multi-signal procedure described below. Bridging the two requires resolving nineteenth-century place names that are frequently abbreviated, inconsistently spelled, duplicated across states, or corrupted in transcription, and reconciling carrier titles that shift with the era’s frequent mergers, leases, and renamings. The difficulty is compounded by the network’s rapid build-out: lines are extended, rerouted, and consolidated from year to year, so each route must be matched against the network as it existed in that year, and the same route label can map to different physical segments over time. Finally, because multiple plausible paths often connect the same pair of termini and routes themselves branch and merge, the specific segments a route traversed are underdetermined by its textual description alone.

I match in two regimes, through 1870 and 1871–1880, primarily because of two features of the railroad data. First, in the digitized railroad GIS a segment carries its operating company’s name only for track in place before Civil War, whereas after 1870 the company-name field is largely empty.⁵⁰ Second, the network expands enormously after 1870: national mileage roughly doubles over the decade, and parallel main lines, branches, and competing roads proliferate. Through 1870 neither problem binds: much of the track is still named, and the network is sparse enough that a route’s endpoints usually pin down a unique line. Each route can therefore be matched directly, by geocoding its origin and destination and linking them to the connecting named lines, and the result can be validated against a hand-verified 1870 cross-section. Splitting at 1870 aligns the matching procedure with both the descriptive information of railroad segments and the network density present in each period.

B.0.2.1 Matching between 1843 - 1870

For the cross-sections through 1870, mail routes are matched to railroad lines in two steps. First, each railroad line operating in a given year is summarized by its origin and destination cities, together with the sequence of cities and towns located along the line, yielding a description directly comparable to the route’s reported termini. Second, a large language model, GPT-4o mini, is prompted on a state-by-state basis with a batch of mail routes, including their termini, contracting railroad company, and reported distance, as well as the set of candidate railroad lines in the state. The model is then asked to assign each route to the railroad line or lines it traverses, identified by the corresponding line-year group (see prompt B.2). Because the matching procedure is automated, I take another step to validate its output against an independently hand-matched 1870 cross-section, which serves as the ground-truth benchmark. The verified 1870 matches are then used to anchor and guide the matching procedure for the 1871–1880 cross-sections described below.

⁵⁰The names come from [Taylor and Neu \(1956\)](#) for the 1850s and from the Hallberg database for 1850 and earlier, and no railroad name is assigned to track built after 1861 ([Atack, 2015](#))

B.0.2.2 Matching between 1871 - 1880

For these years, matching is anchored on the hand-verified 1870 cross-section and implemented through a layered procedure that prioritizes accuracy while systematically documenting the provenance of every match:

1. **Termini matching to the 1870 anchor.** Routes whose termini match a 1870 anchor route inherit that route’s matched segment set, filtered to segments in operation by the route’s year. Matching proceeds in two tiers: an exact match on normalized termini (lowercased, punctuation removed), followed by a state-bounded fuzzy match that recovers routes whose termini differ only in spelling, punctuation, or word order. Anchoring to a single reviewed year avoids compounding automated matches across years.
2. **Spatial (GIS) matching and verification.** Each route’s geocoded origin and destination are mapped onto the contemporary Open Street Map railway and road networks, and the resulting corridor is intersected with candidate railroad segments. When the inferred corridor and the candidate railroad geometry identify the same line, the match is accepted. This procedure both recovers routes that were missed during the anchor-matching stage and independently validates routes that were previously identified, following the hierarchy of trust tiers described above.
3. **Ensemble matching of residuals.** Routes not resolved through anchoring are matched using an ensemble approach that combines three independent classifiers: a TF-IDF text-similarity model, a corporate-name lookup algorithm, and a large language model classifier (Claude Opus 4.7). Candidate matches are aggregated using majority voting. I further impose a distance-plausibility filter that excludes matches whose geometric length is implausible given the reported route distance.
4. **Manual verification.** For routes that remained unresolved after the automated

matching procedures, or for which the available sources provided conflicting information, I conducted manual verification and propagate these validations to adjacent years where appropriate.

Routes that no method resolves are flagged as the low-confidence and unmatched tail. Overall, 14,580 of 18,220 route-years (80.0%) are matched to at least one railroad segment.layer.

Figure B.1: Railroad mail service on the *Annual Report of the Postmaster General*

Generated at Georgia State University on 2024-08-18 16:47 GMT / https://hdl.handle.net/2027/hvd.hwc6p
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by
Google

Original from
HARVARD UNIVERSITY

D.								
Railroad service as in operation on the 1st of October, 1850.								
State.	Number of route.	Termini.	Distance.	Total distance in each State.	Number of trips per week.	Annual pay.	Annual pay in each State.	Remarks.
			Miles.	Miles.				
Maine.....	61a	From Bangor to Upper Stillwater..	9		6	\$50		
	91	From Danville Junction to Waterville..	55		6	3,472		
	94	From Portland to Portsmouth, N. H.	52½		12	6,718		
	97	From Portland to South Paris.....	48		6	3,257		
New Hampshire.....	99	From Portland to Bath.....	33	197½	12	1,900	\$15,394	Under coach contract.
	201	From Concord to Lowell, Mass.....	50		18	6,429		Embraces side supply.
	206	From Concord to Plymouth.....	50		6	2,393		do.....do.
	207	From Concord to West Lebanon, with branch to Bristol.....	82		6	6,471		do.....do.
Vermont.....	237a	From Dover to Farmington.....	18	200	6	80	15,373	
	301	From Montpelier to Windsor.....	77		6	6,600		
	308	From Montpelier to Burlington.....	40		6	2,950		Under coach contract.
	360	From West Lebanon, New Hampshire, to Wells River.....	41½		6	557		
Massachusetts.....	365	From Bellows Falls to Windsor.....	24½		6	2,332		Embraces side supply.
	377	From Bellows Falls to Burlington...	119	302½	6	11,200	26,629	do.....do.
	401	From Boston to Portsmouth, New Hampshire.....	54½		12 }	8,324		do.....do.
	402	Branch to Marblehead.....	4		6 }			
		From Boston to South Berwick Junction, Maine, with 3 miles branch from Dover, New Hampshire, to Great Falls.....	77½		12 }	6,921		do.....do.
	403	From Boston to Lowell.....	26		18 }			
		Branch to Woburn.....	3		6 }	3,600		

Doc. No. 1.

331

Table B.2: Prompt used to match mail routes to railroad lines.

Task. You are an expert historical data analyst. Your goal is to match Mail Routes (Table A) to Rail Lines (Table B) from the year 1846. You must be aggressive in finding matches, understanding that historical rail endpoints were often:

1. **Suburbs / nearby towns** (e.g., “Chelsea” or “Cambridge” instead of “Boston”; “Green Island” instead of “Troy”).
2. **Intermodal transfer points** (e.g., “South Amboy” or “Camden” for “New York to Philadelphia” routes; “Aquia Creek” for Richmond routes).
3. **Segments of a chain:** a mail route might cover three different rail groups.

Input data.

- *Mail Routes (Table A):* Number of route, Origin, Destination, Distance.
- *Rail Lines (Table B):* YrGroupNumber, Origin_City, Dest_City.

Matching logic (priority order).

1. **Direct match (exact or fuzzy):** Origin/Dest in Table A match Origin/Dest in Table B. Example: “Auburn to Rochester” matches “Auburn” and “Rochester.”
2. **Metro area / suburb match (high-priority fix):** if Table A says “Boston,” Table B might say “Cambridge,” “Chelsea,” or “Somerville”; “Troy” $\leftrightarrow$ “Green Island”; “Columbia, SC” $\leftrightarrow$ “Eau Claire”; “New York” $\leftrightarrow$ “Jersey City,” “Piermont,” or “South Amboy.” Treat these as matches.
3. **Intermodal / water-connection match:** routes involving major bays (Chesapeake, Long Island Sound, NY Harbor) often list the ultimate city (e.g., “Philadelphia”), while the rail line ends at the water (e.g., “Camden” or “Perryville”). Rule:

if a rail line ends at a known ferry point across from the Mail Destination, it is a match.

4. **Multi-segment chain:** if a Mail Route covers a long distance (e.g., “Baltimore to Philadelphia”), look for a series of rail groups that link up to form that path (e.g., Baltimore → Havre de Grace → Elkton → Wilmington → Philadelphia). List all rail groups that form the chain.

Output format. Generate a CSV with these columns:

- Mail_Route_ID;
- Mail_Origin;
- Mail_Destination;
- Matched_Rail_Group_ID (comma-separated if multiple);
- Rail_Endpoints (the city pairs from Table B);
- Match_Type (“Exact,” “Metro/Suburb,” “Intermodal,” “Multi-Segment,” “No Match”);
- Reasoning (explain the geography, e.g., “Rail Group 140 connects Green Island to Schenectady; Green Island is directly across the river from Troy”)

Prompt shown for the 1846 cross-section; analogous prompts were used for the other years through 1870.

Table B.1: Prompt used to extract the mail-route tables from the scanned reports (Google Gemini, 1870 report shown).

Role. You are an expert data-entry specialist and historical-document analyst.

Task. Analyze the provided image (“Railroad service as in operation on the 30th of June, 1870”). Extract all tabular data into a structured JSON format.

Layout rules.

- *State section headers:* there is no “State” column; states appear as centered, capitalized headers (e.g., “MAINE.,” “NEW HAMPSHIRE.”). Apply each state name to every subsequent row until the next state header.
- *Ditto marks:* in the corporate-title column, replace “...do...,” “—do—,” or similar marks with the text from the row immediately above; do not output “do.”
- *Split rows:* some routes have one route number but multiple lines of data for distance or trips (indicated by curly brackets). Treat these as separate rows sharing the same route number and termini.

Column mapping.

1. *Number of route:* extract the number (e.g., “1,” “9a,” “116”).
2. *Termini:* extract the full text (e.g., “Augusta to Skowhegan”); split on “to” to populate Origin and Destination.
3. *Corporate title of company:* extract the company name; apply fill-down if the text is “do” or dots.
4. *Distance:* extract the miles, including fractions (e.g., “70 3/4”).
5. *Total distance in each State:* extract only if present (usually the last row of the state).
6. *Number of trips per week:* extract the number.
7. *Annual pay:* extract the currency amount.
8. *Annual pay in each State:* extract only if present.
9. *Annual cost per mile:* extract the value (e.g., “\$75 00”).
10. *Remarks:* extract any text in the rightmost column.

Output. Return a valid JSON array, one object per route, with fields: State; Number of route; Termini; Origin; Destination; Corporate title of company; Distance; Total distance in each State; Number of trips per week; Annual pay; Annual pay in each State; Annual cost per mile; Remarks (required: State, Number of route, Termini, Distance, Annual pay). Then convert the JSON to a CSV block.