Proposal for the NBER Productivity, Innovation and Entrepreneurship Program's Post-Doctoral Fellowship in Innovation Policy

Elisabeth Ruth Perlman

I seek to understand how information networks affect innovative activity by examining the geography of innovation in nineteenth-century America. I am currently working on three papers: The first two papers examine transportation and urbanization, and are pieces of my dissertation. The last, which I would investigate as a post-doctoral fellow at the NBER, uses the knowledge that can be gained from the text of patents with modern data-mining techniques to understand the historical record. In the first paper, I test the Sokoloff (1988) hypothesis that increasing market access through the spread of transportation infrastructure leads to an acceleration of innovation, using the spread of the railroad across the United States as an example. In the second, I map the relationship between patents and population for all counties in the United States, asking whether population size or density creates innovative agglomerations, and investigating the relationship between transportation spread and urbanization. For the third paper, I plan to use textual data to explore the generation and spread of interrelated ideas across time and space. To accomplish this, I have compiled a full-text, searchable database of all patents issued between 1836 and 1897. Patents are by no means a perfect measure of innovative activity (Trajtenberg, 1990; Moser, 2004), but their paper trail is extensive, including descriptions of each invention. These records allow for their systematic use as evidence for historical innovation.

I will focus on using the full text of patents, along with the inventors' locations, to examine if the railroad changed how information spread once a place became connected to the larger network. It has historically been difficult to identify how earlier ideas are related to later ones or to measure the speed of ideas spreading from one place to another, since ideas cannot be observed directly. However, when people write about technology as they do in a patent their sentence structure and word usage can help reveal these underlying information flows. Patents drawing on common sources will use shared language by referring to the same concepts and antecedents. These commonalities between patents can be identified through searching for sequences of words and phrases (n-grams). When cross-referenced with the time series and geographic information about patents, the spread of these n-grams should reveal how well connected disparate places are to the same information scores, how quickly, and over what distance, innovation transmits. Following Takahashi et al. (2012), computer scientists studying natural language processing have been investigating tools to determine when particular topics become popular. Much as Packalen and Bhattacharya (2012) examines the use of new phrases in patents 1920-2010 to get a sense of what fields are newly important, this paper will locate centers of innovation in particular technologies by mapping the spread of n-grams as they appear in patents. I will detect words that are "bursty" that is, are used more than expected, and trace the geography of these bursts.

These texts will allow the estimation the effect of centrality on innovation. While it is presumed that increasing the flow of information between places decreases the advantage of being in any particular location, it may also be that greater information access increases the advantage of those centrally located in the network. Given the general diffusion of innovation caused by the spread of the railroad this seems unlikely, but the magnitude of the effect in either direction can be estimated with my data set.

The production of a novel, patentable innovation requires knowledge of the state of the art in order to avoid duplicating a known improvement. Patentable innovations thus require knowledge of where the technological frontier is: which problems are interesting and how these might be solved; and which lines of research have been or are being explored. While duplicative innovations may well enhance productivity, they cannot be patented. Contact with the state of the art requires information to be available and accessible. One large barrier to accessibility is filtering; while a library may contain the information a potential inventor needs to know, locating relevant work has always presented a challenge. Contact with others in the same field, leading to social learning, has the effect of making information more accessible, in part by improving filtering. This can take the form of Marshallian "information in the air" or information gleaned from individual connections. In the modern period, questions about the co-location of technologies have followed the work of Jaffe et al. (1993), which showed that patents in the same city are far more likely to cite each other, although later work has shed doubt on the generality of these claims.

Despite their flaws as a way to measure innovation, the text of patents provides an interesting source for exploring innovation because of their limited scope and obvious interconnectedness. There are many other automated methods, in addition to looking for word burst, for exploring the content of historical texts. These have the potential to discover dispersed patterns in the patent data that would be otherwise unobservable. A vector representing the words used in the text (bag of words) is often revelatory, and texts can be easily compared to each other by taking the dot product of these vectors or vectors weighted for frequency of word use (see Egesdal et al. (2014) for an example). One can also look at text for new n-grams, in a naive way, was well as looking for bursty n-grams. Texts can also be clustered according to common word use, often with each cluster representing a topic (see Dittmar (2011) for an example), this highlights common topics in otherwise unexplored text, and can be used to understand how often topics are written about over time. More formal topic models can also be used. One of the most common techniques used is document classification, which using general features of a text to assign it to a category (this is most often used for sentiment analysis). This paper will combine simple bag of words comparisons with detection of new n-grams, with the intention of exploring other techniques as well.

Further, I will combine what is gathered from the text of patents with data available on Google books, which can be used to see where the same n-grams appear in general discourse (see Koike et al. (2013) for an example of cross-medium topic identification). This geographically-linked text database will enable a much closer look at the relationship between agglomeration, innovation, and the transmission ideas.

In addition, these data on the full text of patents can also be used to explore the details of the locational specificity of patenting. I will analyze data for patterns of urban/rural specialization and industrial clustering. The role of general-purpose or foundational technologies in creating innovation can also be addressed by looking at how related patents cluster around, or spread outward from, the point of invention. Furthermore, one could look at the relationship between industrial agglomerations, innovation, and long-run viability of cities. Areas highly concentrated in one industry make productivity gains through agglomeration forces, but are also vulnerable to industry-specific shocks, which might lead to a troubling relationship between productivity and robustness. Consider the current plight of Pittsburgh or Detroit: innovation concentrated in one sector may make a city vulnerable to that sector's fluctuations. These data can be used to identify which areas are persistently innovative, and which are only innovative in specific sectors, or during particular times.

I am investigating how innovation responds to increases in connectedness, in particular how places on the periphery respond to connections to a larger markets through transportation improvements. As part of my dissertation I revisit Sokoloff (1988), who argued that between 1805 and 1835 counties along the newly-built canals in the eastern United States, particularly the Erie, saw a sharp increase in patenting activity. I construct a dataset to include the data used in Sokoloff (1988), linking all patents issued between 1790 and 1836 to the counties in which the named inventors resided. This is linked to the patent data assembled by Tom Nicholas (Akcigit et al., 2013) and to Jeremy Atack's transportation data (Atack, 2013). With the greater time granularity my patent data provide, one can see that the gradual increase in patenting activity happens after the arrival of the canal or the railroad. Over the course twenty years following the arrival of a railroad in a county the number of patents per capita increased to approximately twice the pre-railroad amount. This effect is dominated by the area of a county that is close enough make a round trip to the new transportation within a day, and not by area further away, even controlling for urbanization. As the increase is not a sudden shock, but rather a slow adjustment to a new state and to increased economic growth, this suggests that access to a larger market spurs development, which in turn increases innovation in the area.

Also as part of my dissertation, I examined the relationship between innovation and population distribution. The largest cities always patent at approximately the same rate, per capita as the other large cities in that time. This suggests either equilibrium among these large cities or limits to the economies of scale available in innovation at any given point in time. However, relationship between population and patenting does not remain constant; in later years, the relationship has a steeper slope—a unit increase in population corresponds with a larger increase in patenting in 1870 than it does in 1840.

The work that I propose will help uncover how information flows, something that is very hard to observe, change the distribution of innovation. It will show how increased transportation relate to changes in information diffusion, and how the that people are working on change with transportation links. In addition, future projects with this data will give us a new way to understand the relationship between innovation diversity and city prosperity.

References

Akcigit, U., Kerr, W. R., and Nicholas, T. (2013). The Mechanics of Endogenous Innovation and Growth: Evidence from Historical U.S. Patents.

Atack, J. (2013). On the use of geographic information systems in economic history: The american transportation revolution revisited. *The Journal of Economic History*, 73:313–338.

Dittmar, J. E. (2011). Information technology and economic change: The impact of the printing press. *The Quarterly Journal of Economics*, 126(3):1133–1172.

Egesdal, M., Gill, M., and Rotemberg, M. (2014). Here comes the sunshine act.

Jaffe, A. B., Trajtenberg, M., and Henderson, R. (1993). Geographic localization of knowledge spillovers as evidenced by patent citations. *The Quarterly Journal of Economics*, 108(3):577–98.

Koike, D., Takahashi, Y., Utsuro, T., Yoshioka, M., and Kando, N. (2013). Time series topic modeling and bursty topic detection of correlated news and twitter. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 917–921. Asian Federation of Natural Language Processing.

Moser, P. (2004). Determinants of Innovation Evidence from 19th Century World Fairs. *The Journal of Economic History*, 64(2):548–552.

Packalen, M. and Bhattacharya, J. (2012). Words in Patents: Research Inputs and the Value of Innovativeness in Invention.

Sokoloff, K. L. (1988). Inventive Activity in Early Industrial America: Evidence From Patent Records, 1790–1846. *The Journal of Economic History*, 48(04):813–850.

Takahashi, Y., Utsuro, T., Yoshioka, M., Kando, N., Fukuhara, T., Nakagawa, H., and Kiyota, Y. (2012). Applying a burst model to detect bursty topics in a topic model. In *Advances in Natural Language Processing*, pages 239–249.

Trajtenberg, M. (1990). A Penny for Your Quotes: Patent Citations and the Value of Innovations. *The RAND Journal of Economics*, 21(1):172.