10.5465/AMBPP.2017.123

ON THE MEASUREMENT OF KNOWLEDGE FLOWS WITH PATENT CITATIONS

MARCO CORSINO
Department of Management
Bologna University,
Via Capo di Lucca 34, Bologna (Italy)

MYRIAM MARIANI
Department of Policy Analysis and Public Management and ICRIOS
Bocconi University, Milan (Italy)

SALVATORE TORRISI
Department of Management
Bologna University, Bologna (Italy)

ABSTRACT

This paper investigates the validity of patent citations as indicators of knowledge flows between business organizations. It compares patent citations with inventors' assessment of the importance of business organizations as knowledge sources. We find sizeable measurement errors in citation-based indicators, which highlight strategic rationales for citing rather than knowledge transfers.

INTRODUCTION

Patent citations are a widely used indicator of knowledge flows - e.g., knowledge flows from universities and public research (Gittelman & Kogut, 2003; Roach & Cohen, 2013; Sorenson & Fleming, 2004) and the geographical extent of knowledge spillovers (Alcácer & Gittelman, 2006; Jaffe, Trajtenberg, & Henderson, 1993; Thomson & Fox-Kean, 2005).

Despite their popularity, however, two streams of research contend that patent citations are a noisy indicator of knowledge flows. As a consequence, empirical studies on the determinants and implications of knowledge transfer based on such indicators may produce biased findings. A first set of contributions uncovers confounding factors—for example, the large share of citations added by patent examiners—that dampen the suitability of patent citations as a proxy for knowledge flows (Alcácer, Gittelman, & Sampat, 2009; Criscuolo & Verspagen, 2008; Lampe, 2012; Sampat, 2010; Steensma, Chari, & Heidl, 2014). A second set of studies addresses the validity of patent citations by comparing them with other measures of knowledge flows between organizations and individuals (Callaer, Pellens, & Van Looy, 2014; Duguet & MacGarvie, 2005; Jaffe, Trajtenberg, & Fogarty, 2000; Nelson, 2009; Roach & Cohen, 2013; Tijssen, 2002).

Our research contributes to these two streams of literature. First, it investigates the validity of citation-based indicators with an application to knowledge flows that occur among business organizations (e.g., competitors, suppliers, customers), which account for the majority of patented inventions and represent the most important source of knowledge in several industries (Almeida & Kogut, 1999; Klevorick, Levin, Nelson, & Winter, 1995; Pavitt, 1984; Rosenkopf & Almeida, 2003; Singh & Agrawal, 2011). Earlier studies either focus on knowledge transfer from public research (e.g., Roach & Cohen, 2013) or do not distinguish between public and private sources of

10.5465/AMBPP.2017.123

knowledge (e.g., Duguet & MacGarvie, 2005). Second, our research adopts a fine-grained analysis at the level of the patent application, which allows us to observe knowledge flows at the level of the project leading to a patent application and the individual inventor working on it. Third, it distinguishes between examiner-added and applicant-added citations to isolate knowledge flows tightly coupled with the invention process.

The empirical analysis uses a database comprising about 13,500 research projects that led to patented inventions in business organizations. It compares survey data about knowledge exchanges that occurred during the inventive process with citations to prior art listed in the resulting patents. The estimated results show evidence of systematic measurement error associated with backward citations to other firms' patents. They also indicate that this measurement error is larger for citations from business firms than for those from public research organizations (Roach & Cohen, 2013). We find, indeed, that firm-to-firm backward citations reflect to a larger extent dimensions of strategic behavior in firms' patenting and citing activities.

LITERATURE REVIEW

The Role of Examiners and the Applicant Search for Prior Art

Previous studies uncover two sets of confounding factors that reduce the suitability of patent citations as a proxy for knowledge flows: (i) the large share of citations added by patent examiners, and (ii) the incentives of the applicant to add citations for strategic reasons. Based on USPTO data, Alcácer et al. (2009) and Sampat (2010) find that examiners add 63% of the citations on average and that 39% of the patents include only citations inserted by the examiners. Likewise, Criscuolo and Verspagen (2008) find that the examiners add over 90% of all citations in European Patent Office (EPO) patent applications. The large share of examiner-added citations suggests that the inventors may not be aware of the cited inventions at the time of the inventive process. Moreover, examiner-added citations are not randomly distributed, as they vary across technology fields, firms, and individual examiners, with some of them displaying a set of 'favorite citations' added as prior art (Alcácer & Gittelman, 2008).

Applicants too may cite prior art for reasons unrelated to knowledge flows. On one hand, they have an incentive to cite a large number of prior-art patents to increase the enforceability and validity of a patent in case of litigation (Allison, Lemley, Moore, & Trunkey, 2004; Akers, 2000; Moore, 2003). Sampat (2010) and Lampe (2012), for instance, show that applicants disclose more prior art for higher-expected-value patents. Other studies show that low-value-patent applications filed for purely strategic reasons (e.g., preempting the granting of other patents) also have more prior-art citations than other patents (Guellec, Martinez, & Zuniga, 2012). On the other hand, applicants have also an incentive to omit specific prior art in their applications to obtain broader patent scope. Lampe (2012) finds that inventors of U.S. patents withhold between 21% and 33% of relevant citations of which they are aware. This 'strategic' nondisclosure is positively associated with inventors' patenting experience to estimate the risk of patent invalidation (Steensma et al., 2014).

Validation Studies

This stream of research typically relies on small-scale surveys of inventors or R&D managers to assess the reliability of patent citations as a measure of knowledge flows. By comparing citation counts with the respondents' assessment of the importance of external sources

10.5465/AMBPP.2017.123

of knowledge for the inventive process, these studies suggest that citations are subject to significant biases. For example, Jaffe et al. (2000) find that half of participants in a small-scale survey of U.S. inventors report a low degree of familiarity with the inventions cited in their patents, and about one-third indicate that they did not know about the cited invention before the interview. More recently, Roach and Cohen (2013) use data on U.S. firms provided by the Carnegie Mellon Survey and, by comparing survey-based indicators with citations to public research organizations, report different sources of measurement errors.

DATA & METHODOLOGY

Method

We employ the methodology discussed in Roach and Cohen (2013) to establish the presence of measurement errors in patent citations as indicators of knowledge flows. The approach distinguishes between two sources of measurement errors: (i) errors of omission, that is, the failure of citations to capture important dimensions of knowledge flows; and (ii) errors of commission, that is, the possibility that citations correlate with factors extraneous to knowledge flows. As in Roach and Cohen (2013), we estimate the following two regression models:

$$k_c = \alpha_1 X_1 + \alpha_2 X_2 + \gamma_c P + \varepsilon_c \qquad (1)$$

$$k_s = \theta_1 X_1 + \theta_2 X_2 + \gamma_s P + \varepsilon_s \qquad (2)$$

The dependent variable of model (1), k_c, is a measure of knowledge flows based on patent citations, whereas the dependent variable of model (2), k_s, is an indicator of the "true" knowledge flows obtained from a survey-based measure of the importance of knowledge sources during the invention process. The vector X_1 includes correlates of knowledge flows reflected by patent citations, while the vector X_2 includes correlates of knowledge flows that are not reflected by patent citations and that represent sources of errors of omission. P is a vector of factors influencing patent citations that do not reflect knowledge flows and that represent sources of errors of commission.

Data

Information about the knowledge sources used during the inventive projects comes from the InnoS&T survey that collects information on the inventors, the inventive processes, and the resulting inventions of 22,557 randomly selected patent applications filed at the EPO with priority dates between 2003 and 2005, and their inventors located in 20 European countries, the United States, Israel, and Japan (Torrisi, Gambardella, Giuri, Harhoff, Hoisl, & Mariani, 2016). For the purposes of this study, we focus on 20,825 patent applications held by business organizations.

The survey data were matched with secondary data from the CRIOS-PatStat database that contains standardized information on patents, inventors, and applicants of EPO patent applications (Coffano & Tarasconi, 2014). The CRIOS-PatStat database tracks the citations of all DOCDB patent families—that is, families of equivalents of patent documents filed in different patent offices and sharing the same priority date. After dropping duplicates in citing-cited pairs

10.5465/AMBPP.2017.123

within the same DOCDB family, cited patents without EPO equivalents, self-citations, and observations with missing values in the key covariates, we end up with a working sample of 13,470 patents. At the DOCDB family level, these patents are associated with 73,745 backward citations, that is, a patent average of 5.47 references to prior-art patents, excluding self-citations.

Variables

Dependent variables. We use two types of dependent variables: citations-based and survey-based indicators of knowledge flows from other firms.

The first type of dependent variable is labeled *BACK CITS,* the total number of citations to other business organizations. To depurate this indicator from examiners' added citations, we also constructed the variable *APP BACK CITS,* the total number of only applicant-added citations. The average patent in our setting comprises slightly more than 5 backward citations, 2 of which are added by the applicant.

The second type of dependent variable, *KF SURVEY,* is a survey-based measure that indicates the importance of other firms as knowledge sources. The InnoS&T survey asks inventors to rate the importance of five sources of knowledge: (i.e., customers, users, suppliers, competitors, consulting or contract R&D firms) for the invention process on a 5-point Likert scale (from 1 = not important to 5 = very important). Customers are the most important source of knowledge from private firms (average score 2.72), and consulting or contract R&D firms are the least important (average score 1.45). We construct the variable *KF SURVEY* as the maximum score between the five sources (Shugan & Mitra, 2009).

Correlates of knowledge flows. The first set of covariates involves the search for errors of omission in patent citations to measure knowledge flows. A set of these covariates reflects the use or importance of different channels of knowledge flows: open innovation; private interactions; and employees with a PhD degree. Specifically, we consider three measures of open innovation: (i) patent documents (*Patents*), (ii) participation in technical conferences and workshops (*Conferences*), and (iii) informal interactions in the form of discussions, meetings, and exchange of ideas with people belonging to other, unaffiliated organizations during the invention process (*Informal interactions*). The InnoS&T inventors rated the importance of these three channels of information on a 5-point Likert scale (from 1 = not important to 5 = very important). The use of private interactions is measured by the variable *Formal collaborations,* which equals 1 if the organization engaged in formal collaborations (involving written contracts) with other firms during the inventive process, and 0 otherwise. The variable *Industrial scientist* equals 1 if the inventor holds a master's, a PhD, or a postdoctoral degree at the time of the invention, and 0 otherwise.

Beyond the channels of knowledge flows, we include information about the nature of the inventive output that results from the research project: the variable *Published output* equals 1 if the results related to the focal invention are also published in scientific journals, and 0 otherwise. This type of invention may rely on basic research more than others, and therefore benefit from knowledge flows from basic research.

Correlates of patent citations. This set of covariates comprises correlates of patenting and citing behaviors, but uncorrelated with the survey-based indicator of knowledge flows. They are, therefore, sources of errors of commission. Firms that consider patents to be an effective means to protect their inventions have a higher propensity to patent and, as a result, to cite, than firms

10.5465/AMBPP.2017.123

for which patents are not an effective appropriability mechanism. Thus, the variable *Commercial exploitation* measures (on a 5-point Likert scale) the importance of obtaining patent rights to exploit the invention commercially. In addition, when applying for a patent, firms make strategic decisions about the disclosure of patent references that are relevant for patent enforcement and litigation but that may not reflect any knowledge flow. To account for this possibility, we factor into the model three variables: *Citing propensity*, *Family size*, and *Claims*. The variable *Citing propensity* is computed as the ratio between the stock of backward citations of the applicant (excluding self-citations) divided by its patent stock, both calculated before the priority year of the focal patent. Earlier studies have found that the size of the patent family and the number of claims reflect the expected profitability of the patent and may induce the applicant to over-cite previous inventions to protect themselves from the risk of litigation and patent invalidation (Allison et al., 2004). We therefore include the variable *Family size,* the number of equivalents of the focal patent in the DOCDB family, and the variable *Claims,* the number of claims in the focal patent.

Controls. Our estimates include the applicant's patent stock before the priority of the focal patent (*Patent stock*), a variable equal to 1 if there is a U.S. equivalent in the DOCDB family of the focal patent and 0 otherwise (*US equivalent*), 8 macro-region dummies that identify the patent office of the first filing, and 30 dummy variables for the technological class of the focal patent.

RESULTS

We estimate separate regression models for the three dependent variables. In particular, we use an ordered logit model when the dependent variable is the survey-based measure *KFs SURVEY*. We employ a negative binomial model for the two citation-based measures of knowledge flows (*BACK CITS* and *APP BACK CITS*). To compare the results for the survey-based and the citation-based dependent variables, we discuss the estimated effects of the covariates in terms of the percentage changes in the dependent variables for a one-standard-deviation change in the continuous covariates and a change from 0 to 1 for the discrete variables.

The analysis unveils large errors of omission. All channels of knowledge flows are positively correlated with the survey-based measure of knowledge flows. The three dimensions of open innovation show comparable magnitudes: a one-standard-deviation increase in the variables *Patents*, *Conferences*, and *Informal interactions* implies a 43.8%, 38.4%, and 44.8% increase in *KFs SURVEY*, respectively. Notably, the variable *Formal collaboration* has a large effect: a one-standard-deviation increase in the importance of formal collaborations implies a 113.8% increase in *KFs SURVEY*. For the correlates of citation-based measures of knowledge flows, only the variable *Patents* produces a statistically significant coefficient: a one-standard-deviation increase in the importance of patents as an information source implies a 6.8% increase in *APP BACK CITS* and a 2.9% increase in *BACK CITS*. It is worth noting that the inclusion of the examiner-added citations in the dependent variable would underestimate the influence of patents on the citation-based indicator.

The analysis also reveals substantial errors of commission in citations as a measure of knowledge flows from other firms. Specifically, a one-standard-deviation increase in the variables *Claims*, *Family size,* and *Citing propensity* leads to a 35.9%, 23.9%, and 34% rise in *APP BACK CITS*, respectively. However, these variables are uncorrelated with *KFs SURVEY*. Contrary to expectations, the variable *Commercial exploitation* (a proxy for appropriability) is

10.5465/AMBPP.2017.123

positively associated with both the survey-based and the citation-based measures of knowledge flows. A one-standard-deviation increase in the variable *Commercial exploitation* implies a 14.4% increase in *KFs SURVEY* and a 10.8% increase in *APP BACK CITS*, respectively. This finding suggests that the knowledge provided by other firms can be readily exploited for developing patentable inventions that are close to the market. The inclusion of examiner-added citations produces an underestimation of the effect of *Claims*, *Family size*, *Citing propensity,* and *Commercial exploitation* on the citation-based indicator of knowledge flows.

To check whether our results are not specific to the empirical settings of our study, we quasi-replicated the analysis of Roach and Cohen (2013) with an application to knowledge flows from public research. Our results are fully consistent with those obtained by Roach and Cohen. A comparison with firm-to-firm knowledge flows also shows that citation-based measures reflect dimensions of knowledge flows to a much larger extent when the source is a public research organization than in the case of knowledge flows from business firms.

CONCLUSIONS

The analysis carried out in this study shows that patent citations suffer from systematic measurement errors that may hinder their use as an indicator of knowledge flows among firms. Moreover, the quasi-replication of Roach and Cohen's (2013) results indicates that this measurement error is larger when citations are employed to measure knowledge flows between firms than between firms and public research organizations. Our findings also suggest that citations are more likely added or withheld because of strategic considerations and that knowledge flows across firms are more constrained by secrecy compared with knowledge developed in public research organizations. There may be indeed strategic reasons to maintain secrecy and confidentiality in interfirm collaborations—for example, to avoid information spillover to the benefits of competitors.

REFERENCES AVAILABLE FROM THE AUTHORS