Research Policy xxx (2013) xxx–xxx

Contents lists available at ScienceDirect

Research Policy

journal homepage: www.elsevier.com/locate/respol

Getting patents and economic data to speak to each other: An 'Algorithmic Links with Probabilities' approach for joint analyses of patenting and economic activity☆,☆☆

Travis J. Lybbert[a,*], Nikolas J. Zolas[b]

[a] Department of Agricultural & Resource Economics, University of California, Davis, United States
[b] Center for Economic Studies, United States Census Bureau, United States

ARTICLE INFO

Article history:
Received 24 August 2012
Received in revised form 1 May 2013
Accepted 4 September 2013
Available online xxx

Keywords:
Patents
Trade
Industry
Concordances
Technology
International Patent Classification

ABSTRACT

International technological diffusion is a key determinant of cross-country differences in economic performance. While patents can be a useful proxy for innovation and technological change and diffusion, fully exploiting patent data for such economic analyses requires patents to be tied to measures of economic activity. In this paper, we describe and explore a new algorithmic approach to constructing concordances between the International Patent Classification (IPC) system that organizes patents by technical features and industry classification systems that organize economic data, such as the Standard International Trade Classification (SITC) and the International Standard Industrial Classification (ISIC). This 'Algorithmic Links with Probabilities' (ALP) approach mines patent data using keywords extracted from industry descriptions and processes the resulting matches using a probabilistic framework. We compare the results of this ALP concordance to existing technology concordances. Based on these comparisons, we discuss advantages of this approach relative to conventional approaches. ALP concordances provide a meso-level mapping to industries that complements existing macro- and firm-level mappings – and open new possibilities for empirical patent analysis.

1. Introduction

International technological diffusion is an important driver of technological change, which is in turn a key determinant of cross-country differences in income and economic growth (Romer, 1990; Aghion and Howitt, 1992; Grossman and Helpman, 1991; Keller, 2004). International trade and foreign direct investment are often considered to be key catalysts of technology transfer (Coe and Helpman, 1995; Eaton and Kortum, 2002; Branstetter et al., 2006; Acharya and Keller, 2009), but directly studying this process is often hampered by the fact that measuring transferred technology empirically is challenging. While patent data often serve as useful proxies for technological change (Griliches, 1990; Basberg, 1987) and diffusion (Jaffe et al., 1993), fully exploiting patent data in economic analyses would require that patents be linked to economic activity at a level of disaggregation that allows for different technological, industrial and spatial patterns. Such a detailed link between technological and economic activity would further improve our assessment of policies that aim to promote innovation, as well as assess the relationship between technological change and economic development. In this paper, we propose an algorithmic approach to constructing such a link.

Patent statistics have frequently been used as both technological and economic indicators due to the widespread availability of patent data and the assumption that patents reflect direct inventive activity and innovation. Basberg (1987) describes how patents have been incorporated into innovation models to measure technology diffusion and to evaluate the output of research activity. In a similar survey, Griliches (1990) documents how patents have been used as economic indicators for inter alia R&D output, stock market activity and total factor productivity. Within this literature, however, the empirical validity of patents as technological or economic indicators remains a matter of debate largely because of concerns about how patents are used, enforced and valued differently across

☆ *Note*: The ALP concordances described in this paper can be downloaded from the WIPO website at http://www.wipo.int/econ_stat/en/economics/publications.html.

☆☆ We thank Prantik Bhattarchayya for superb research assistance. We are grateful for the assistance and guidance provided by researchers and programmers at the World Intellectual Property Organization (WIPO), including Carsten Fink, Hao Zhou, Christophe Mazenc, Sacha Wunsch-Vincent, and others. We thank seminar participants at WIPO and Morrison Foerster and attendees at the 2012 "Patent Statistics for Decision-Makers" Conference at the OECD-Paris. We also acknowledge the financial support the project received from the National Science Foundation. All opinions and views expressed are those of the authors and do not represent those of the NSF or the U.S. Census Bureau. All results have been reviewed to ensure that no confidential information is disclosed.

* Corresponding author.
E-mail address: tlybbert@ucdavis.edu (T.J. Lybbert).

0048-7333/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.respol.2013.09.001

different industries, jurisdictions and time periods. We believe that more disaggregated analyses of patent statistics – particularly when matched with equally disaggregate economic data – will alleviate some of these concerns and open new research possibilities.

In general, there are three levels at which patents can be linked to economic activity. At the coarsest macro-level, aggregate patent data taken from a specific country in a specific year can be associated with aggregate economic data, respectively. Linking patent and economic data at this aggregate level is based simply on the country-year unit of analysis and has enabled research on questions such as measuring the rate of innovation (Porter and Stern, 2000), a country's innovative capacity (Furman et al., 2002) and the effects of patent harmonization (McCalman, 2001). Analyses of foreign patent flows and economic activity (Eaton and Kortum, 1996; Xu and Chiang, 2005; Falvey and Foster, 2006; Harhoff et al., 2009) is similarly based on an aggregate association of patents to economic data through a shared country-year unit of analysis.

At the finest level, patents and economic activity can be linked at the firm-level. While this micro-linkage between patent and economic data enables rigorous and insightful research on patenting as part of firm-level strategies (Brouwer and Kleinknecht, 1999; Austin, 1993), constructing and maintaining such a firm-level database requires substantial effort, is only feasible for a fraction of the firms represented in patent databases, and may miss broader considerations regarding related products, competitors and industrial dynamics. Although progress will continue to be made at this level, these limitations constrain our ability to link patents to economic activity at the firm-level, especially in emerging economies where firm-level data is relatively sparse.

Between these macro- and micro-level linkages is a meso- or industry-level linkage that associates patents and economic data based on the domain of goods and services they represent. At this level, patents on biomedical and semiconductor inventions, for example, are linked to industry or product classes that use biomedical and semiconductor inventions, respectively. We argue that a robust industry-level linkage – perhaps in conjunction with macro- and micro-level analyses – will enable researchers to better understand the relationship between patenting and economic activity over time and across space and technology classes. Most industry-level linkages are based on concordances. For example, the Yale Technology Concordance (YTC) (Kortum and Putnam, 1997) links the International Patenting Classification code (IPC) to the Canadian Standardized Industrial Classification system. Thus, with the YTC a researcher can link patent data organized by IPC, country and year to the value of production organized by Canada SIC, country and year. Unfortunately, conventional concordance approaches like the YTC suffer from a host of flaws that limit their usefulness in empirical research. After describing these limitations, we propose new methods for constructing concordances and, thereby, industry-level linkages between patent and economic data. These methods use data mining and probabilistic matching to build links that can be applied broadly or narrowly across time and space, can be easily updated, and can create direct linkages between patent data and a variety of economic classification schemes.

We refer to the general approach we develop in this paper as an Algorithmic Links with Probabilities (ALP) approach to constructing concordances. This approach identifies patents in the PATSTAT database that contain keywords extracted from industry classifications in the text of the title and abstract. Tabulated by IPC code, these retrieved patents reveal frequency matches between the industry and IPC classifications. We then process these frequencies to generate a probabilistic mapping that works in two directions: from IPC to the industrial classifications and vice versa. Researchers can use these direct ALP concordances for industry and technology-level analyses of the relationships between patents and economic activity organized by different classification schemes such as SITC, ISIC, North American Industry Classification System (NAICS), and Harmonized System (HS). Given that these methods require minimal manual or subjective intervention, the concordances they generate are also easy to update and refine when new patent data becomes available or when classification systems undergo revisions as they frequently do.

After providing a brief background of related patent concordance research, we discuss the prevailing IPC concordances in some detail and describe their limitations when applied to economic data. We then describe our ALP approach to constructing more useful concordances and generate IPC concordances for both trade (SITC) and industry (ISIC) classification schemes. To test our approach, we compare the ALP concordance with two prevailing concordances, including the YTC.

2. Background

Patents are a potentially powerful data source for technology and innovation analyses because the patents themselves contain a wealth of information, including the names of the inventee, date, prior art, technologies used, as well as a full description of the embedded technology with numerous figures and references. Recently, there has been a large push initiated by the private sector to develop novel ways of analyzing, organizing and making this patent information accessible to firms interested in exploiting or diversifying their patent portfolios and formulating R&D strategies (Trippe, 2003; Moehrle et al., 2010). This form of patent analysis – called "patinformatics" – aims to reveal relationships between individual patents and broader technological fields in order to inform commercial, legal and policy decisions and includes grouping similar concepts and technologies, creating patent landscape maps, tracking the evolution of these maps over time, and analyzing and interpreting citation networks. These approaches typically use recent developments in text analysis and text clustering software, and then use the findings from these programs to create different visualization and mapping schemes (see Yang et al., 2008, for a good overview of the various softwares that exists). Among some of the studies that use patent-based indicators and relate it to economic phenomena are the numerous patent citation studies used to assess everything from knowledge spillovers (Jaffe et al., 1993; Jaffe and Trajtenberg, 1999), firm evaluation (Hall et al., 2005) and institution types (Trajtenberg et al., 1992; Jaffe et al., 1998). Other studies have used patent-based indicators by tabulating IPCs to assess the networks of technologies (Leydesdorff, 2008) and quantifying inventor competence (Moehrle et al., 2005).

While the methods we develop are conceptually similar to these tools and could ultimately provide a valuable economic layer to patent landscapes, networks and other patinformatic analyses, the ALP concordances we construct are designed to go beyond descriptive analysis and to enable more rigorous econometric analysis at the industry-level. By doing this, we continue to build on other efforts to link patent and economic data through technology-industry associations. While these industry-level linkages are facilitated by the fact that the IPC and economic classification systems share a detailed hierarchical structure, they are complicated by the fact that these classification systems are motivated by different objectives. Whereas economic classification systems are intended to disaggregate goods and services into meaningful and related sub-groups, the IPC system is intended to facilitate the patent examination process by enabling patent examiners to precisely identify the novel technical features of the disclosed invention and to define the prior art against which they can assess novelty. Since goods or services in very different economic classifications can use the same technical feature (e.g., an electronic motion control device may be used in washing machines and satellites), this difference in intended usage implies that linking patents

Please cite this article in press as: Lybbert, T.J., Zolas, N.J., Getting patents and economic data to speak to each other: An 'Algorithmic Links with Probabilities' approach for joint analyses of patenting and economic activity. Res. Policy (2013), http://dx.doi.org/10.1016/j.respol.2013.09.001