Open vs. closed innovation: using online network data to measure innovation

Benjamin Snow and Oliver Bott

24 October 2014

============

0.1 Introduction

Policy makers worldwide have a profound interest in innovation for its significance for economic development and prospertiy. Taylor (2004) views innovation as "the driving force behind modern economic growth, relative industrial power, and competitive advantage" (p.222). Numerous studies, for example the Innovation Union Scoreboard[1] and the OECD Science, Technology and Innovation Scoreboard[2], have attempted to measure and compare the innovation performance on the national level. However, until today most examinations of innovation have put their analytical emphasis on national level patent data, relying on this form of registering of proprietary data as a means of measuring innovation. This leaves largely unexplored other, more open measures of innovation. The relatively recent emergence of network-based research systems offer new and potentially more instructive metrics by which to measure innovation, compared to protected closed knowledge.

This research project examines the potential preferability of using collaborative online network data on the city-level as an innovation indicator. By doing so, this work will critically assess current innovation indicators in the hope of offering new alternatives for measuring and understanding innovation. Since Freeman and Soete (2009), among other scholars, called for the continuous improvement of innovation measurement, this work seeks to go beyond the widespread use of patent data to contribute to the refinement of innovation indicators, and the field as a whole.

0.2 State of the Field

Defining Innovation

Using a rather grand view in his understanding of innovation, Schumpeter (1942) sees innovation as "a process of industrial mutation, that incessantly revolutionizes the economic structure from within". In a more understated characterization, Smith (2005) defines innovation as "the creation of something qualitatively new, via processes of learning and knowledge building. It involves changing competences and capabilities, and producing qualitatively new performance outcomes" (Smith 2005, 149). While it is widely accepted that innovation can take many forms, e.g. product, process, marketing and proccess innovations, Frankelius (2009), in his extensive literature review of innovation studies, criticizes the widespread underlying assumption that innovation is limited to technological innovation. While accepting Frankelius (2009)'s critique of innovation as taking place outside of the technological realm, for the purpose of this study technological innovation, and specifically software innovation, will be the primary focus following relatively closely to Smith (2005)'s definition.

Measuring Innovation

The frequent technological focus when measuring innovation can partly be explained by the difficulties associated with innovation's measurement. Smith (2005) notes the measuring challenge, as innovation is by definition a novelty and thus commensurability is a demanding task. For these reasons innovation has traditionally though controversially been measured by looking either at its inputs, outputs, our throughputs.

[1] For the latest edition see [http://ec.europa.eu/enterprise/policies/innovation/policy/innovation-scoreboard/index_en.htm].

[2] For the latest edition see [http://www.oecd.org/sti/scoreboard.htm].

Attempting to measure patents by inputs often focuses on resources, such as personnel and equipment allocated to R&D, which Freeman and Soete (2009) notes is often overestimating innovation in research and development by including the routine with the novel. Freeman and Soete (2009) compares this with output oriented measures, which are often based on what he concludes are the already inadequate measures such as GDP.

An indicator most often found in innovation research is patent data (see Taylor 2004). A patent is a "temporary legal monopoly granted by the government to an inventor for the commerical use of his or her invention [Taylor (2004), 229). A patent constitutes a property right awarded when an invention is shown to be non-trivial, useful, and novel (Taylor 2004, 230). Patents were first used to measure demand-side determinants of innovation, and have been used in the analysis of innovation activity for over three decades [Taylor (2004), 230). While long hindered by technological limitations requiring labor intensive patent analysis, recent improvements due to machine-readable patent data have spurred recent econometric analysis, as did for example Hall, Jaffe, and Trajtenberg (2005) in their large-scale patent analysis. Taylor (2004) uses patent data taken from 1963 to 1999 in six different industries and their future citation levels and uses Ordinary Least Squares (OLS) model to test what he terms the 'industry-innovation assumption'.

Limitations of Patent Data

Despite the usefulness of patent data, Taylor (2004) found several limitations. In addition to the 'classification problem' related to assigning a specific industries to patents, patents may vary widely in significance, both technically and economically (Taylor 2004, 231). Most significantly for the purpose on this study, Taylor (2004) as well as Pakes and Griliches (1980) found that "patents are a flawed measure particularly since not all new innovations are patented and since patents differ greatly in their economic impact" (Taylor (2004), 378). Thus, while for some considerable time patents have been considered to be the most effective proxy with which to measure innovation, even recent studies have begun to examine alternatives due to patents limitations in measuring innovation. This is why for example Taylor (2004) also used publication data and the number of their citations as an innovation proxy. Still, both data on patents and academic publishing include only proprietary or closed forms of innovation.

Using Network Data as Innovation Indicator

Current developments in research indicate that "characteristics that were important last century may well no longer be so relevant today and indeed may even be positively misleading" (Freeman and Soete 2009, 3). Today we witness a shift away from the belief that innovation only occurs in professional R&D labs, a change towards what Freeman and Soete (2009) calls "research without frontiers" (p.13). Even though networks and research collaborations become increasingly important, there have been relatively few studies focusing on network data (see for example Breschi and Malerba 2005). Even where research networks have been analysed, the focus is too often on economically useful knowledge (see Acs, Anselin, and Varga 2002). Other studies focusing on research networks focus on other protected collaborative networks (see Ponds, Van Oort, and Frenken 2010). In an exception to this standard, Senghore et al. (2014) and team attempt to answer whether social network statistics act as indicators of innovation performance within a network, and which statistics could predict innovation performance. Using Gnyawali and Srivastava (2013)'s model on cluster and network effects to analyze miltipartite social networks at mass collaboration events, gathering their data from NASA's International Space Apps Challenge. They use graph theory models constructed from affiliation networks in the same challenge over two different years, where 63 curated projects represent nodes, and submitted solutions represent edges, finding (preliminarily) that distributions likely correlate to key aspects, thought at the point of publication the manner in which correlation is measured is not mentioned (Senghore et al. (2014)).

Since Freeman and Soete (2009) among others called for the continuous improvement of innovation measurement, this work seeks to contribute to the refinement of innovation indicators. The purpose of this study is to explore the conceptual and statistical viability of a new metric by which we can measure innovation.

In light of the above mentioned state of innovation research we want to examine the following research question:
To what extent can open innovation network data add to the measurement of innovation performance? Technological advances related to the increasing use of the internet and open research

platforms like GitHub, we want to explore whether open knowledge networks can help refine currently limited innovation performance measurements.

1 Methodology

To examin open network data against patent data, this study will rely largely on two data sets and use the statistical tool R (R Core Team 2014) for the data analysis.

API network data

The first data set is obtained by using the Application Programming Interface data for open networks. To examine open data innovation, data will be obtained from the the git repository web-based hosting service GitHub[3]. GitHub is a web-based hosting service used for collaborative research. Its use source code management make it a commonly used software development collaboration tool. Since most of the repositories are openly accessible one can use API tools to track the popularity of repositories, measured by the repository user counts. The R (R Core Team 2014) packages Wickham (2014), Wickham and Francois (2014) and Couture-Beil (2014) allow us to compile data on the user counts and locations associated with different repositories. We are working on the underlying assumption that work done in GitHub repositories is openly accessible innovation.

For the analysis, first we will create location vectors for different cities with the `locations` code. Since many GitHub users can be located, we will identify different open innovation clusters, for example Berlin, New York and San Francisco. In addition, we use the `vector()` code to get information on user counts, focusing on repositories with more than 20 or so followers. By combining locations and user counts data by using `data.frame()`, we will be able to construct data sets for different location clusters and user count numbers, where users of highly relevant repositories are located.

Closed innovation OECD patent data

For closed innovation we use city-level patent data, taken from the Organization for Economic Co-operation and Development[4]. The R (R Core Team 2014) package (Blondel 2014) necessary for obtaining the OECD dataset. Patent Cooperation Treaty (PCT) patent data are used to track internationally patentend inventions. For our preliminary sample, we took patent data of 20 cities overall, ranging from six different countries, including their country level patent data, for general comparison over the time period 2000 until 2012. We pull the OECD data from the online database using (Blondel 2014). As we are interested in patent data, we work with data indicating the PCT patent applications per 10,000 inhabitants. From the same database, we also use GDP per capita data and environmental data, as other variables which could prove significant in explaining differences in innovation. The OECD data will be in a format that allows us to locate patent activity to individual cities.

The code to pull OECD into R is shown below:

To clean this data for effective statistical comparison, the deletion of several empty rows, and the renaming of relevant variables with their city name rather than an attributed code, for clarity. Past this, all datasets must be effectively merged.

To further clean up the table we rename the city and country IDs using this code:

Descriptive Statistics of OECD data set

To give a brief overview of the OECD data already obtained, the summary statistics give some first impression of the OECD city data set.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.24    1.41    3.45    3.44    5.40    8.81
```

[3] Online accessilbe on [https://github.com/].
[4] Online accessible on [http://stats.oecd.org].

In the data set, the mean for PCT patent application is 3.44 per 10,000 inhabitants. The histogram indicates that the mean is not very meaningful as there are many locations with patent applications much lower and some with around 5.5 patent applications. This finding supports our claim that on the regional level there are great variations in patent activity, which supports our approach of focusing on city-level data. The histogram indicates a weak relationship between GDP and patent intensity.

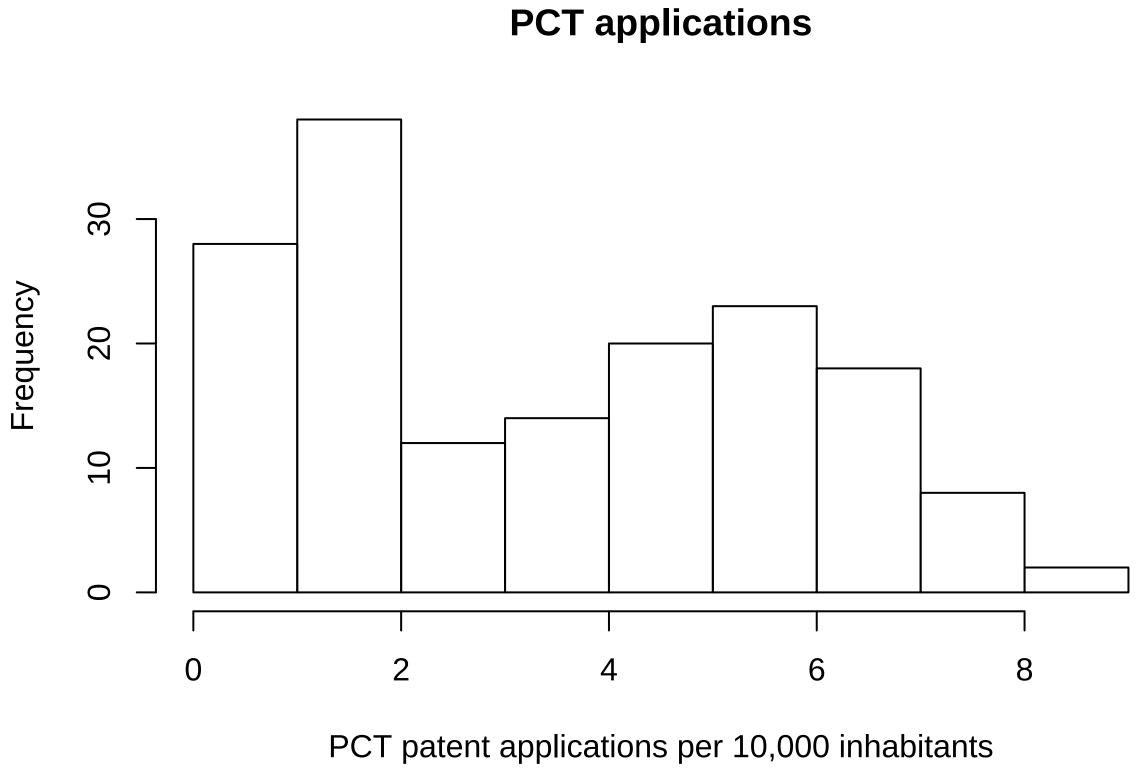

Figure 1: plot of chunk unnamed-chunk-7

Statistical Model

On the type of analysis and question, this study will use an ordinary least squares (OLS) model by using `plot` and possibly `rcorr( , type="pearson")` to examine the relationship between patent and open data. If there is a relationship this would presumably demonstrate that open data shows the same innovation patent data show, but shows the 'throughput' of innovation rather than the 'output'. Open data having a relationship to patent data would presumably show innovation as a throughput, since it is measured by people finding those contributing in open data as innovators (followers on github), rather than looking at the specific innovation at completion (patents). One limitation we are facing is that the patent data will not be up to date compared to the network data. Still we believe that a general comparison is possible and could lead to valid results.

References

Acs, Zoltan J, Luc Anselin, and Attila Varga. 2002. "Patents and Innovation Counts as Measures of Regional Production of New Knowledge." *Research Policy* 31 (7). Elsevier: 1069–85.

Blondel, Emmanuel. 2014. *Rsdmx: Tools for Reading SDMX Data and Metadata*. http://CRAN.R-project.org/package=rsdmx.

Breschi, Stefano, and Franco Malerba. 2005. *Clusters, Networks and Innovation*. Oxford University Press.

Couture-Beil, Alex. 2014. *Rjson: JSON for R*. http://CRAN.R-project.org/package=rjson.

Frankelius, Per. 2009. "Questioning Two Myths in Innovation Literature." *The Journal of High Technology Management Research* 20 (1). Elsevier: 40–51.

Freeman, Christopher, and Luc Soete. 2009. "Developing Science, Technology and Innovation Indicators: What We Can Learn from the Past." *Research Policy* 38 (4). Elsevier: 583–89.

Gnyawali, Devi, and Manish Srivastava. 2013. "Complementary Effects of Clusters and Networks on Firm Innovation: A Conceptual Model." *Journal of Engineering Management*, no. 30: 1–20.

Hall, Bronwyn H, Adam Jaffe, and Manuel Trajtenberg. 2005. "Market Value and Patent Citations." *RAND Journal of Economics*. JSTOR, 16–38.

Pakes, Ariel, and Zvi Griliches. 1980. "Patents and R&D at the Firm Level: A First Report." *Economics Letters* 5 (4). Elsevier: 377–81.

Ponds, Roderik, Frank Van Oort, and Koen Frenken. 2010. "Innovation, Spillovers and University–industry Collaboration: An Extended Knowledge Production Function Approach." *Journal of Economic Geography* 10 (2). Oxford Univ Press: 231–55.

R Core Team. 2014. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/.

Schumpeter, Joseph. 1942. "Creative Destruction." *Capitalism, Socialism and Democracy*.

Senghore, Fatima, Enrique Campos-Nanez, Pavel Fomin, and James S Wasek. 2014. "Using Social Network Analysis to Investigate the Potential of Innovation Networks: Lessons Learned from NASA's International Space Apps Challenge." *Procedia Computer Science* 28. Elsevier: 380–88.

Smith, K. H. 2005. "Measuring Innovation." PhD thesis, Oxford University Press.

Taylor, M. Z. 2004. "Empirical Evidence Against Varieties of Capitalism's Theory of Technological Innovation." *International Organization* 58 (03). Cambridge Univ Press: 601–31.

Wickham, Hadley. 2014. *Httr: Tools for Working with URLs and HTTP*. http://CRAN.R-project.org/package=httr.

Wickham, Hadley, and Romain Francois. 2014. *Dplyr: Dplyr: A Grammar of Data Manipulation*. http://CRAN.R-project.org/package=dplyr.