KU LEUVEN

FACULTY OF ECONOMICS AND BUSINESS

Text matching to measure patent similarity

by Sam Arts, Bruno Cassiman and Juan Carlos Gomez

MSI_1708

TEXT MATCHING TO MEASURE PATENT SIMILARITY[1]

Sam Arts
KU Leuven
Korte Nieuwstraat 33, 2000 Antwerp, Belgium
sam.arts@kuleuven.be

Bruno Cassiman
IESE Business School, KU Leuven and CEPR
Avenida Pearson 21, 08034 Barcelona, Spain
bcassiman@iese.edu

Juan Carlos Gomez
University of Guanajuato Campus Irapuato-Salamanca
Carretera Salamanca - Valle de Santiago km 3.5 + 1.8, Salamanca, Mexico
jc.gomez@ugto.mx

Abstract: We propose using text matching to measure the technological similarity between patents. Technology experts from different fields validate the new similarity measure and its improvement on measures based on the United States Patent Classification System, and identify its limitations. As an application, we replicate prior findings on the localization of knowledge spillovers by constructing a case-control group of text-matched patents. We also provide open access to the code and data to calculate the similarity between any two utility patents granted by the United States Patent and Trademark Office between 1976 and 2013, or between any two patent portfolios.

Keywords: text mining, matching, patent, patent classification, technological similarity

Acknowledgements: This work is supported by KU Leuven grant IMP/16/002, FWO grant G071417N from the Flemish government and MEC grant ECO2015-71173-P from the Spanish government. We would like to thank the technology experts and the participants in the presentations at the 2016 Academy of Management Conference, the OECD Blue Sky Forum on Science and Innovation Indicators, the R&D management conference, as well as Juan Alcacer, Sen Chai, Andrea Fosfuri, Alfonso Gambardella, Giovanni Valentini, and Bart Van Looy. Errors and omissions remain those of the authors.

[1] The appendix provides a description of the code and the data files. Code and data are available at https://dataverse.harvard.edu/dataverse/patenttext.

INTRODUCTION

Measuring the similarity between particular bodies of knowledge is a critical step in many innovation and strategy studies: Are knowledge spillovers geographically localized (Jaffe, Trajtenberg, and Henderson, 1993)? Does inter-firm mobility of engineers influence the transfer of knowledge between firms or regions (Singh and Agrawal, 2011; Almeida and Kogut, 1999)? Do alliance partners draw on related technological knowledge from each other (Rosenkopf and Almeida, 2003)? Are mergers and acquisitions between firms with similar technological knowledge more successful (Makri, Hitt, and Lane, 2010)? How do knowledge spillovers affect R&D investments and the productivity of firms (Bloom, Schankerman, and Van Reenen, 2013)? In order to answer these and related questions convincingly, it is necessary to carefully measure the technological similarity between patents or patent portfolios.

Prior and current research on innovation and strategy traditionally relies on the classification system of patent offices to measure the similarity between patents (e.g., Singh and Marx, 2013; Aharonson and Schilling, 2016), to construct a case-control group of similar patents (e.g., Jaffe *et al.,* 1993; Almeida, 1996; Almeida and Kogut, 1999; Agrawal, Cockburn, and Rosell, 2010; Belenzon and Schankerman, 2013), or to measure the similarity between patent portfolios of firms (e.g., Ahuja, 2000; Rosenkopf and Almeida, 2003; Makri *et al.*, 2010; Bloom *et al.*, 2013). However, this aggregated classification system might not capture all the technological characteristics of an invention (Thompson and Fox-Kean, 2005; Singh and Agrawal, 2011). Moreover, different classes and subclasses might contain significant overlap so that technologically similar patents can have a different classification (McNamee, 2013).

In this paper, we use a text-mining technique based on common keywords to develop a new measure of technological similarity for all utility patents granted by the United States

Patent and Trademark Office (USPTO) between 1976 and 2013. This new similarity measure and its improvement on measures based on the United States Patent Classification System (USPC) are validated in two ways.[2] First, by means of an expert panel whereby thirteen independent experts from five different fields assess the technological similarity of a random sample of patents from their field of expertise. Second, we start from the assumption that patents are more likely to be technologically similar if they belong to the same patent family, list the same inventors, are owned by the same assignees, or cite each other. Using the full population of patents, we confirm that text-matched patents are more likely to cite each other, to belong to the same patent family, or have a common inventor or assignee compared to patents that are matched based on their patent classification. Moreover, we find significant differences across different groups of text-matched patent pairs depending on the degree of similarity in text.

To identify the limitations of the new similarity measure, we asked feedback from the experts on the discrepancies between our text-based similarity measure and their personal rating. Not surprisingly, patents with only few keywords with little discriminatory power — such as method, system, process, and material — increase the likelihood of false positives or type I errors. Different spelling variants and synonyms increase the likelihood of false negatives or type II errors.

In this paper, we mainly focus on the use of text matching to construct a case-control sample of technologically similar patents filed in the same year — a frequently recurring effort in the strategy and innovation field (e.g., Jaffe *et al.,* 1993; Almeida, 1996; Agrawal *et al.*, 2010; Belenzon and Schankerman, 2013). An alternative and more effective approach was introduced by Thompson (2006), who studied the within-patent variation in geographic

[2] The USPTO stopped classifying patents according to the USPC and switched to the Cooperative Patent Classification (CPC) system in January 2015. The CPC is based on a harmonization of the existing classification systems of the European Patent Office and the USPTO, ECLA and USPC respectively. Nonetheless, the large majority of studies relies on U.S. patents filed before 2015 and classified according to the USPC.

localization between examiner- and inventor-added citations. Yet, patent fixed effects cannot be used for most other applications and research questions. Furthermore, not all patents have both examiner- and inventor-added citations, and this information is only available for patents granted since 2001 (Lemley and Sampat, 2012). Moreover, examiners rarely insert citations to scientific prior art so that patent fixed effects cannot be used to study the localization of patent-to-paper citations (e.g., Belenzon and Schankerman, 2013). Therefore, text matching provides a useful alternative for applications where patent fixed effects cannot be used.[3] As an application of our new text-based similarity measure, we use text-matched case-control patents to replicate Thompson's (2006) findings on the localization of knowledge spillovers. Both identification strategies — text-matching and within-patent variation — lead to the conclusion that knowledge spillovers are geographically localized at the country, state, and metropolitan level, or measured as spatial proximity.

Besides the use of text matching to select a case-control sample, we discuss other potential applications for both future research and practitioners, and provide open access to the code and data to calculate the similarity between any two utility patents granted by the USPTO between 1976 and 2013 or between any two patent portfolios (available at https://dataverse.harvard.edu/dataverse/patenttext).

TEXT MATCHING TO MEASURE TECHNOLOGICAL SIMILARITY

Sample and data collection

We retrieve the titles and abstracts of all utility patents granted by the USPTO between 1976 and 2013 from PATSTAT (October 2013 edition). The data is available for 4,422,009 patents, approximately 97.5 percent of the utility patents granted by the USPTO.[4] We

[3] In a related effort, Younge and Kuhn (2016) use a vector space model to measure patent similarity.
[4] The first patent in our sample is 3,930,271; granted on January 6 1976. The last patent in our sample is 8,495,761; granted on July 23, 2013.

preprocess the data by concatenating the title and abstract, lowercasing the text, tokenizing all words, and eliminating stop words, words with only one character, numbers, and words that appear only once across all patents — arguably because of spelling mistakes.[5] What remains is a collection of unique keywords for each patent that represents the technical content of the patent. The full patent corpus contains 526,561 unique keywords and the average number of unique keywords per patent is 37. Hence, the overall number of keywords and the number per patent document are significantly larger than the number of classes and subclasses. Figure 1 shows the histogram of the number of unique keywords per patent.

'Insert Figure 1 here.'

A text-based measure of patent similarity

To measure the similarity between any two patents, we calculate a Jaccard index by dividing the number of unique keywords in the intersection of the two patents by the number of unique keywords in the union.[6] Thus, we have a continuous measure ranging from zero to one. Because patent classification changes over time and because we want to compare text matching to USPC matching, we only compare the similarity between patents filed in the same year in the remainder of the paper.[7] For each baseline patent in our sample, we select the 200 most similar text-matched patents filed in the same year. To do so, we compare the keywords of the baseline patent with the keywords of all other patents filed in the same year, irrespective of the classification of the patents.[8] To make sure that we have sufficient information on the content of a patent and that text-matched patents have at least one or more

[5] Appendix 1 provides more detailed information on the code.

[6] Cosine similarity is an alternative measure of similarity based on common keywords. We calculate the cosine similarity for a random sample of 25,000 patent pairs with varying degrees of Jaccard similarity. The correlation between the Jaccard and cosine similarity is 0.91, suggesting that both similarity measures lead to similar results (Magerman, Van Looy, and Debackere, 2015).

[7] We provide open access to the code and data so that anyone can calculate the similarity between any two patents or between any two groups of patents irrespective of the time of filing.

[8] For n patents, the number of unique patent pairs is $\frac{n!}{2 * (n-2)!}$. Comparing the similarity between 50 patents requires 1,225 pairwise calculations. Comparing the similarity between 150 patents requires as many as 11,175 calculations.

keywords in common, we restrict the sample to patents with at least ten unique keywords and to text-matched patents with a minimum Jaccard index of 0.05 (0.8 percent of the sample drops, resulting in 4,386,405 baseline patents and 855,413,378 text-matched patent pairs).

In line with the traditional matching method pioneered by Jaffe *et al.* (1993), we select for each baseline patent the closest text-matched patent — with the highest Jaccard — filed in the same year. In cases where multiple patents have an identical Jaccard index, we select the one with the closest filing date. The average Jaccard index for the 4,386,405 closest text-matched patent pairs is 0.24, corresponding to approximately 14 common keywords for two patents with an average number of 37 keywords. Besides selecting the closest text-matched patent, we select for each baseline patent all corresponding patents without any overlap in keywords (Jaccard index equals zero), filed in the same year, and we select the patent with the closest filing date. Thus, for each baseline patent, we have the closest text-matched patent, the 200 most similar text-matched patents, and a distant text-matched patent with a Jaccard index of zero.

Validation

We validate the text-based measure of technological similarity in two ways. First, we use an expert panel and a random sample of text-matched patents to test the externally validity of the new measure. Second, we analyze the joint characteristics of the full population of closest and distant text-matched patents to assess the face validity of the new measure.

Expert validation

We recruited thirteen paid experts from five different fields to rate the technological similarity of different patents in their field of expertise on a Likert scale from one (completely dissimilar, no content in common) to seven (very similar, almost identical content). The rating is exclusively based on the technical description available in the full

patent document of the two patents. The group of experts consists of both academic and industrial scientists and engineers specialized in five different fields: three R&D scientists and engineers with at least six years of working experience in optical inspection systems who work for a large multinational company, four R&D scientists and engineers specialized in semiconductor devices with at least four years of working experience in a large public research organization, four final-year PhD students specialized in molecular biology and microbiology working in a university lab, one R&D engineer with four years of working experience in chemical engineering in a large multinational chemicals company, and one R&D engineer specialized in power systems who worked for more than twenty years in a large specialty chemicals company. We restrict the actual analysis to the three fields with multiple experts.[9]

Relying on the technological categories of Hall, Jaffe, and Trajtenberg (2002) and the USPC, we select for each of the five fields a random sample of baseline patents,[10] and randomly match each baseline patent with five text-matched patents with varying degrees of Jaccard similarity: 0; 0.05–0.25; 0.25–0.50; 0.50–0.75; and larger than 0.75. As a result, we have for each field a sample of baseline patents, and for each baseline patent five text-matched patents with varying degrees of Jaccard similarity. Although sampling on expected similarity might result in bias, random sampling without conditioning on Jaccard similarity would result in a very large share of patent pairs without any content in common and would make the exercise fruitless.[11] We randomize the order in which we present the text-matched patent pairs to the experts. Experts from the same field rate the same set of patent pairs in order to assess inter-rater agreement. The experts rated on average approximately 65 patent pairs, resulting in a total number of 850 ratings. Rating a particular patent pair lasted between

[9] None of our findings changes if we include the two fields with only one expert.
[10] We manually check each selected baseline patent to make sure that it matches the field of expertise, and redraw a new random baseline patent in cases where there is doubt.
[11] We would like to thank a reviewer who pointed out the potential sampling bias.

15 and 45 minutes depending on the similarity of the patents and the length of the full patent document. In the actual analysis, we discard text-matched patent pairs from the same patent family as they might bias our findings.[12]

The average inter-item correlation between the ratings of different experts is 0.812 and Cronbach's alpha is 0.945.[13] Thus, multiple experts from the same field consistently agree on the similarity of two patents. We find no significant differences in ratings between experts with different levels of experience. Most importantly, the correlation between our text-based similarity measure and the expert ratings of technological similarity — based on the full technical description of the patents — is 0.838 and Cronbach's alpha is 0.912. These results verify that the Jaccard index can be used to measure the similarity of two patents on a continuous scale from zero to one.

'Insert Table 1 here.'

'Insert Figure 2 here.'

'Insert Figure 3 here.'

Table 1 provides summary statistics of the expert ratings for the five subsamples of text-matched patent pairs split by Jaccard similarity. Figure 2 plots the expert ratings, and Figure 3 displays the means of expert ratings and the corresponding confidence intervals. Unreported t-tests indicate that the means of expert ratings of the subgroups are significantly different from each other at the one percent level. Table 2 shows the results of ordered logit and linear regression models with expert ratings of similarity as outcome. The unit of observation is a single text-matched patent pair rated by a single expert. Regressions demonstrate that the expert ratings are significantly higher for patent pairs with a higher

[12] We rely on the DOCDB patent family identifier from PATSTAT. The DOCDB identifier groups all patents that share the same priority application (including divisional, continuation, and continuation-in-part applications). The findings of the expert validation do not change if we include patents from the same family.
[13] We use the *alpha* command in STATA with the *std* option to standardize items in the scale to mean 0 and variance 1, and calculate inter-item correlation and Cronbach's alpha.

Jaccard index. All estimated coefficients are significantly different from each other at the one percent level. Using random or fixed effects to control for unobserved heterogeneity across baseline patents or experts does not affect our results. The high R-squared across the different models — ranging from 0.702 to 0.766 — illustrates that the text-based similarity measure explains a large proportion of the variance in expert ratings, and that the keywords in the title and abstract of a patent provide the required information to assess the technological similarity of patents.

'Insert Table 2 here.'

To better understand the limitations of the text-based similarity measure, we identify false positives and false negatives — or type I and type II errors — and ask the experts for detailed feedback on the differences between our text-based similarity measure and their personal rating. First, we identify false positives as patent pairs with a high Jaccard index but a low expert rating of similarity (11 ratings, 3.5 percent of the ratings for pairs with a Jaccard larger than 0.25).[14] Although the pairs are split by Jaccard similarity, we find that the average Jaccard index of these false positives is somewhat lower — but still within the same boundaries by construction — compared to the patents pairs from the same group. In addition, the text-matched patents associated with false positives have a significantly lower number of unique keywords in the title and abstract. A given Jaccard index corresponds with fewer common keywords for patents with less keywords compared to patents with more keywords. Because the experts use the full technical description rather than just the title and abstract, the likelihood of false positives decreases with the number of keywords in the title and abstract. Across all ratings by all experts, we indeed find that the correlation between the text-based similarity measure and the expert rating is higher for patents with more keywords. Moreover, the experts pointed out that false positives are often matched on more general

[14] two patent pairs with jaccard>0.75 and expert rating<=3, eight patent pairs with 0.50<jaccard<=0.75 and expert rating<=3, one patent pair with 0.25<jaccard<=0.50 and expert rating=1

keywords such as method, system, device, apparatus, process, material, image, and sensor, which have many different applications across different fields. False positive matches often share the same tools and methods that are used for different applications in different contexts. Hence, the fact that patents have few keywords with little discriminatory power increases the likelihood of false positives or type I errors.

Second, we identify false negatives as patent pairs with a low Jaccard index but a high expert rating of similarity (6 ratings, 1.6 percent of the ratings for pairs with a Jaccard lower than 0.50).[15] Although the pairs are split by Jaccard index by construction, we find that the average Jaccard index of these false negatives is significantly higher compared to the patents pairs from the same group. We find no differences in the number of keywords, but the experts point out that false negatives often correspond to patents with different yet closely related keywords, such as '*system for monitoring errors*' versus '*defect inspection method and apparatus*'. Thus, keywords with different spellings and synonyms increase the likelihood of false negatives or type II errors.

Face validity

We analyze the joint characteristics of the full population of closest and distant text-matched patents to assess the face validity of the new measure. We select for each baseline patent the closest text-matched patent filed in the same year and a distant text-matched patent with Jaccard similarity equal to zero and filed in the same year.

'Insert Table 3 here.'

We rely on the assumption that two patents of the same patent family, developed by the same inventor(s), assigned to the same assignee(s), or which cite each other, are similar to a certain degree. For all text-matched patent pairs, we calculate binary indicators equal to one

[15] three patent pairs with 0.05>=jaccard<0.25 and expert rating>=6, three patent pairs with 0.25<=jaccard<0.50 and expert rating=7

in cases where the patents share the *same patent family,* have at least one inventor in common (*same inventor(s)*), have at least one assignee in common (*same assignee(s)*), and in cases where one patent cites the other (*citation link*). Patent family identifiers are collected from PATSTAT; information on inventors, assignees, and citations is collected from the disambiguated inventor database (Li *et al.*, 2014). Columns 1 to 5 of Table 3 provide summary statistics for different subsamples of text-matched patent pairs split by Jaccard index. For patent pairs with a Jaccard index of zero (column 1), 0.0 percent belong to the same patent family, 0.0 percent to the same inventors, 0.1 percent to the same assignees and, in 0.0 percent of the cases, one cites the other. For text-matched patent pairs with a Jaccard index equal to or larger than 0.75 (column 5), 40.7 percent belong to the same patent family, 92.7 percent to the same inventor(s), 86.5 percent to the same assignee(s) and, in 8.5 percent of the cases, one cites the other. While the latter numbers might seem large at first sight, it should be noted that a minimum Jaccard index of 0.75 corresponds to highly similar patents, i.e. two average patents with 37 unique keywords having at least 31 keywords in common. Moreover, the average Jaccard index for our sample of closest text-matched patent pairs is 0.24. Unreported t-tests indicate significant differences at the one percent level in *same patent family*, *same inventor(s)*, *same assignee(s)*, and *citation link* across the five different subsamples split by Jaccard index. The only non-significant difference is in *same assignee(s)* and *citation link* for pairs with Jaccard>=0.50 and <0.75 (column 4) versus pairs with Jaccard>=0.75 (column 5). Because assignees and inventors tend to work on related technologies, and because citing patents and patents of the same family tend to be similar, we interpret these results as indirect evidence that text matching can be used to measure the technological similarity between patents on a continuous scale.

TECHNOLOGICAL SIMILARITY BASED ON THE UNITED STATES PATENT CLASSIFICATION SYSTEM

Text-based versus USPC-based similarity measures

To compare the accuracy of our text-based similarity measure with a similarity measure based on the USPC, we calculate *subclass similarity* as the number of unique subclasses two patents have in common divided by the total number of unique subclasses in the union (e.g., Singh and Marx, 2013; Agrawal *et al.*, 2010). Subclass similarity is calculated in the same way as text-based similarity except for using subclasses instead of keywords. Subclasses are nested within classes and provide the most granular level of classification. For the sample of patent pairs rated by the experts, subclass similarity ranges from zero to one, and has an average of 0.203 and a standard deviation of 0.326. We find a correlation of 0.448 between subclass similarity and expert rating of similarity (compared to 0.838 for the text-based similarity measure), and a Cronbach's alpha of 0.619 (compared to 0.912 for the text-based measure). Findings in Table 2 (columns 5–7) illustrate that text similarity is a more significant and precise predictor of expert ratings. The coefficient of subclass similarity drops and becomes insignificant when including text similarity, and the explanatory power of the regression increases dramatically. Regressions including fixed effects or ordered logit models give similar results. Therefore, we conclude that the text-based Jaccard index provides a more accurate measure for the technological similarity of patents. It should be noted that scholars and practitioners may well want to combine both USPC and text to measure patent similarity and identify closely related patents.

Text-matched versus class- and subclass-matched patents

Prior research traditionally relies on the USPC to construct a matched control group of similar patents (e.g., Jaffe *et al.,* 1993; Almeida, 1996; Almeida and Kogut, 1999; Agrawal, Cockburn, and Rosell, 2010; Belenzon and Schankerman, 2013). In line with common practice, we select three alternative groups of USPC-matched patents. We follow the method pioneered by Jaffe *et al.* (1993), matching each of the baseline patents in our sample

(n=4,386,405) to all other patents with the same primary class and filing year, and selecting the patent with the closest filing date. This results in a sample of 4,279,839 *primary-class-matched* patent pairs (we find no match for two percent of the sample). Alternatively, scholars relied on primary subclass, filing year, and approximate filing date to select technologically similar patents (e.g., Almeida, 1996). Following this procedure results in 3,492,480 *primary-subclass-matched* patent pairs (we find no match for 20 percent of the sample). Finally, we use all subclasses of a patent (e.g., Agrawal *et al.*, 2010), calculate a Jaccard index based on the overlap in subclasses with all other patents filed in the same year, select the patents with the highest subclass similarity, and select the patent with the closest filing date in cases where there are multiple matches. This renders a sample of 4,229,647 *subclasses-matched* patent pairs (we find no match for four percent of the sample).

We validate the improvement of text matching on matching based on the USPC in two ways. First, we use an expert panel and a random sample of text-matched and corresponding USPC-matched patents to validate the improvement externally. Second, we analyze the joint characteristics of the full population of text-matched and corresponding USPC-matched patents to assess the face validity of the improvement.

Expert validation

For each field of expertise, we selected a new random sample of baseline patents and, for each baseline patent, the corresponding closest text-matched, primary-class matched, and subclasses-matched patent. To reduce the workload for the experts, we excluded the primary-subclass matched patents. The eleven experts rated on average approximately 27 patent pairs, resulting in a total number of 300 ratings. We randomized the order of the patent pairs, and we asked the experts to rate the technological similarity of the patents on a Likert scale from one (completely dissimilar, no content in common) to seven (very similar, almost identical

content) using the full patent document. In the analysis, we discard patent pairs from the same patent family since they might bias our findings.

'Insert Table 4 here.'

'Insert Figure 4 here.'

'Insert Figure 5 here.'

The average inter-item correlation between the ratings of different experts for the text-matched, primary-class matched and subclasses-matched patents is 0.635, and Cronbach's alpha is 0.874. Hence, multiple experts from the same field generally agree on the similarity of text-matched and USPC-matched patents. The correlation between our text-based similarity measure and the expert ratings of technological similarity is 0.618, and Cronbach's alpha is 0.890. While these numbers are somewhat lower compared to our initial validation exercise, presumably because there is less variation in Jaccard similarity, they remain reasonably high. Table 4 provides summary statistics of the expert ratings for the primary-class matched, subclasses-matched, and text-matched patent pairs. Figure 4 plots the expert ratings, and Figure 5 displays the means of expert ratings and the corresponding confidence intervals. The text-matched patents receive the highest average expert rating of similarity (unreported t-tests significant at the one percent level). Table 5 displays the results of ordered logit and OLS regression models, and shows that the expert rating is significantly higher for the text-matched patents compared to both primary-class-matched and subclasses-matched patents. All differences are significant at the one percent level. Using random or fixed effects to control for unobserved heterogeneity across baseline patents or experts does not affect our results.

'Insert Table 5 here.'

To obtain a deeper insight into the limitations of matching on primary class, subclasses and text, we calculate the likelihood of false positives or type I errors as the share of matched patent pairs that received a low expert rating of similarity. It should be noted that

classes and subclasses may be constructed by the USPTO to capture related but not identical patents, and examiners may add insights besides the patent text alone. Consequently, false positives do not necessarily reflect classification bias, but potential bias in research that uses the USPC to select technologically similar patents. For the *primary-class matched* patents, 44 percent receive the lowest rating on a scale of one to seven. Ignoring any potential errors made by the experts, the likelihood that primary-class matched patents are completely dissimilar and have no content in common is estimated at 44 percent. Seventy four percent of the primary-class matched patent pairs receive a rating of one or two. For the *subclasses-matched* patents, 24 percent receive the lowest rating of one and 46 percent receive a maximum rating of two. For the *text-matched* patents, 10 percent receive the lowest rating and 24 percent receive a maximum rating of two.

While text matching significantly reduces the likelihood of false positives, a significant share of errors remains present. In line with our previous results, we find that these false positives correspond to text-matched patents with a smaller number of keywords and a lower Jaccard index compared to the other text-matched patents (all differences are significant, at least at the five percent level). Not every baseline patent has a close text-matched patent filed in the same year. Depending on the research question at hand, scholars may want to enlarge the sample of potential matches to patents filed in different years, or restrict the analysis to patents for which a sufficiently close text-matched patent can be found (e.g., Thompson and Fox-Kean, 2005). The findings from the expert validation discussed in the first part of this paper (especially Figure 3) can be used to determine a lower bound of Jaccard similarity.

Face validity

Finally, we analyze joint characteristics of the full population of text-matched, primary-class matched, primary-subclass matched, and subclasses-matched patents to compare the accuracy

of different matching methods. As illustrated in column 1 of Table 6, the average Jaccard index for the 4,386,405 closest *text-matched* patent pairs is 0.238, corresponding to approximately 14 common keywords for two patents with an average number of 37 keywords. The average Jaccard index for the *primary-class-matched* patents is 0.054 (column 2). This corresponds to approximately three common keywords for two average patents with 37 unique keywords, which is arguably low. Twelve percent of the primary-class-matched patents have not a single keyword in common. As shown in column 3, the average Jaccard index for the *primary-subclass-matched* patent pairs is 0.092, corresponding to approximately six common keywords for two average patents with 37 keywords. The percentage of primary-subclass-matched patents with not a single keyword in common is 4.3. Finally, the *subclasses-matched* patents have an average Jaccard index of 0.097, corresponding to approximately seven common keywords for two average patents with 37 keywords. Nevertheless, 4.0 percent of the subclasses-matched patents have a Jaccard index of zero. Not surprisingly, text matching results in closer matches compared to matching based on classes or subclasses. We find a text-matched patent with a higher Jaccard index for 98.3 percent of the 4,279,839 primary-class-matched patents, for 95.8 percent of the 3,492,480 primary-subclass-matched patents, and for 96.4 percent of the 4,229,647 subclasses-matched patents.

'Insert Table 6 here.'

Figure 6 plots the non-parametric kernel density estimation of the text-based Jaccard index for the different groups of matched patents. A large share of the USPC-matched patents has a low similarity in text. Matching on primary subclass improves the similarity compared to matching on primary class; matching on all subclasses only marginally improves the similarity compared to matching on primary subclass only.

'Insert Figure 6 here.'

Finally, paired t-tests displayed in columns 2, 3, and 4 of Table 6 illustrate that text-matched patents are significantly more likely to belong to the same patent family, to be developed by the same inventor(s), to be assigned to the same assignee(s), and to cite each other, in comparison to primary-class-matched, primary-subclass-matched, and subclasses-matched patents. All differences are significant at the one percent level. These findings indirectly illustrate the strength of text-based matching for the full population of patents. Together with the expert validation, the results demonstrate the improvement of text-based matching on matching based on the USPC. It should be noted that we only compared text-matched to USPC-matched patents to illustrate the improvement of text matching. In practice, scholars might want to match on both USPC and text, which would arguably result in the closest match.

TEXT MATCHING TO STUDY THE LOCALIZATION OF KNOWLEDGE SPILLOVERS

As an application of our new similarity measure, we use text-matched case-control patents to replicate prior findings on the localization of knowledge spillovers. Technological progress is a search process whereby inventors rely on prior knowledge to solve new problems and develop new technologies (Mokyr, 1990). Knowledge spillovers, the non-rival transfer of knowledge across individuals from different organizations, are key for technological progress and economic growth (Romer, 1986; Krugman, 1991; Grosman and Helpman, 1991). There is a long-standing debate fueled by mixed evidence on whether and to what extent these spillovers are geographically localized. Starting with Jaffe *et al.* (1993), prior research has predominantly used patent citations as an indicator of knowledge spillovers, and the geographical proximity between inventors of the citing and the cited patents as evidence of the localization of these spillovers (e.g., Thompson, 2006; Singh and Marx, 2013; Murata, Nakajima, Okamoto, and Tamura, 2014).

Because industries and related technological activities are spatially clustered rather than randomly dispersed across the globe, as with information technology in Silicon Valley and biotechnology in Boston, it is critical to control for the pre-existing concentration of knowledge production while estimating the localization of spillovers (Audretsch and Feldman, 1996). To solve this identification problem, scholars used a case-control matching method, selecting a control patent matched to the citing patent on primary class or subclass and approximate filing date (e.g., Jaffe *et al.,* 1993; Almeida, 1996; Agrawal, Cockburn, and Rosell, 2010; Belenzon and Schankerman, 2013). A random control patent belonging to the same class but not citing the same patent would arguably control for the pre-existing geographical distribution of technological activities. In this research design, a significant difference in localization between the cited and the citing patent versus between the cited and the non-citing control patent is interpreted as evidence for the localization of knowledge spillovers. Therefore, it is critical to find control patents that are technologically similar to the citing patents.

Thompson (2006) introduced a new and more effective identification strategy by studying the within-citing-patent variation in the geographic localization of citations added by inventors and citations added by examiners. This new approach solves the imperfect matching problem and the key identification assumption present in all studies using the case-control matching method (Thompson and Fox-Kean, 2005). The new identification strategy rests on two main assumptions. First, inventor-added citations are more likely to represent true knowledge spillovers than examiner-added citations. Second, examiners cannot learn about prior art because of geographic proximity to related technological activities as they are typically recruited directly after college and work in a single campus in Alexandria, Virginia. The new approach also has a number of limitations. First, examiners might add citations that represent true knowledge spillovers but are omitted by the inventors, for instance for strategic

reasons (Lampe, 2012). Second, detailed citation information is only available for patents granted since 2001, and the analysis is limited to patents with both examiner- and inventor-added citations, which might result in selection bias. Approximately 40 percent of all patents have all citations added by examiners, and approximately eight percent of all patents have no citations added by examiners (Alcacer and Gittelman, 2006). Third, examiners rarely insert citations to scientific prior art so that the approach cannot be used to study the localization of patent-to-paper citations (Lemley and Sampat, 2012; Belenzon and Schankerman, 2013). Finally, the new approach cannot be used for most other research questions and applications. Thus, text matching provides a useful alternative for applications where within-patent variation cannot be used.

As an application of the text-based similarity measure, we use text-matched case-control patents to replicate prior findings on the localization of knowledge spillovers. For each citing patent, we select a non-citing case-control patent based on similarity in text and filing date. Our statistical test compares (1) the geographic localization of inventor-cited patents and citing patents, with (2) the geographic localization of inventor-cited patents and text-matched control patents. Although imperfect matching based on text is necessarily inferior to within-patent variation, comparing the results of the three different approaches — case-control patents based on USPC, case-control patents based on text, within-patent variation — allows us to assess the accuracy of text matching.

We use the replication data provided by Thompson (2006), which include a sample of 2,670 *citing patents* granted in the first week of January 2003 and having an institutional assignee, and 27,665 citations made by these patents to *cited patents* granted after January 1 1976. Self-citations — between the same assignees — are excluded. In line with Thompson (2006), we select the first inventor for each patent, but use the disambiguated address information from Li *et al.* (2014) — rather than manual cleaning — to determine the

geographic matching and spatial proximity between two inventors on two different patents. We calculate binary indicators for *matched country* (the two inventors reside in the same country), *matched state* (conditional on the patent having a U.S. inventor), and *matched cbsa* (conditional on the patent coming from a CBSA). To assess the localization of knowledge spillovers at the metropolitan level, we follow Singh and Marx (2013) and map cities to core-based statistical areas (CBSAs), a U.S. geographic area that consists of one or more counties of at least 10,000 people plus adjacent counties within commuting distance. CBSA definitions are intended to cover reasonable commuting distances and replace the prior MSA/CMSA definitions. We rely on the 2003 definition by the U.S. Office of Management and Budget. Using CBSA instead of MSA/CMSA results in a larger share of patents for which the location can be defined (Singh and Marx, 2013). Finally, we calculate *distance in miles* using data from Li *et al.* (2014) that maps cities to latitudes and longitudes, and the *vincenty* command in Stata.

First, we compare the selection of control patents based on USPC versus the selection of control patents based on text. We follow prior research and select for each citing patent a control patent based on primary class and approximate filing date (e.g., Jaffe *et al.,* 1993), and a control patent based on primary subclass and approximate filing date (e.g., Almeida, 1996). In addition, we select for each citing patent one text-matched control patent that is most similar to the citing patent as measured by the Jaccard index based on keywords and approximate filing date. Table 7 displays the geographic matching rates and spatial proximity between the first inventor of the citing patent and the first inventor of respectively the text-matched, primary-class matched, and primary-subclass matched control patents. The numbers between parentheses give the t-statistic for the test of equality in geographic localization to the citing patents. All differences are significant at the one percent level. Compared to class-matched (subclass-matched) patents, text-matched patents are 1.72 (1.52) times more likely

to match the country of the citing patent, 3.34 (2.19) times more likely to match the state, 4.07 (2.40) times more likely to match the CBSA, and 34 (33) percent closer in miles. Hence, text-matched patents provide a better control for the pre-existing geographic concentration of related technological activities compared to USPC-matched patents, and thus offer a more rigorous test for the localization of knowledge spillovers.

'Insert Table 7 here.'

Second, we use text-matched case-control patents to replicate Thompson's (2006) findings on the localization of knowledge spillovers, and to compare the text-matching approach to the within-patent approach. For each inventor-added citation from the Thompson (2006) sample, we select for each citing patent the population of patents filed in the same year that are not part of the same patent family and are not citing the same patent. Of the remaining patents, we select the patent most similar to the citing patent as measured by the Jaccard index based on keywords as *text-matched control*. In case multiple patents have identical Jaccard scores, we select the patent with the closest filing date.

'Insert Table 8 here.'

Table 8 displays the geographic matching rates and spatial proximity between the inventor-cited and citing patents (column 2), between the examiner-cited and citing patents (column 3), and between the inventor-cited and text-matched control patents (column 4). The numbers between parentheses give the t-statistic for the test of equality in geographic localization. Despite the fact that text-matched patents provide a more rigorous control for the pre-existing geographic concentration of related technological activities compared to USPC-matched patents, we continue to find strong support for the localization of knowledge spillovers. Compared to text-matched control patents, citing patents are 1.16 times more likely to match the country of the cited patent, 1.17 times more likely to match the state, 1.32 times more likely to match the CBSA, and 14 percent closer in miles. All differences are

significant at the one percent level. As shown in column 3, using within-citing-patent variation between citations added by inventors and citations added by examiners leads to the same conclusion of localized knowledge spillovers at the country, state, and metropolitan level, or measured as spatial proximity. It should be noted that it is difficult to compare the effect sizes or economic significance of both approaches. The reason is that the text-matching approach compares the localization of — on the one hand — the cited patent and — on the other hand — the citing and text-matched control patents. By contrast, the within-patent variation approach compares the localization of — on the one hand — the citing patent and — on the other hand — the examiner-cited and inventor-cited patents. Nevertheless, comparing columns (3) and (4) of Table 8 indicates that there are no significant differences (not even at the ten percent level) in *match state*, *match cbsa*, and *distance in miles* for text-matched patents versus examiner-added citations as control. Yet, *match country* is significantly higher for text-matched patents compared to examiner-added citations. The fact that both identification strategies — within-patent variation and case-control patents based on text — render the same results illustrates the effectiveness of text matching. As illustrated in Table 7, selecting case-control patents based on text or based on USPC results in different findings.

Table 9 displays the results of the regressions estimating the localization of spillovers using the within-patent approach (panel A) and the text-matching approach (panel B). Besides a binary indicator for *inventor citation*, the regressions include controls for *cited non-institutional* (binary indicator equal to one in cases where the cited patent has no institutional assignee) and *cited patent age* (difference in number of years between the filing date of the cited patent and the filing date of the citing/text-matched control patent). The regression results are generally in line with the t-tests displayed in Table 8. Both the text-matching approach and the within-patent approach lead to the conclusion that knowledge spillovers are

geographically localized at all levels of analysis. The only exception is that inventor citation is only significant at the 0.17 level in the regression estimating the localization of spillovers at the metropolitan level for the within-patent approach (Panel A). This lack of significance is presumably driven by the limited variation in *match cbsa* between examiner- and inventor-added citations within the same patent, and the associated drop in observations in the fixed-effects logit model. Estimating the same logit model without fixed effects doubles the number of observations and turns the effect of *inventor citation* significant at the three percent level. In conclusion, our results differ from Thompson and Fox-Kean (2005), who only found support for localization at the country level, but are in line with the original findings of Jaffe *et al.* (1993) and more recent findings of Thompson (2006) and Singh and Marx (2013). Both identification strategies — text-matching and within-patent variation — lead to the conclusion that knowledge spillovers are geographically localized.

'Insert Table 9 here.'

DISCUSSION AND CONCLUSION

We make three main contributions to the literature. First, we illustrate how a text-mining tool can be used to more accurately measure the technological similarity between patents on a continuous scale. Our method and data can also be used to calculate the similarity between two groups of patents, for instance between the patent portfolios of two companies (e.g., Makri *et al.*, 2010; Bloom *et al.*, 2013). A more fine-grained measure of the similarity of two patent portfolios can be obtained by aggregating keywords at the portfolio level and calculating the Jaccard, cosine, or Mahalanobis similarity between the two patent portfolios. Text mining can also be used to identify technologically dissimilar or novel patents, which have little overlap in content compared to all prior patents or contain (a combination of) words or topics that appear for the first time (e.g., Kaplan and Vakili, 2015; Balsmeier *et al.*, 2017; Arts and Fleming 2017). By using text to identify technological novelty, it is possible

to avoid the use of patent (sub)classes (e.g., Arts and Veugelers, 2015) or citations (e.g., Dahlin and Behrens, 2005), which may suffer from examiner bias, are correlated with patent value, and are subject to temporal changes.

Second, we demonstrate how case-control matching based on text improves on matching based on the United States Patent Classification System, and reduces the likelihood of false positives or type I errors. While many prior studies underline the potential accuracy issues related to matching on the USPC (e.g., Thompson and Fox-Kean, 2005; Benner and Waldfogel, 2008; Belenzon and Schankerman, 2013), we are, to our knowledge, the first to estimate the likelihood of type I errors by means of expert assessments across five different fields. Scholars should be aware of the potential bias in prior research and could use our method and data to replicate key findings in both economics and strategy (Ethiraj, Gambardella, and Helfat, 2016). Depending on the research question at hand, we would encourage future studies to use text matching — potentially in combination with USPC — to measure the similarity between patents or patent portfolios and to construct a case-control sample of patents.

Third, we provide open access to the code and data for all utility patents granted by the USPTO between 1976 and July 2013. In Appendix 1, we provide a description of the code. Appendix 2 describes the different data files. Code and data are available at https://dataverse.harvard.edu/dataverse/patenttext.

Besides contributing to the literature on strategy and innovation, our method could be used by various practitioners such as inventors, attorneys, patent examiners, and managers to search for closely related prior art, to assess the novelty of a patent, to identify R&D opportunities in less crowded areas, to detect in- or out-licensing opportunities, to map companies in technology space, and to find acquisition targets.

Our exercise has several limitations. First, we only used title and abstract to measure patent similarity. Yet, our text-based similarity measure explains a large proportion of the variation in expert ratings based on the full patent document. Hence, the title and abstract arguably provide the necessary information on the technological content of a patent. Nevertheless, future research could take the description of claims into account. Second, future work might consider more sophisticated similarity measures that take into consideration the number of times a certain keyword occurs in a single patent or in the full patent corpus. We experimented with cosine similarity for a random sample of 25,000 patent pairs with varying levels of Jaccard similarity and found a very high correlation between the two alternative similarity measures. However, researchers can use the list of cleaned keywords of each patent we provide to construct their own measures. Third, text matching has a number of limitations as shown by the feedback from the experts on the differences between our text-based similarity measure and their personal rating. Patents with few keywords with little discriminatory power increase the likelihood of false positives, while different spelling variants and synonyms increase the likelihood of false negatives. Different words might have the same meaning and the same word might have a different meaning in a different context. Moreover, it is difficult to correct for spelling errors. Arguably, these problems should be limited because patents have a relatively large number of keywords, making the occasional occurrence of synonyms, homographs or spelling errors less of a problem. Nonetheless, preprocessing and text-mining tools such as stemming, latent semantic analysis or probabilistic topic modeling might help to overcome some of these limitations. Finally, in line with the traditional method of constructing a case-control group of matched patents (Jaffe *et al.*, 1993), and because patent classifications change over time, we only compared patents filed during the same year. However, we see no reason to believe that restricting the comparison to patents filed in the same year would result in bias. It seems

unlikely that matching patents from different years would change the validation results. We provide open access to the source data so that others can calculate the similarity between any two patents or between any two groups of patents filed at different points in time. By more carefully measuring technological similarity without relying on examiner-given classifications, we hope our method opens up opportunities for future research.

REFERENCES

Agrawal A, Cockburn I, Rosell C. 2010. Not invented here? Innovation in company towns. *Journal of Urban Economics* **67**(1): 78-89.

Aharonson BS, Schilling MA. 2016. Mapping the technological landscape: Measuring technology distance, technological footprints, and technology evolution. *Research Policy* **45**(1): 81-96.

Ahuja G. 2000. Collaboration networks, structural holes, and innovation: a longitudinal study. *Administrative Science Quarterly* **45**(3): 425-455.

Alcacer J, Gittelman M. 2006. Patent citations as a measure of knowledge flows: The influence of examiner citations. *The Review of Economics and Statistics* **88**(4): 774-779.

Almeida P. 1996. Knowledge sourcing by foreign multinationals: patent citation analysis in the US semiconductor industry. *Strategic Management Journal*, Winter Special Issue **17**: 155-165.

Almeida P, Kogut B. 1999. Localization of knowledge and the mobility of engineers in regional networks. *Management Science* **45**(7): 905-917.

Arts S, Veugelers R. 2015. Technology familiarity, recombinant novelty, and breakthrough invention. *Industrial and Corporate Change* **24**(6): 1215-1246.

Arts S, Fleming L. 2017. Paradise of Novelty – or Loss of Human Capital? Exploring New Fields and Inventive Output. Paper presented at the annual DRUID conference, New York.

Audretsch DB, Feldman MP. 1996. R&D spillovers and the geography of innovation and production. *The American Economic Review*: 630-640.

Balsmeier B, Fierro G, Li G, Johnson K, Kaulagi A, O'Reagan D, Yeh B, Lueck S. 2017. Machine learning and natural language processing applied to the patent corpus. Working paper.

Belenzon S, Schankerman M. 2013. Spreading the word: geography, policy, and knowledge spillovers. *Review of Economics and Statistics* **95**(3): 884-903.

Benner M, Waldfogel J. 2008. Close to you? Bias and precision in patent-based measures of technological proximity. *Research Policy* **37**(9): 1556-1567.

Bloom N, Schankerman M, Van Reenen J. 2013. Identifying technology spillovers and product market rivalry. *Econometrica* **81**(4): 1347-1393.

Chamberlain G. 1980. Analysis of Covariance with Qualitative Data. *The Review of Economic Studies* **47**(1): 225-238.

Dahlin KB, Behrens DM. 2005. When is an invention really radical?: Defining and measuring technological radicalness. *Research Policy* **34**(5): 717-737.

Ethiraj S, Gambardella A, Helfat C. 2016. Replication in strategic management. *Strategic Management Journal* **37**(11): 2191-2192.

Grossman GM, Helpman E. 1991. *Innovation and Growth in the Global Economy*. MIT Press: Cambridge, MA.

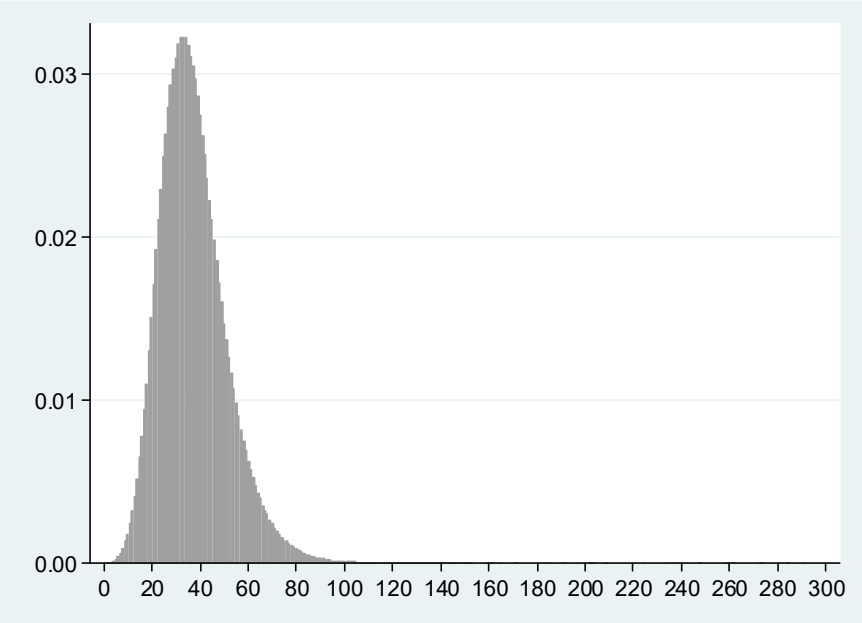
Hall B, Jaffe A, Trajtenberg M. 2002. The NBER patent citations data file: lessons, insights and methodological tools. In *Patents, Citations, & Innovations: A Window on the Knowledge Economy*, Jaffe A, Trajtenberg M. (eds.). MIT Press: Cambridge, MA: 403-459.

Jaffe AB, Trajtenberg M, Henderson R. 1993. Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations. *Quarterly Journal of Economics* **108**(3): 577-98.

Kaplan S, Vakili K. 2015. The double-edged sword of recombination in breakthrough innovation. *Strategic Management Journal* **36**(10): 1435-1457.

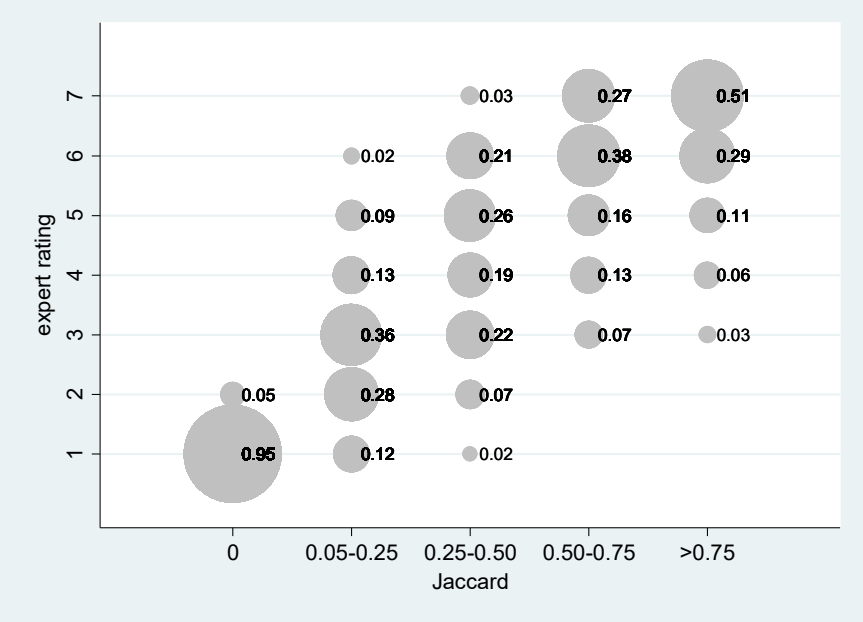
Krugman PR. 1991. *Geography and Trade*. MIT Press: Cambridge, MA.

Lampe R. 2012. Strategic citation. *Review of Economics and Statistics* **94**(1): 320-333.
Lemley MA, Sampat B. 2012. Examiner characteristics and patent office outcomes. *Review of Economics and Statistics* **94**(3): 817-827.
Li GC, Lai R, D'Amour A, Doolin DM, Sun Y, Torvik VI, Yu AZ, Fleming L. 2014. Disambiguation and co-authorship networks of the US patent inventor database 1975–2010, *Research Policy* **43**(6): 941-955.
Magerman T, Van Looy B, Debackere K. 2015. Does involvement in patenting jeopardize one's academic footprint? An analysis of patent-paper pairs in biotechnology. *Research Policy* **44**(9): 1702-1713.
Makri M, Hitt MA, Lane PJ. 2010. Complementary technologies, knowledge relatedness, and invention outcomes in high technology mergers and acquisitions. *Strategic Management Journal* **31**(6): 602-628.
McNamee RC. 2013. Can't see the forest for the leaves: similarity and distance measures for hierarchical taxonomies with a patent classification example. *Research Policy* **42**(4): 855-873.
Mokyr J. 1990. *The lever of the riches: Technological Creativity and Economic Progress*. Oxford University Press.
Murata Y, Nakajima R, Okamoto R, Tamura R. 2014. Localized knowledge spillovers and patent citations: A distance-based approach. *Review of Economics and Statistics* **96**(5): 967-985.
Romer PM. 1986. Increasing returns and long-run growth. *The Journal of Political Economy*: 1002-1037.
Rosenkopf L, Almeida P. 2003. Overcoming local search through alliances and mobility. *Management Science* **49**(6): 751-766.
Singh J, Agrawal A. 2011. Recruiting for ideas: how firms exploit the prior inventions of new hires. *Management Science* **57**(1): 129-150.
Singh J, Marx M. 2013. Geographic constraints on knowledge spillovers: political borders vs. spatial proximity. *Management Science* **59**(9): 2056-2078.
Thompson P, Fox-Kean M. 2005. Patent citations and the geography of knowledge spillovers: a reassessment. *American Economic Review*: 450-460.
Thompson P. 2006. Patent citations and the geography of knowledge spillovers: evidence from inventor-and examiner-added citations. *The Review of Economics and Statistics* **88**(2): 383-388.
Younge K, Kuhn J. 2016. Patent-to-Patent Similarity: A Vector Space Model, working paper.



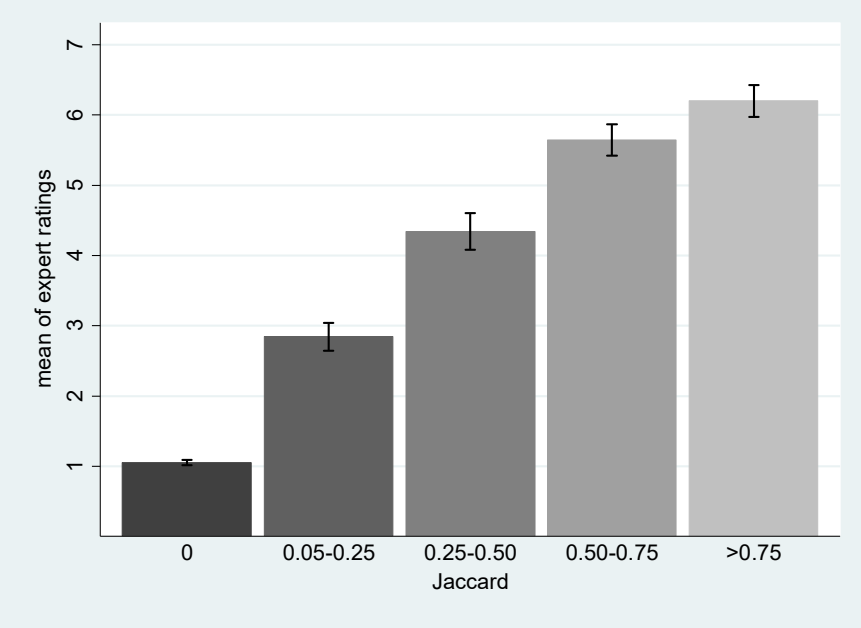
Notes: 4,422,009 U.S. utility patents, granted in 1976–2013

Figure 1. Histogram number of unique keywords per patent



Notes: The sample is constructed by selecting for each field of expertise a random sample of baseline patents, and for each baseline patent five random patents with varying degrees of Jaccard similarity (0, 0.05–0.25, 0.25–0.50, 0.50–0.75, 0.75 onwards). The order of the patent pairs is randomized, and the experts rate the similarity of the patent pairs in their field on a Likert scale from 1 to 7. Matched patents from the same patent family are excluded. The sample consists of 574 ratings conducted by eleven experts from three different fields. The size of the plotted circles is proportional to the share of patent pairs with a certain expert rating among the subset of pairs with a certain Jaccard similarity.

Figure 2. Expert ratings of technological similarity for text-matched patent pairs

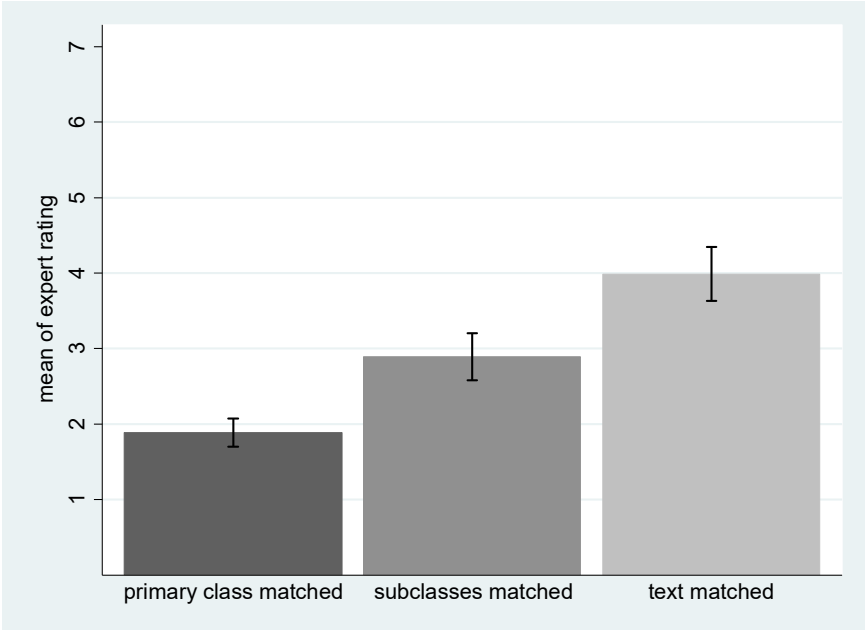


Notes: The sample is constructed by selecting for each field of expertise a random sample of baseline patents, and for each baseline patent five random patents with varying degrees of Jaccard similarity (0, 0.05–0.25, 0.25–0.50, 0.50–0.75, 0.75 onwards). The order of the patent pairs is randomized, and the experts rate the similarity of the patent pairs in their field on a Likert scale from 1 to 7. Matched patents from the same patent family are excluded. The sample consists of 574 ratings conducted by eleven experts from three different fields. The figure displays the means of expert ratings and the corresponding 95 percent confidence intervals for the five different subgroups of patent pairs with varying degrees of Jaccard similarity.

Figure 3. Means of expert ratings of technological similarity for text-matched patent pairs

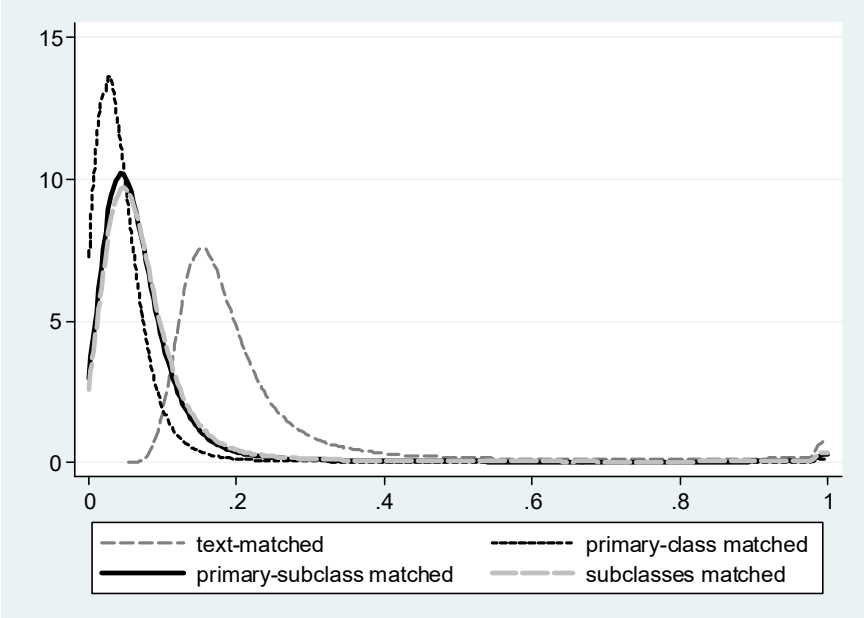
Notes: The sample is constructed by selecting for each field of expertise a random sample of baseline patents, and for each baseline patent a primary-class-matched, a subclasses-matched, and a text-matched patent. The order of the patent pairs is randomized, and the experts rate the similarity of the patent pairs in their field on a Likert scale from 1 to 7. Matched patents from the same patent family are excluded. The sample consists of 297 ratings conducted by eleven experts from three different fields. The size of the plotted circles is proportional to the share of patent pairs with a certain expert rating among the subset of primary-class-matched, subclasses-matched, and a text-matched patent pairs respectively.

Figure 4. Expert ratings of technological similarity of primary-class-matched, subclasses-matched, and text-matched patent pairs



Notes: The sample is constructed by selecting for each field of expertise a random sample of baseline patents, and for each baseline patent a primary-class-matched, a subclasses-matched, and a text-matched patent. The order of the patent pairs is randomized, and the experts rate the similarity of the patent pairs in their field on a Likert scale from 1 to 7. Matched patents from the same patent family are excluded. The sample consists of 297 ratings conducted by eleven experts from three different fields. The figure displays the means of expert ratings and the corresponding 95 percent confidence intervals for the primary-class-matched, subclasses-matched, and text-matched patent pairs respectively.

Figure 5. Means of expert ratings of technological similarity of primary-class-matched, subclasses-matched, and text-matched patent pairs



Notes: The sample only includes baseline patents for which both a text-matched patent, a primary-class-matched patent, a primary-subclass-matched patent, and a subclasses-matched patent are found. The sample consists of 3,492,480 text-matched patent pairs, 3,492,480 primary-class-matched patent pairs, 3,492,480 primary-subclass-matched patent pairs, and 3,492,480 subclasses-matched patent pairs.

Figure 6. Kernel density plot Jaccard similarity

Table 1. Summary statistics expert ratings of technological similarity

	(1) Mean	(2) Median	(3) Stdev	(4) Min	(5) Max
Jaccard=0	1.053	1.000	0.224	1.000	2.000
Jaccard>=0.05 and<0.25	2.843	3.000	1.189	1.000	6.000
Jaccard>=0.25 and <0.50	4.339	4.000	1.389	1.000	7.000
Jaccard>=0.50 and <0.75	5.643	6.000	1.207	3.000	7.000
Jaccard>=0.75	6.200	7.000	1.036	3.000	7.000

Notes: The sample is constructed by selecting for each field of expertise a random sample of baseline patents, and for each baseline patent five random patents with varying degrees of Jaccard similarity (0, 0.05–0.25, 0.25–0.50, 0.50–0.75, 0.75 onwards). The order of the patent pairs is randomized, and the experts rate the similarity of the patent pairs in their field on a Likert scale from 1 to 7. Matched patents from the same patent family are excluded. The sample consists of 574 ratings conducted by eleven experts from three different fields.

Table 2. Regression of expert ratings of technological similarity on Jaccard similarity

	(1) Ordered logit	(2) Ordered logit baseline patent random effects	(3) OLS	(4) Baseline patent fixed effects	(5) OLS	(6) OLS	(7) OLS
Jaccard>=0.05and<0.25	4.974	5.419	1.790	1.787			
	(0.567)	(0.581)	(0.119)	(0.120)			
Jaccard>=0.25 and <0.50	7.018	7.660	3.287	3.281			
	(0.610)	(0.651)	(0.171)	(0.176)			
Jaccard>=0.50 and <0.75	8.767	9.565	4.590	4.583			
	(0.558)	(0.616)	(0.137)	(0.145)			
Jaccard>=0.75	9.790	10.803	5.147	5.207			
	(0.622)	(0.702)	(0.125)	(0.133)			
Jaccard					6.000		5.838
					(0.226)		(0.241)
Subclass similarity						4.290	0.527
						(0.497)	(0.431)
Constant			1.053	1.047	1.662	3.304	1.666
			(0.024)	(0.072)	(0.066)	(0.122)	(0.065)
Number of ratings	574	574	574	574	574	574	574
Number of baseline patents	45	45	45	45	45	45	45
Log pseudolikelihood	-697.158	-683.408					
R-squared			0.753	0.766	0.702	0.171	0.704
	chi2(4)= 407.13	chi2(4)= 337.11	F(4,44)= 680.84	F(4,44)= 543.06	F(1,44)= 706.46	F(1,44)= 74.59	F(2,44)= 362.50

Notes: The outcome variable is the expert rating of technological similarity for a pair of patents on a Likert scale from 1 to 7. The explanatory variables in models (1) to (4) are binary indicators equal to one in cases where the Jaccard index of the patent pair lies respectively between 0.05–0.25, 0.25–0.50, 0.50–0.75, or is equal to or larger than 0.75. Jaccard index equal to zero is the excluded category. In models (5) and (7), the raw Jaccard index is used as an explanatory variable. Models (6) and (7) use the subclass similarity measure (number of unique subclasses two patents have in common divided by total number of unique subclasses in the union). Robust standard errors in parentheses, clustered at the baseline patent level. All coefficients are significantly different from each other at the one percent level across all models. Results are robust to clustering standard errors at the expert level, and to using expert-level random and fixed effects (results not shown).

Table 3. Summary statistics for subsamples of text-matched patent pairs with varying degrees of Jaccard similarity

	(1) Jaccard=0	(2) Jaccard>=0.05 and<0.25	(3) Jaccard >=0.25 and <0.50	(4) Jaccard >=0.50 and <0.75	(5) Jaccard >=0.75
	Mean	Mean	Mean	Mean	Mean
Jaccard	0.000	0.164	0.322	0.609	0.928
Binary: same patent family	0.000	0.002	0.024	0.086	0.407
Binary: same inventor(s)	0.000	0.083	0.496	0.867	0.927
Binary: same assignee(s)	0.001	0.148	0.651	0.866	0.865
Binary: citation link	0.000	0.008	0.044	0.085	0.085
N	4,386,405	3,426,228	601,947	137,551	220,679

Notes: 4,386,405 text-matched patent pairs for patents granted between 1976 and 2013. Each baseline patent is matched to the patent with the highest Jaccard index based on keywords and filed in the same year. In cases where there are multiple matches, patents are matched on approximate filing date. Patents with less than ten keywords are excluded and a minimum Jaccard of 0.05 is imposed. Column 1 includes an additional set of 4,386,405 patent pairs with no overlap in keywords, i.e. Jaccard index of zero, filed in the same year, and matched on approximate filing date. Unreported t tests indicate significant differences in same patent family, same inventor(s), same assignee(s), and citation link across the five different subsamples (columns 1–5). All differences are significant at the one percent level. The only non-significant difference is in same assignee(s) and citation link for pairs with Jaccard>=0.50 and <0.75 (column 4) versus pairs with Jaccard>=0.75 (column 5).

Table 4. Summary statistics expert ratings of technological similarity of primary-class-matched, subclasses-matched, and text-matched patent pairs

	(1) Mean	(2) Median	(3) Stdev	(4) Min	(5) Max
Primary-class-matched patent pairs	1.890	2.000	0.952	1.000	4.000
Subclasses-matched patent pairs	2.890	3.000	1.588	1.000	7.000
Text-matched patent pairs	3.990	4.000	1.800	1.000	7.000

Notes: The sample is constructed by selecting for each field of expertise a random sample of baseline patents, and for each baseline patent a primary-class-matched, a subclasses-matched, and a text-matched patent. The order of the patent pairs is randomized, and the experts rate the similarity of the patent pairs in their field on a Likert scale from 1 to 7. Matched patents from the same patent family are excluded. The sample consists of 297 ratings conducted by eleven experts from three different fields.

Table 5. Regression of expert ratings of technological similarity on indicators for primary-class matched, subclasses-matched, and text-matched patent pairs

	(1) Ordered logit	(2) Ordered logit baseline patent random effects	(3) OLS	(4) Baseline patent fixed effects
Subclasses-matched patent pairs	1.145	1.348	1.000	1.000
	(0.241)	(0.288)	(0.213)	(0.213)
Text-matched patent pairs	2.347	2.789	2.100	2.082
	(0.340)	(0.409)	(0.285)	(0.285)
Constant			1.890	1.896
			(0.107)	(0.137)
Number of expert ratings	297	297	297	297
Number of baseline patents	30	30	30	30
Log pseudolikelihood	-489.346	-469.189		
R-squared			0.250	0.326
	chi2(2)=52.18	chi2(2)=51.56	F(2,29)=29.38	F(2,29)=28.96

Notes: The outcome variable is the expert rating of technological similarity for a pair of patents on a Likert scale from 1 to 7. The explanatory variables are binary indicators for subclasses-matched and text-matched patent pairs. Primary-class matched patent pairs are the excluded category. Robust standard errors in parentheses, clustered at the baseline patent level. All coefficients are significantly different from each other at the one percent level across all models. Results are robust to clustering standard errors at the expert level, and to using expert-level fixed and random effects.

Table 6. Summary statistics for text-matched, primary-class-matched, primary-subclass-matched, and subclasses-matched patent pairs

	(1)	(2)			(3)			(4)		
	Text-matched patent pairs	Primary-class-matched patent pairs			Primary-subclass-matched patent pairs			Subclasses-matched patent pairs		
	(n=4,386,405)	(n=4,279,839)			(n= 3,492,480)			(n= 4,229,647)		
	Mean	Mean	|t|	Pr(|T| > |t|)	Mean	|t|	Pr(|T| > |t|)	Mean	|t|	Pr(|T| > |t|)
Jaccard	0.238	0.054	2200.000	0.000	0.092	1800.000	0.000	0.097	1900.000	0.000
Binary: Jaccard=0	0.000	0.120	-760.000	0.000	0.043	-390.000	0.000	0.040	-420.000	0.000
Binary: same patent family	0.028	0.005	308.439	0.000	0.012	255.467	0.000	0.013	239.195	0.000
Binary: same inventor(s)	0.207	0.037	889.674	0.000	0.079	690.830	0.000	0.085	677.012	0.000
Binary: same assignee(s)	0.276	0.059	992.268	0.000	0.114	752.643	0.000	0.118	743.565	0.000
Binary: citation link	0.019	0.002	267.639	0.000	0.008	165.982	0.000	0.013	92.972	0.000

Notes: t tests assess the mean difference between the text-matched pairs and the primary-class-matched, primary-subclass-matched, and subclasses-matched pairs in columns 2, 3, and 4 respectively. Only the subset of baseline patents for which both a text-matched and a primary-class-matched patent are found are used in the paired t test in column 2. Only the subset of baseline patents for which both a text-matched and a primary-subclass matched patent are found are used in the paired t test in column 3. Only the subset of baseline patents for which both a text-matched and a subclasses-matched patent are found are used in the paired t test in column 4.

Table 7. Geographic localization of citing and control patents

		Control patents	
	Text-matched	Primary-class matched	Primary-subclass matched
	(n=2,660)	(n=2,658)	(n=2,376)
	(1)	(2)	(3)
Match country	0.678	0.394	0.447
		(21.48)	(16.85)
Match state*	0.397	0.119	0.181
		(17.11)	(12.07)
Match cbsa†	0.346	0.085	0.144
		(15.65)	(10.89)
Distance in miles	2,287.011	3,449.055	3,403.809
		(-15.66)	(-13.86)

Notes: * Conditional on the citing patent having a U.S. inventor, † Conditional on the citing patent coming from a CBSA. The numbers display the geographic matching rates and spatial proximity in miles between the first inventors from respectively the citing patents and the text-matched control patents (column 1), the citing patents and the primary-class matched control patents (column 2), and the citing patents and the primary-subclass matched control patents (column 3). The numbers between parentheses give the t-statistic for the test of equality in geographic matching rates and spatial proximity between respectively the text-matched and the primary-class matched control patents (column 2), and between the text-matched and the primary-subclass matched control patents (column 3).

Table 8. T-tests for geographic localization of knowledge spillovers

	All citations	Inventor citations	Examiner citations	Text-matched control
	(n=27,377)	(n=15,960)	(n=11,417)	(n=15,960)
	(1)	(2)	(3)	(4)
Match country	0.543	0.589	0.479	0.506
			(18.04)	(14.96)
Match state*	0.111	0.116	0.101	0.099
			(2.85)	(8.96)
Match cbsa†	0.063	0.066	0.056	0.050
			(2.32)	(8.34)
Distance in miles	2,848.543	2,673.732	3,092.709	3,097.48
			(13.50)	(15.12)

Notes: Self-citations are excluded. * Conditional on the citing/text-matched patent having a U.S. inventor, † Conditional on the citing/text-matched patent coming from a CBSA. The numbers display the geographic matching rates and spatial proximity in miles between the first inventors from respectively the cited patents and the citing patents (column 1), the patents cited by the inventor(s) and the citing patents (column 2), the patents cited by the examiners and the citing patents (column 3), and the patents cited by the inventor(s) and the patents text-matched to the citing patents (column 4). The numbers between parentheses give the t-statistic for the test of equality in geographic matching rates and spatial proximity between inventor- and examiner-given citations (column 3), and between inventor-given citations and the text-matched control group (column 4). The numbers differ slightly from Thompson (2006) because we used a different dataset (Li *et al.*, 2014) to obtain disambiguated location information.

Table 9. Regressions for geographic localization of knowledge spillovers

	N	Inventor citation	Cited non-institutional	Cited patent age
		(1)	(2)	(3)
		Panel A: examiner-citation control		
Match country	21,872	1.211	1.360	0.990
		(3.77)	(5.23)	(-3.49)
Match state	11,797	1.304	1.104	0.966
		(3.10)	(1.11)	(-6.48)
Match cbsa	8,374	1.177	0.999	0.948
		(1.36)	(-0.00)	(-6.94)
Distance in miles	26,937	-118.818	-229.490	7.578
		(52.57)	(59.83)	(3.21)
		Panel B: text-match control		
Match country	31,920	1.402	1.552	0.988
		(23.53)	(9.52)	(-4.78)
Match state	21,210	1.468	0.937	0.972
		(11.46)	(-0.83)	(-6.36)
Match cbsa	17,916	1.669	0.727	0.973
		(10.40)	(-2.65)	(-4.42)
Distance in miles	31,920	-423.872	-485.957	4.609
		(19.83)	(52.06)	(2.830)

Notes: Self-citations are excluded. Panel A replicates Thompson (2006) and uses examiner-given citations as control. The models with match country, match state, and match cbsa as dependent variable are estimated with a citing-patent fixed-effects logit model (Chamberlain, 1980). Odds ratios are displayed and Z-scores between parentheses. The model with distance in miles as dependent variable is estimated with a linear citing-patent fixed-effects model. The estimated coefficients are displayed and T-scores between parentheses. The results differ slightly from Thompson (2006) because we used a different dataset (Li *et al.*, 2014) to obtain disambiguated location information. Panel B uses text-matched patents as control for inventor-given citations. The models with match country, match state, and match cbsa as dependent variable are estimated with a logit model. Odds ratios are displayed and Z-scores between parentheses. The model with distance in miles as dependent variable is estimated with OLS. The estimated coefficients are displayed and T-scores between parentheses.

ONLINE APPENDIX

Appendix 1: Description code

The code is written in Java and relies on Java standard libraries. We describe in general terms the different processes conducted and the data files created in order to compute the similarities between patents. We start the process in a working directory `[WD]` where we store the original data file `patents_raw.csv` and the subsequent generated files. `Patents_raw.csv` is a raw comma separated value file containing one row for each patent in the following format:

```
[Patent Number],[Filing Year],[Title],[Abstract]
```

As an example of the format, we show a patent in the file:

3631874,1970,FLUIDIC OVERSPEED SENSOR FOR A POWER TURBINE,A fluidic sensor having two parallel frequency-to-analog circuits whose output is summed to provide an error signal is disclosed.

Starting with the raw patent data file, we carry out the following steps:

1. We read each row from the data file `patents_raw.csv` and we do the following:
    a. Parse the row per sections using the comma separator.
    b. Concatenate the title and the abstract sections as a single content string.
    c. Lowercase the content string.
    d. Tokenize the content with `w[[-]w&&[^ ]]+w]` as the regular expression to extract keywords. This will match as keywords strings that are composed of any combination of characters[16] except _, and allows the hyphen (-) in order to consider compound keywords such as chemical names (e.g., bi-color, 4-di-n-oxide). Each unique keyword from the extraction is added to a list.
    e. Sort the list in ascending alphabetical order and clean it by removing stop words[17], keywords formed only by numbers (e.g., 1974, 1-3-4-4) and keywords with one character.
    f. Output the clean list of keywords as a row in the file `patents bow.txt` in the following format:
    `[Patent Number],[Filing Year],{List of Keywords}`
    where the keywords in the list are separated by a space.
    For example, extracting the keywords for the patent shown above, we have:

    3631874,1970,circuits disclosed error fluidic frequency-to-analog output overspeed parallel power provide sensor signal summed turbine

2. We read each row from the file `patents bow.txt` and we execute the following.
    a. Parse it per sections using the comma separator and take the list of keywords.
    b. Aggregate the unique keywords in a list to form a single vocabulary for the whole patent corpus, counting the number of rows (patents) in which each keyword occurs.

[16] See http://www.regular-expressions.info/shorthand.html
[17] See https://gist.github.com/shuson/b3051fae05b312360a18

3. After reading all the rows and forming a general vocabulary, we remove from the vocabulary keywords that occur in only one patent. Keywords appearing in only one patent are uninformative about similarity with other patents and are likely to be spelling errors.
4. We output in the file `vocabulary.txt` the list of keywords, one keyword per line.
5. We load the formed vocabulary in memory, we read each row from the file `patents_bow.txt` and then we carry out the following:
    a. Parse it per sections using the comma separator and take the list of keywords.
    b. Eliminate from the list of keywords the ones that do not appear in the vocabulary.
    c. Output the clean list of keywords as a row in the file `words.csv` in the following format:
    `[Patent Number],{List of Keywords}`
    For example, after cleaning the list of keywords from the patent shown above, we have:

    3631874,circuits disclosed error fluidic frequency-to-analog output overspeed parallel power provide sensor signal summed turbine

    d. Output the patent number and the filing year as rows in two individual files: patents_numbers.txt and patents_years.txt, respectively. There is a correspondence 1 to 1 between rows in files words.csv, patents_numbers.txt and patents_years.txt.
6. Using the files `patents_numbers.txt` and `patents_years.txt` we split the file `words.csv` per filing year, and thus create a set of files `patents_[year].txt`, in order to compute the similarities among patents in the same year.
7. For each file `patents_[year].txt` we take each row as a focus patent A, we parse it using the comma separator and take the list of keywords, and then we execute the following:
    a. Iterate over all the rows of the file `patents_[year].txt`, except the row of patent A, with each subsequent row being patent B, we parse it using the comma separator and take the list of keywords.
    b. Compute the Jaccard coefficient between patents A and B: $J(A,B) = |A \cap B|/|A \cup B|$, where A and B are the set of keywords of the corresponding patent.
    c. Store the pair of patents with their correspondent coefficient in the file `jaccard_[year].txt`, with the format:
    `[Patent Number A],[Patent Number B],[Jaccard Coefficient]`

If we wanted to compute similarities between patent portfolios, we would only need to concatenate the keywords from every patent in each portfolio and assign a portfolio number in a file `words_portfolio.csv`. We could then proceed in the same manner as from step 6.

Appendix 2: Description data files

The following data files are available from https://dataverse.harvard.edu/dataverse/patenttext

The first data file "***words.csv***" contains one row and two columns for each patent. The first column contains the patent number, the second column contains the set of unique and cleaned keywords separated by a space and ordered alphabetically. This database can be used to calculate the similarity between any pair of patents, or between two groups of patents by aggregating the keywords at the group level and calculating the similarity between the two groups.

The second data file "***closest match.csv***" contains for each patent the closest text-matched patent filed in the same year (with a minimum Jaccard index of 0.05). It consists of three columns: The first column contains the patent number of the baseline patent, the second column contains the patent number of the closest text-matched patent filed in the same year, and the third column contains the Jaccard index based on the overlap in keywords between the two patents. This database can be used to select a case-control group for a given set of patents.

The third data file "***200 closest matches.csv***" is constructed in an identical way as closest match.csv but contains the two hundred closest matches (with a minimum Jaccard index of 0.05) filed in the same year. It can be used to select a case-control group for a given set of patents conditional on a number of additional criteria that might exclude the single closest patent. For instance, the control patent needs to be assigned to a different firm or to a different set of inventors.

FACULTY OF ECONOMICS AND BUSINESS
DEPARTMENT OF MANAGERIAL ECONOMICS, STRATEGY AND INNOVATION
Naamsestraat 69 bus 3500
3000 LEUVEN, BELGIË
tel. + 32 16 32 67 00
fax + 32 16 32 67 32
info@econ.kuleuven.be
www.econ.kuleuven.be/MSI