Information and Innovation: Evidence From Railroad Expansion in the Nineteenth Century

Shuting Li*

Abstract: I study the causal effect of information flow on innovation by exploiting the staggered expansion of the U.S. railroad network between 1840 and 1900. Using an event study analysis surrounding the opening of the first railroad in each county, I find that counties where the first railroad was opened generate more patents over the next five years compared to those without one. These patents tend to be more important and impactful, though less novel. The effect is stronger in counties that are more connected to other counties through the railroad network. My findings suggest that improved information exchange efficiency enhances innovation. In my future work, I plan to leverage text-based patent similarities to further substantiate the role of information diffusion as a key channel facilitating innovation activity.

Keywords: Innovation, Information, Knowledge, Inventors, Diffusion.

1 Motivation and Hypotheses

To what extent does the information flow promote innovation? Economic growth theories have long recognized information and knowledge as key drivers of innovation.[1] More recent empirical studies also show that knowledge plays a central role in innovation activities.[2] However, rigorously identifying the causal effect of information has been impeded by two empirical challenges: the unavailability of datasets capturing the information flow, and the difficulty in isolating the specific role of information flow in innovation activities amidst various economic confounding factors. I attempt to address these two identification challenges within the historical context, the expansion of the railroad network across the United States between 1840 and 1900. The historical context provides a well-suited setting for tackling these challenges for three reasons.

First, unlike today when information can be exchanged through various means and traveling is relatively convenient, in the nineteenth century the railroad system was crucial for the exchange of information and knowledge. The United States Postal Department (USPD), which held a

*J. Mack Robinson College of Business, Georgia State University, Email: sli49@gsu.edu

[1]See Change (1990), Weitzman (1998), among others.

[2]See Matray (2021) and Bai, Jin, and Zhou (2023).

monopoly on mail delivery, relied on railroad companies to transport mail along contracted routes known as railroad mail routes. On these routes, postal offices also distributed newspapers and magazines, often with little to no cost, thereby boosting their circulation.[3] Thus, the railroad's role in facilitating the delivery of mail, newspapers, and magazines was pivotal to the dissemination of knowledge.

Second, the emergence of railroads and the improvements in mail distribution significantly improved the efficiency of information exchange. Prior to railroads, mail and goods were transported by horseback or horse-drawn vehicles. Trains, however, were faster, more cost-effective, and could carry larger volumes, even in inclement weather. In addition, improvements in mail sorting and distribution enhanced delivery speed. Initially, mail was sorted only at major post offices. Beginning in the 1830s, sorting occurred on trains during transit. The introduction of Railroad Post Office (RPO) cars further streamlined the process, greatly improving efficiency.[4]

Third, the expansion of the railroad unfolded in several stages, beginning in the Northeast and Mid-Atlantic regions, then extending eastward, and ultimately pushing westward.[5] This gradual development led to a staggered introduction of railroads, causing variations in the efficiency of information exchanges. Examining the timing of the arrival of the railroad in different counties allows me to explore how different counties respond differently to railroad construction that create significant large information shocks to innovation.

The expansion of the railroad network enabled inventors to gain information exposure through various means: mails, including newspapers and magazines, and traveling. The most famous examples of journals include *Scientific American* and *American Inventors*, which published articles on new inventions and details of recently issued patents. Inventors could access comprehensive patent lists and detailed specifications in patent offices and branch offices by traveling via railroad.[6] This increased access to information and knowledge through railroads leads to my first hypothesis: counties with railroads generate more patents than those without.

However, the impact of information flow on innovation quality remains unclear. On the one hand, information and knowledge are diffused along railroads; and knowledge accumulation leads to the production of higher-quality innovations. On the other hand, exposure to periodic information can reduce novelty, as inventors are likely to absorb familiar knowledge rather than generate

[3] Kennedy (1957) estimates that 40 percent of newspapers and magazines traveled essentially postage-free in 1874 because postmasters failed to collect the postage.

[4] The 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.

[5] By 1837, twenty-nine railroads were under construction, and by 1850, most large Eastern cities were connected by rail. National expansion began in the 1850s, reaching the West and establishing overland mail routes to California. In particular, the transcontinental railroad, completed in 1869, reduced travel time from Missouri to California from 16-20 days to just 7-8 days, significantly improving information exchange efficiency.

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

truly novel ideas by themselves (Dai, Donohue, Drechsler, & Jiang, 2023). Therefore, my second hypothesis is that patents generated in counties with railroads are more impactful but less novel compared to those in counties without railroads.

2 Data Collection and Variable Construction

I combine a digital map of the U.S. railroad network with historical U.S. county boundary maps to construct a county-year level panel data spanning the year from 1840 through 1900. My source of railroad data is the historical transportation database created by Jeremy Atack (Atack, 2018). The railroad map includes all constructed railroad segments, along with their respective years of opening, allowing me to identify counties with railroads passing through at the year level and whether the travel between each pair of counties is feasible through the railroad. During the period of 1840-1900, I find that 46% of county-year observations had access to railroads, and the average number of counties connected to a specific county by rail was 1.45.

The key outcome of interest is the patent-based innovation output. I collect patent information for the years 1840-1900 from PatentCity (Bergeaud & Verluise, 2024), which includes all utility patents from this period, along with inventors' names and locations.[7] The first outcome variable is the quantity of patents, the number of patents published each year at the county level. In my sample, the average number of patents generated within a county-year is 4.3.

The second outcome variable is patent quality, measured by patent importance, impact, and novelty.[8] I obtain the patent quality dataset from Kelly, Papanikolaou, Seru, and Taddy (2021). 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 existing patents is more likely to be novel, and a patent greatly influencing the subsequent patents is more likely to be impactful. Based on that, the importance of a patent is measured as the ratio of the forward similarity over the backward similarity.

3 Empirical Test

3.1 Matched sample construction

The historical era provides me with an ideal setting to study the effect of information, but the empirical setup does pose two challenges. First, the construction of railroads was not random: it

[7] At this point, I restrict attention to utility patents because PatentCity covers utilities patents only. The rest of the patents will be included in the later version of the study.

[8] A 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 USPTO in patent documents only after 1947.

is possible that factors such as economic factors could influence both railroad construction and innovation. Second, the treatment effect of railroads on innovation may vary over time, potentially leading to biased estimates in staggered difference-in-difference (DiD) analysis (A. C. Baker, Larcker, & Wang, 2022; A. Baker, Callaway, Cunningham, Goodman-Bacon, & Sant'Anna, 2025).

To address these two concerns, I employ an event study analysis that leverages the year of the opening of the first railroad in each county between 1840 and 1900. Surrounding the opening years, I employ a stacked difference-in-difference framework to compare the changes in innovation output between counties where the first railroad was built and the 'counterfactual' counties with similar characteristics but no railroad, within the same time window.[9] To construct the matched sample, I first identify the years in which any of the counties has the first railroad built and then identify the county-year observations in the five years before and the five years after the event year, i.e., an 11-year event window. Counties with the first railroad built in the event year are flagged as treated counties. Next, I match each treated county in the five years before the opening to a counterfactual county that has similar population size, urban population ratio, and white population ratio, and has no railroad in the 11-year event window.[10] The key identification assumption is that the treated county and the matched 'counterfactual' county would have similar trends in terms of innovation in the absence of the railroad.

3.2 Regression specification and baseline result

Using the matched sample, I compare the change in innovation output between counties with the first railroad opening and 'counterfactual' counties and between the pre- and post-railroad construction period. Specifically, I estimate:

$$Y_{c,t,s} = \beta Treat_{c,s} Post_{c,t,s} + \alpha_1 Treat_{c,s} + \alpha_2 Post_{c,t,s} + Control_{c,t,s} + \gamma_c + \gamma_t + \gamma_s + \epsilon_{c,t,s}, \quad (1)$$

where $Y_{c,t,s}$ is the outcome of interest for county c in year t in event window s, including patent quantity and patent quality. $Treat_{c,s}$ is an indicator variable that equals one if county c has the first railroad opening in the event year. $Post_{c,t,s}$ is an indicator variable that takes a value of one for the years after the first railroad built in county c. 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 $Control_{c,t}$ includes population, urban ratio, and white population ratio. These control variables help absorb the time-varying socioeconomic conditions that

[9] Studies implemented a stacked DiD include Gormley and Matsa (2011, 2016); Cengiz, Dube, Lindner, and Zipperer (2019); Deshpande and Li (2019); Gormley, Jha, and Wang (2024).

[10] I collect county-level population data from the U.S. Censuses of Agriculture and Population. Urban ratio is defined as the total urban population over total population, and the white ratio is defined as the total white population over total population. Between 1840 and 1900 the average population within a county-year is 14,710, the average urban ratio is 6%, and the average white ratio is 77%.

were potentially associated with innovation activity. Since 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 event fixed effect γ_s to mitigate potential omitted variable bias. The county level fixed effect γ_c control for time-invariant omitted county characteristics. The year fixed effect γ_t accounts for transitory economy-wide factors, such as macroeconomic conditions, and any time trend in innovation activity.

The key coefficient of interest is β, which captures the impact of the information flow on innovation output after the construction of the first railroad. If improved information exchange efficiency leads to more innovation, then treated counties should generate more patents after the first railroad than before, and counties should have more patents than counties without the service. Therefore, I expect to see a positive and significant β.

Consistent with my hypothesis, I find that counties with the introduction of railroads generate more patents over the next five years compared to 'counterfactual' counties, and these patents tend to be more important and impactful. However, patents produced following the arrival of the railroads tend to be less novel.

I next adopt a continuous variable, degree centrality $Cen_{c,t}$, to measure the treatment intensity. Degree centrality measures the number of counties connected by railroad links.[11] The rationale is that counties with higher centrality benefit from greater exposure to information and knowledge as a result of connection with more other counties. I regress the innovation output $Y_{c,t,s}$ on $Cen_{c,t,s}$ and the interaction term $Cen_{c,t,s} \times Treat_{c,s}$:

$$Y_{c,t,s} = \beta Treat_{c,s} Cen_{c,t,s} + \alpha_1 Treat_{c,s} + \alpha_2 Cen_{c,t,s} + Control_{c,t,s} + \gamma_c + \gamma_t + \gamma_s + \epsilon_{c,t,s}, \quad (2)$$

where $Cen_{c,t,s}$ is the number of counties county c was linked directly along railroad in year t. I hypothesize that a county with more connections should have more patents, and hence the key coefficient of interest β should be positive and significant. I find that indeed counties with higher degree centrality produce more patents, and those patents tend to be more important, more impactful, but less novel.

3.3 Endogeneity

I perform two additional tests in an attempt to further alleviate the endogeneity concern. First, I use the Postal Act of 1845, an act of Congress that slashed postage rates for letters, as an shock, and see whether reduced communication costs improve innovation output.[12] I find that

[11] In the nineteenth century, overlapping railroads do not connect by default because of differences in railroad gauges and additional transferring cost between different railroads.

[12] For example, prior to the Act sending a letter from Baltimore to New York in 1840 costs 18.75 cents. After the act went into effect, the cost was lowered by 73% to 5 cents.

counties with railroads generate more patents following the Act, and the patents tend to be more important, more impactful, but less novel. These findings indicate that information flow is an important determinant of innovation.

Second, I exclude counties adjacent to treated counties from the counterfactual control groups and repeat the baseline test. The idea is that the control counties that are exposed to spillover knowledge from adjacent treated counties, and the observed effect should be greater if those counties are excluded from control group. Indeed, I find that the observed effect is greater after excluding the adjacent counties.

3.4 Mechanism

If the observed effect is driven by information diffusion along railroads, inventors in counties along the same line tend to be exposed to similar knowledge. As a result, their innovation output should be more closely related. Therefore, I hypothesize that after the railroad's opening, patents from inventors in counties along the same railroad are more similar or connected than those from counties without direct railroad connection.

To test the hypothesis, I calculate cosine similarities between patents based on patent descriptions. I first collect patent text descriptions from USPTO and Google, and then convert each into a 1024-dimensional vector using the Jasper and Stella embedding model by Zhang, Li, Zeng, and Wang (2025). The cosine similarity for each patent pair were calculated using the resulting vectors.[13]

I plan to test this mechanism using a difference-in-differences framework to compare changes in patent similarities between county pairs with railroad lines and those without, before and after railroad construction.[14] Specifically, I estimate:

$$Similarity_{c,k,t} = \beta_t Route_{c,k,t} \times Post + \alpha_1 Route_r + \alpha_2 Post + \boldsymbol{X_{c,k,t}} + \gamma_{c,k} + \gamma_t + \epsilon_{c,k,t}, \quad (3)$$

where the dependent variable $Similarity_{c,k,t}$ is average patent similarities between county c and county k in year t. $Route_{c,k,t}$ is a dummy variable that takes the value of one if county c and county k are directly connected. $Post$ is the indicator variable that takes the value of one for the years after the initiation $Route_{c,k,t}$ following the completion of the railroad construction. I include county-pair fixed effect $\gamma_{c,k}$ and year fixed effect γ_t. β_t estimates the effect of information flow on between-county patent similarities, and I expect to see a positive and significant β_t.

[13] The high-dimensional embeddings created by Jasper and Stella model capture the semantic meaning and the context of text, making it more effective to reflect similar ideas even when different words are used in the patent description.

[14] Due to the large-scale patent-level data, the results has not been generated yet.

4 Future Work

In this section, I list the potential work that I will explore in the future and the results will be reflected in the later version of the study.

First, I will examine how the establishment of USPD's railroad mail routes affects the innovation output, as the establishment offers three advantages for my study. First, a comparison of innovation output linked through railroads versus mail routes allows me to understand whether the observed effect is driven by information channel. Second, USPD records the contract cost and weekly trips per route, allowing me to analyze the effect of information using treatment intensity while controlling for information costs. Third, combining the mail routes with the Postal Act of 1845 offers a cleaner framework to study how postal rate reductions affect innovation output.

Second, I will examine co-inventorship, exploring whether counties connected by railroads are more likely to produce collaborative patents and if these patents are of higher quality. The idea is that the railroad enabled inventors to communicate via mail or in person, facilitating co-inventorship.

Third, I will include waterways and telegraphs to capture additional information flows and evaluate the marginal impact of railroads relative to these other communication modes.

Fourth, I plan to examine the knowledge spillover effect from counties with railroads to neighboring counties without them. To quantify this, I will exclude the original treated counties and consider the adjacent counties as the new treated ones. I expect these neighboring counties to produce more patents with improved quality.

Fifth, I will analyze cargo-only and passenger-only railroad lines to assess the marginal impact of face-to-face interactions versus mail exchanges on innovation output. Cargo cars transport mail, while passenger coaches enable in-person travel, allowing inventors to exchange ideas both directly and via mail. If counties along passenger-only lines have more patents than those along cargo-only lines, it would suggest that face-to-face interactions marginally boost innovation.

Sixth, I intend to examine the types of patents generated by each county to disentangle the information channel from the market access channel. The idea of the market access channel is that the increased innovation output is a response to the expanded market access facilitated by railroads. Specifically, the expansion of the railroad network lowers shipping costs for goods and products and expands the market for them by connecting to more regions. As a result, companies face increased demand to enhance productivity and efficiency, prompting inventors to respond to these demand by developing innovations, such as exploitation-type patents. If a county generates more patents focused on productivity enhancement (i.e., exploitation-type), the observed effect is likely driven by the market access channel. Conversely, if a county produces patents centered on exploration, the effect is more likely attributed to the information channel.

5 Contribution

My proposal attempts to contribute to several strands of literature. It is most directly linked with a growing literature that evaluates the effects of information and knowledge on innovation. Previous literature document geographical area borders and proximity are important barrier for knowledge diffusion (Jaffe, Trajtenberg, & Henderson, 1993; Thompson & Fox-Kean, 2005; Matray, 2021; Bai et al., 2023), whereas a number of studies document innovation spillovers occur with information and knowledge transferring, such as inventor mobility (Hombert & Matray, 2017; Bernstein, Diamond, Jiranaphawiboon, McQuade, & Pousada, 2022) and merge and acquisition (Li & Wang, 2023). This proposal attempts to contribute the literature by exploiting the variation in information flow resulted from railroad construction.

Second, this proposal attempts to contribute to the literature on the determinants of innovation. Previous studies recognize the importance of a wide range of factors, such as analyst coverage(He & Tian, 2013), stock liquidity (Fang, Tian, & Tice, 2014), and whether the firm choose to go public (Bernstein, 2015).[15] This proposal attempts to contribute to this large body of research by providing evidence on the causal effect of information exchange on innovation activity.

Third, this proposal attempts to contribute to our understanding of the the knowledge accumulation and production. Previous studies show that knowledge is transferred through coauthorship (Azoulay, Graff Zivin, & Wang, 2010)[16], within institution (Furman & Stern, 2011), and the access to prior knowledge is particularly crucial for idea production (Williams, 2013; Galasso & Schankerman, 2015; Iaria, Schwarz, & Waldinger, 2018). This proposal contribute to the literature by showing how railroad network stimulate the knowledge diffusion and technological development progress. In particular, I focus on how railroad networks promote the collaborations between inventors.

References

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.

Azoulay, P., Graff Zivin, J. S., & Wang, J. (2010). Superstar extinction. *The Quarterly Journal of Economics*, *125*(2), 549–589.

Bai, J., Jin, W., & Zhou, S. (2023). Proximity and knowledge spillovers: Evidence from the introduction of new airline routes. *Management Science*.

[15] Other factors include the health of financial system(Nanda & Nicholas, 2014).

[16] They show that when academic superstars die unexpectedly, the research productivity of collaborators falls significantly.

Baker, A., Callaway, B., Cunningham, S., Goodman-Bacon, A., & Sant'Anna, P. H. (2025). Difference-in-differences designs: A practitioner's guide. *arXiv preprint arXiv:2503.13323*.

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.

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. (2015). Does going public affect innovation? *The Journal of finance*, *70*(4), 1365–1403.

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.

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.

Change, E. T. (1990). Endogenous technological change. *Journal of Political Economy*, *98*(5), 2.

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.

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.

Fang, V. W., Tian, X., & Tice, S. (2014). Does stock liquidity enhance or impede firm innovation? *The Journal of Finance*, *69*(5), 2085–2125.

Furman, J. L., & Stern, S. (2011). Climbing atop the shoulders of giants: The impact of institutions on cumulative research. *American Economic Review*, *101*(5), 1933–1963.

Galasso, A., & Schankerman, M. (2015). Patents and cumulative innovation: Causal evidence from the courts. *The Quarterly Journal of Economics*, *130*(1), 317–369.

Gormley, T. A., Jha, M., & Wang, M. (2024). *The politicization of social responsibility* (Tech. Rep.). National Bureau of Economic Research.

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.

He, J. J., & Tian, X. (2013). The dark side of analyst coverage: The case of innovation. *Journal of Financial Economics*, *109*(3), 856–878.

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.

Iaria, A., Schwarz, C., & Waldinger, F. (2018). Frontier knowledge and scientific production:

Evidence from the collapse of international science. *The Quarterly Journal of Economics*, *133*(2), 927–991.

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.

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.

Li, K., & Wang, J. (2023). Inter-firm inventor collaboration and path-breaking innovation: Evidence from inventor teams post-merger. *Journal of Financial and Quantitative Analysis*, *58*(3), 1144–1171.

Matray, A. (2021). The local innovation spillovers of listed firms. *Journal of Financial Economics*, *141*(2), 395–412.

Nanda, R., & Nicholas, T. (2014). Did bank distress stifle innovation during the great depression? *Journal of Financial Economics*, *114*(2), 273–292.

Thompson, P., & Fox-Kean, M. (2005). Patent citations and the geography of knowledge spillovers: A reassessment. *American Economic Review*, *95*(1), 450–460.

Weitzman, M. L. (1998). Recombinant growth. *The Quarterly Journal of Economics*, *113*(2), 331–360.

Williams, H. L. (2013). Intellectual property rights and innovation: Evidence from the human genome. *Journal of Political Economy*, *121*(1), 1–27.

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