In search of anti-commons: Patent-Paper Pairs in Biotechnology. An Analysis of Citation Flows.

Tom Magerman[1,2,3], Bart Van Looy[1,2,3] and Koenraad Debackere[1,2,3]

tom.magerman@econ.kuleuven.be
[1] Managerial Economics, Strategy and Innovation, Faculty of Business and Economics, K.U.Leuven, Leuven (Belgium)
[2] Research Division INCENTIM, K.U.Leuven, Leuven (Belgium)
[3] ECOOM, Leuven (Belgium)

Abstract
In this paper, we examine the possible presence of anti-commons dynamics in biotechnology by comparing citation patterns of patent-paper pairs, i.e. scientific publications from which the contents (subject, methodology, findings, discovery) is part of a patent application. Patent-paper pairs have been detected by relying on text mining algorithms. Starting from a dataset consisting of 948,432 scientific biotechnology publications and 88,994 EPO and USPTO biotechnology patent documents, a total of 584 patent-paper pairs have been identified. Forward citations patterns of those patent-paper pairs have been compared with biotechnology patents and publications that are not part of a patent-paper pairs. In terms of scientific citations, patent-paper pairs receive considerable more citations than publication without a patent counterpart. This is not the case for the technological impact of patent-paper pairs; patent citation rates do not differ significantly between patents with or without a scientific counterpart. As such our findings do not provide evidence for the presence of anti-commons effects stemming from the introduction of IP within scientific activities in the field of biotechnology.

Introduction: Entrepreneurial Universities
Collaboration between science and industry, and the phenomenon of 'enterprising universities', have been studied extensively over the last few decades. This growing interest is connected to the increasing acknowledgement of the fundamental role of knowledge and innovation in stimulating technological performance, international competitiveness and economic growth. Researchers in the domain of innovation (including Freeman, 1987 and 1994; Lundvall, 1992; Nelson, 1993; Nelson and Rosenberg, 1993; Mansfield and Lee, 1996; Mansfield, 1995; Mowery and Nelson, 1999; Dosi, 2000) stress the role of science and the importance of interaction between a variety of institutional actors underlying the innovative capacity and consequent economic performance of an economical system.
Contributing effectively to the innovative capacity of an innovation system requires a willingness of universities to become more 'entrepreneurial'. The notion of 'entrepreneurial universities' (Branscomb, Kodama and Florida, 1999; Etzkowitz, Webster and Healy, 1998) refers to the development of the following spectrum of activities: more intense commercialization of research results, patent and license activities, spin-off activities, collaboration projects with the industry, and greater involvement in economic and social development. Indeed, nowadays an increasing activity of academic researchers in exploiting their discoveries can be observed (Henderson et al., 1998; Thursby and Thursby, 2002; Meyer et al., 2003; Lissoni et al., 2008) and university patents become an important – and visible - method of technology transfer (Basberg, 1987; Boitani and Ciciotti, 1990; Trajtenberg et al., 1990; Archibugi, 1992).
At the same time some concerns arise due to the increasing commercialization of scientific activities undertaken by universities. While most empirical evidence – on the level of individual scientists – reports a positive relationship between patenting activities and publication outcomes (quantity as well as quality: Fabrizio and DiMinin, 2008; Van Looy et al., 2006, Breschi et al., 2007; Czarnitzki et al., 2007; Stephan et al., 2007), the expansion of IPR might still result in 'privatizing' the scientific commons and potentially limiting scientific progress (Argyres and Liebeskind, 1998; David, 2000; Krimsky, 2003). This fear is nicely

expressed by the metaphor of the "Tragedy the anti-commons", introduced by Heller (Heller, 1998). Heller states that the presence of too many owners with blocking power can lead to the underutilization of scarce resources, or, translated to the world of IPR, more intellectual property rights may lead paradoxically to fewer useful products (too many owners hold rights in previous discoveries creating obstacles for future research - see also Heller and Eisenberg, 1998).

Although anecdotal evidence exists of problematic use of IPR on scientific findings (e.g. the 'OncoMouse' or 'Havard mouse' of Leder and Stewart; or patents on human genes associated with breast and ovarian cancer owned by Myriad Genetics), large scale evidence of the presence of an anti-commons effect in biotechnology patenting is rare. One notable exception is the study of Murray and Stern (2007) suggesting a modest anti-commons effect based on a decline in citation rate – after granting of the patent - by 10 to 20% for a set of 169 patent-paper pairs published in Nature Biotech between 1997 and 1999.

In our study, we want to contribute to the research on an anti-commons effect in biotechnology by comparing forward citation patterns of patents and scientific publications for a large dataset containing all biotechnology patents (EPO and USPTO) and scientific publications (published in ISI Web of Science covered journals) from 1991 to 2008. First we investigate whether biotechnology publications for which a counterpart exists in the patent system (so called 'patent-paper pairs') are cited differently (more/less) within scientific journals, compared to similar biotechnology publications which are not related to a patent document. Next, we engage in a similar analysis focusing this time on 'technological' citations: to what extent are patents closely related to scientific publications cited differently by other patents compared to biotechnology patents without scientific counterpart. The former will allow us to shed some light on the fear that exploitation of scientific findings is hampering scientific development by pruning promising developments due to the introduction of (potentially blocking) patents. The latter will allow us to look at the technological impact of scientific developments that become translated into a patent.

An important methodological aspect for this kind of studies relates to the identification of patent-paper pairs, i.e. scientific publications from which the contents (subject, methodology, findings, discovery) is part of a patent application. To obtain a set of patent-paper pairs, we stepped down from a manually guided process of mapping patent and scientific publications and applied text mining algorithms to match automatically patents and scientific publications based on content similarity.

Within the next pages, we first outline the selection of the data used for this analysis, followed by a description of the methodology adopted to assess the similarity between patents and scientific publications. This section is followed by reporting the findings, for scientific citations and patent citations respectively. We conclude with outlining the limitations of our work and suggest avenues for further research in this area.

Data and methodology

Field selection

We focus on patents and scientific publications in the field of biotechnology because it is a field well known for the presence of science-technology linkages and because the large scale exploitation of biomedical research makes it more susceptible to an anti-commons effect (Heller and Eisenberg, 1998). Patents and publications are selected based on technological and scientific classification schemes respectively.

Selection of biotechnology patents

On the patent side, the OECD definition of biotechnology is used to identify biotechnology patents (OECD, 2005), defining 30 International Patent Classification subclasses/groups related to biotechnology. We use PATSTAT (EPO Worldwide Patent Statistical Database) to retrieve all EPO and USPTO granted patents with application and grant year between 1991 and 2008 according to the 30 defined IPC-subclasses/groups related to biotechnology. This led to a set of 27,241 EPO and 91,775 USPTO patents (PATSTAT edition October 2009). As text mining techniques are applied for the further identification of patent-paper pairs, only patents with titles and a minimum abstract length of 250 characters were withheld, resulting in a final patent data set of 7,254 EPO and 80,994 USPTO biotechnology patents (hence 88,248 patents in total).

Selection of scientific publications

On the publication side, we select biotechnology publications (articles, letters, notes, reviews)[53] from the Thomson Reuters ISI Web Of Science database based on the Web of Science subject classification, for the same time period 1991-2008 (volume year). 243,361 publications are revealed from subject category 'Biotechnology and Applied Microbiology'. However, to ensure that all potentially related scientific publications are present in the data set, we extend this 'core' publication set with publications from nine related subject categories: 'Biochemical Research Methods'; 'Biochemistry & Molecular Biology'; 'Biophysics'; 'Plant sciences'; 'Cell Biology'; 'Developmental Biology'; 'Food sciences & Technology'; 'Genetics & Heredity' and 'Microbiology Materials'[54] which are cited or citing this 'core' set (683,674). Finally we also add all multidisciplinary publications from 'Nature', 'Science' and 'Proceedings of the National Academy of Sciences of the United States of America', resulting in 97,970 additional publications. Again we only retain publication documents with titles and a minimum abstract length of 250 characters, resulting in a final publication set of 948,432 biotechnology related publications.

Text mining oriented identification of patent-paper pairs

The identification of patent-paper pairs is based on the content similarity of titles and abstracts of patents and publications. For all patents, the similarity with all publications is derived based on content similarity metrics. An extensive study has been set up to derive the best content based metric to detect patent-paper pairs; 40 measures based on LSA/SVD[55] (4 weighting methods in combination with 10 levels of dimensionality reduction and a cosine metric) and 3 measures based on the number of common terms were compared and thoroughly validated to check whether derived similarity measures really grasp content relatedness in terms of subject, methodology, findings and discovery mentioned in the patent and publication documents (see Magerman et al., 2011)[56].

In practice, first all titles and abstracts of all biotechnology patents and scientific publications where indexed[57]. During indexing, a limited number of stop words were removed[58], stemming

[53] Articles are by far the biggest category (90% articles compared to 1.5% letters, 2% notes and 6.5% reviews)

[54] The authors want to thank Wolfgang Glänzel for his kind help in the development of a search strategy for biotechnology publications.

[55] Landauer, T. K., McNamara, D. S., Dennis, S. & Kintsch, W. (Eds.) (2007). *Handbook of Latent Semantic Analysis.* Mahwah (NJ): Lawrence Erlbaum Associates

[56] It turned out that simple measures based on the number of common terms outperform LSA/SVD based cosine measures when it comes to identifying patent-paper pairs.

[57] Apache-Lucene™, an open source text search engine library, was used for the indexing (http://lucene.apache.org/java/docs/index.html)

[58] Based on the Snowball English stop word list (http://snowball.tartarus.org/algorithms/englisch/stop.txt)

was applied (Porter stemmer) and all terms only occurring once in the corpus were removed. Next, a vector space (Salton, 1975) was created based on the full text index[59]. This vector space consists of a document by term matrix, whereby rows are defined by all included documents and columns consists of all (stemmed) terms identified within the set of documents. Overall, the matrix used in this analysis consists of 1,066,632 rows (documents) and 301,697 columns (terms). Two metrics were combined for the classification of patent-paper combinations. The number of common terms, divided by the minimum of the number of terms of the patent document on the one hand and of the publication document on the other hand, is used for a first selection of patent-paper combinations with significant content similarity (CommonTermsMin $\geq$ 0.60). A second criterion, based on the number of common terms divided by the maximum of the number of terms of the patent document and publication document, is used to filter out ambiguous cases (CommonTermsMax $\geq$ 0.30). These two content-based criteria are combined with an additional criterion: at least one of the patent inventors has to be listed as a publication author. Together those three criteria allow an accurate identification of patent-paper pairs (precision equals to 0.96 and recall equals to 0.84 up to 0.90 – depending of using a conservative or optimistic definition of similarity - based on the data of the validation sample of 300 cases). For a more elaborate discussion of the technical details on content based similarity measures best suited for the identification of patent-paper pairs, we refer to Magerman et al., (2010, 2011).

Identified patent-paper pairs

The starting point for the identification of patent-paper pairs is the combined dataset of 88,248 biotech patents and 948,432 biotech publications[60]. Application of the first matching criterion, a content similarity of at least 0.60 based on the number of common terms weighted for the minimum of the number of terms of both documents, yields 27,250 related patent-paper combinations out of the more than 80 billion combinations under examination. Application of the second matching criterion, a content similarity of at least 0.30 based on the number of common terms weighted for the maximum of the number of terms of both documents, resulted in 645 patent-paper pairs. Application of the last criterion, at least one patent inventor being listed as a publication author, resulted in a final set of 584 patent-paper pairs. 17 patents are matched with multiple publications (up to three publications), which seems to be cases of (partly) disclosure of the same results in multiple scientific articles. At the same time, 115 publications are matched to multiple patents (up to seven patents), which revealed to be members of the same patent family. Hence we have 566 distinct biotechnology patents having a paired biotechnology publication, and 400 distinct biotechnology publications having a paired biotechnology patent.

Note that we deliberately opted for a very conservative selection to identify patent-paper pairs. Especially the second criterion filters out a lot of ambiguous cases, so we can be confident that the described patent-paper matching method reveals real patent-paper combinations. We also want to stress that, although identification of pairs is based on the number of common terms, thorough validation proved that this metric is truly able to grasp the real similarity between patent and publication document in terms of subject, methodology, findings and discovery mentioned in the patent and publication document (again see Magerman et al., 2011).

[59] MatWorks Matlab™, a commercial packet for technical computing, was used for the construction of the vector space and further data handling and similarity calculation (http://www.mathworks.com/products/matlab/). The authors want to thank Frizo Janssens who was so kind to share his propriety Matlab code for the import of the full text index into a document-by-term matrix.

[60] Only patents and publications with titles and abstracts of sufficient length are retained to allow for content-based matching.

Findings on citation patterns of scientific publications (publication-to-publication citations)

Within this section we report and discuss the empirical results obtained when analysing scientific citations - i.e. citations from other scientific articles - to scientific publications that are part of a patent-paper pair (publication-to-publication citations). This analysis implies a comparison with scientific citations to scientific articles which do not belong to a patent-paper pair.

Descriptive Statistics.

Table 1 provides a summary overview of the number of biotechnology publications under study as well as the observed forward citations, organized by publication year.
Within the period 1991-2000, we have identified 328 publications that are part of a patent-paper pair. On average we clearly observe substantially higher forward citation counts for publications that are part of a patent-paper pair and other publications (mean of 161.8 versus 46.9, median of 65 versus 20). But not only the average numbers are higher, the complete distribution of forward citation counts is shifted to the right in favour of publications that are part of a patent-paper pair as becomes apparent in figure 1.

Table 1. Number of biotechnology publications and forward citations per year (all publication matching our search key without elimination of publications having no or small abstract)

YEAR	*NUMBER OF BIOTECHNOLOGY PUBLICATIONS*	*NUMBER OF FORWARD PUBLICATION CITATIONS*	*AVERAGE NUMBER OF FORWARD CITATIONS*
1991	31,381	1,585,560	50.53
1992	35,185	1,734,412	49.29
1993	38,677	1,913,155	49.46
1994	42,764	2,014,535	47.11
1995	48,092	2,210,601	45.97
1996	50,788	2,256,455	44.43
1997	53,175	2,441,374	45.91
1998	57,361	2,638,305	45.99
1999	59,866	2,739,699	45.76
2000	61,072	2,877,433	47.12
	478,361	**22,411,529**	**46.85**

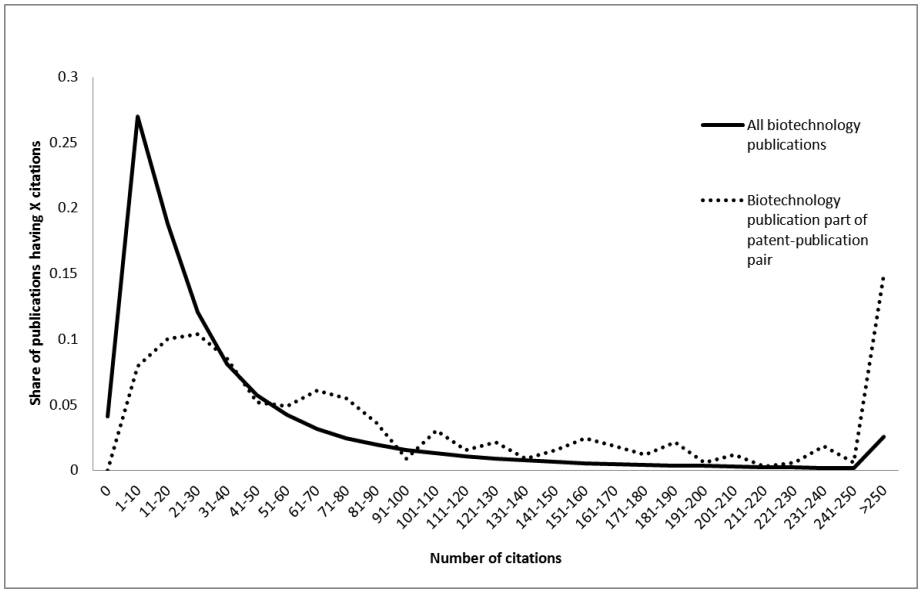

Figure 1. Distribution of the number of forward publication citations for all biotechnology publications and biotechnology publication part of a patent-paper pair (1991-2000).

Figure 1 shows the distribution of the number of forward publication citations for all biotechnology publications and biotechnology publication part of a patent-paper pair for the period 1991-2000. 25% of paired biotechnology publications have 27 or less citations compared to 7 or less citations for the first quartile for all biotechnology publications; 50% of paired biotechnology publications have 65 citations or less (20 citations for all biotechnology publications) and 75% of paired citations have 160 citations (48 citations for all biotechnology publications).

At the right side of the distribution we observe substantial outliers, especially for publications that are related to a patent.

One potential explanation for the higher number of forward citations might be the difference in number of authors. Publications having more authors tend to have more forward citations - as is confirmed by our data (an average of 38 forward citations for single authored papers up to 46 citations for publications with 5 authors and 86 citations for publications with 10 authors)[61]. Another, more important, consideration when observing the difference in forward citation counts is the presence of a selection bias for paired publications towards "higher quality" publications. For the large overall biotechnology publication sample, all kind of quality levels will be present in the dataset. For publications that are part of a patent-paper pair, one can expect to find more publication of higher quality than average, i.e. publications valuable enough to justify costs and effort to apply for a patent. We correct this by taking into account the journal in which publications are published as an indication of the quality level of publications (i.e. we assume underlying journal impact factors are a good indication of the average quality of publications appearing in that journal).

Multivariate Analysis

To verify the significance of the observed difference when controlling for other factors multivariate analysis have been performed. Given the nature of the data (citation data) we opted for a negative binomial regression with the number of forward citations as dependent

[61] 78% of the biotechnology publications in our sample have 5 or less authors, 20% have 6 to 10 authors.

variable and a dummy variable indicating whether a publication is part of a patent-paper pair as independent variable.

To adjust for the expected difference in average quality between paired and non-paired publications (due to the potential selection bias of publications that are part of a patent-paper pair), we only include publications from journals that have at least one paired publication, i.e., we only use publications that are comparable in average impact factor because they originate from the same set of journals. For this analysis, we use net citation counts, i.e. citations counts corrected for self-citations, as independent variable. This leaves 400 biotechnology publications that are part of a patent-paper pairs, and 451,803 biotechnology publications that are not part of a patent-paper pair.

We further control for journal of publications (105 distinct journals), publication document type (article, letter, note, review), number of backward publication-to-publication citations, and finally, the number of authors. We also include a time variable (1 for the first year, 1991, up to 18 for the last year, 2008) and a squared time variable to accommodate evolutions over time.

Table 2 reports the results of the regression analysis of forward publication citations of publications. Publications being part of patent-paper pairs have significantly more forward publication citations (Pair Y/N). One also notices a positive relationship between forward citations and the number of authors as well as the number of backward citations. Citation rates differ between document types: reviews receive more citations compared to articles, letters and notes. The number of forward citations differ significantly between journals (journal dummies have been included, but not reported, n=104). Finally, the observed citation rates reflect an inverse U pattern over time.

When removing outliers, i.e. all publications with a forward citation count larger than the mean plus three times the standard deviation, similar results are obtained then the ones reported in Table 2.

Table 2. Results of negative binomial regression - Number of forward publication citations of publications (net, i.e. with exclusion of self-citations) (1991-2008)

Parameter Estimates

Parameter	*B*	*Std. Error*	*95% Wald Confidence Interval*		*Hypothesis Test*		
			Lower	*Upper*	*Wald Chi-Square*	*df*	*Sig.*
(Intercept)	2.966	.1258	2.719	3.213	555.643	1	.000
Pair (Y/N)	.450	.0506	.350	.549	78.945	1	.000
Document type:							
Article	-.574	.0113	-.596	-.552	2589.688	1	.000
Letter	-.774	.0590	-.890	-.659	172.469	1	.000
Note	-.567	.0175	-.601	-.533	1051.989	1	.000
Review	0	.	.	.	.	.	.
Number of backward publication citations	.013	.0001	.013	.014	10416.453	1	.000
Number of authors	.033	.0005	.032	.034	4613.407	1	.000
Time	.125	.0015	.122	.128	7191.199	1	.000
Time²	-.012	.0001	-0.13	-.012	29450.994	1	.000
Journal dummies (n=105)	Included						

Comparison of citation counts before and after patent grant

Inspired by the observations of Murray and Stern (2007) - a relative decline in citation patterns after patents have been granted – we verify whether the citation rates differ before and after a patent has been granted. We follow the reasoning of Murray and Stern stating that if a patent grant comes to a complete surprise to follow-on researchers, i.e. if researchers that continue working on previous discoveries are not aware of pending patent applications on those previous discoveries, a drop in citation rate can be an indication of the presence of an anti-commons effect. The reasoning behind this construct is that if researchers are not aware whether a given piece of knowledge is subject to patent filing, they will use (cite) this knowledge (publication) in a normal way. As soon as a patent covering that piece of knowledge is granted, those follow-on researchers might stop using (citing) this knowledge because of the perceived "price" (patent rights) of building on the prior discovery. Hence in case of the presence of an anti-commons effect, forward citations of publications that are part of a patent-paper pair are expected to drop as soon as the corresponding patent is granted.

To test this we split up forward citation counts for all pairs into the number of citations received before and after the grant year of their corresponding patent[62] [63] [64]. These numbers are aggregated at the level of journals and publication years, resulting in two average citations counts; one for the pre-grant period, and one for the post-grant period. Next, for every observed journal and publication year, we construct a control group that consists of all non-paired biotechnology publications published in that given journal and year. For these publications, forward citations are split up in exactly the same manner as to reflect the pre- and post-grant period. This is done as follows: if for a given journal and publication year only one paired publication is present, we split citation counts for all non-paired publications published in the same year and journal based on the lag between the publication year (journal) and grant year (corresponding patent). This is again aggregated at the level of the journal and publication year, resulting in an average citation count pre- and post-grant for non-paired patents for the given journal and publication year. If a given journal has multiple publication with a paired patent in a given publication year, we split up forward publication citation counts for the non-paired publications multiple times, once for every lag between the publication year and the grant year of the corresponding patents. All these number are aggregated at the level of the journal and publication year, resulting in an average citation count pre- and post-grant for non-paired patents. Finally, for all combinations of journals and publication years in which pairs have been observed, we calculate the ratio between citation received by pairs versus non-pairs two times: for the pre-grant period as well as the post-grant period. If an anti-commons effect would manifest itself, after granting the patent, the ratio between pairs and non-pairs would drop significantly.

As table 3 indicates, the ratio of average citations received by pairs versus non-pairs equals to 1,71 and 1,74 before and after granting respectively. Controlling for journal and publication year, this figure means that papers that are part of a pair receive on average 71% and 74% more citations than their counterparts not belonging to a pair. While these descriptive statistics do not indicate a decline, a formal t-test reveals that both ratios are not significantly

[62] For those publications linked to multiple patents (multiple members of patent families), the earliest patent grant data was used to split citations into a pre-grant and post-grant period.

[63] For this analysis, the total number of citations was used, not the net number of citations (excluding self-citations)

[64] Only publications of period 1991-2000 are included to have a full 10-year citation window for all publications and to make use of the fact that USPTO applications were not made public before 2001, making the changes of a 'surprise' grant to follow-up researchers more likely.

different (p=0,86). As such, our data do not show any sign of anti-common effects that become visible after patent rights have been granted.

Table 3. Results of independent T-test – Ratio average citations pairs/non-pairs pre-grant versus post-grant (1991-2000)

VARIABLE	*CLASS*	*N*	*LOWER CL MEAN*	*MEAN*	*UPPER CL MEAN*
Ratio average citations pairs/non-pairs	Pre-grant	288	1.42	1.71	2.00
Ratio average citations pairs/non- pairs	Post-grant	288	1.48	1.74	2.00
Diff	(1-2)		-0.43	-0.03	0.36

Findings on citation patterns of patents (patent-to-patent citations)

Within this section we report and discuss the empirical results obtained when analysing patent citations - i.e. citations from other patent document - to patent documents that are part of a patent-paper pair (patent-to-patent citations). This analysis implies a comparison with patent citations to patent documents which do not belong to a patent-paper pair.

Descriptive results

Table 4 provides a summary overview of the number of biotechnology patents under study as well as the observed average number of forward patent citations, organized by application year, for all biotechnology patents and for the paired biotechnology patents.
Within this period (1991-2000) we have identified 435 patent-paper pairs. For the period 1991-2000, the average number of forward citations is 9.5 (median is equal to 4). These averages are about 8% lower compared to non-paired patents (11.38).

Table 4. Number of biotechnology patents and forward citations per year (only patents with substantial abstract)

	ALL BIOTECHNOLOGY PATENTS		*PAIRED BIOTECHNOLOGY PATENTS*	
APPLICATION YEAR	*NUMBER OF PATENTS*	*AVERAGE NUMBER OF FORWARD PATENT CITATIONS*	*NUMBER OF PATENTS*	*AVERAGE NUMBER OF FORWARD PATENT CITATIONS*
1991	3,069	16.21	9	14.56
1992	3,727	16.14	11	24.09
1993	4,392	16.01	25	12.68
1994	6,170	14.39	37	11.16
1995	9,881	14.60	71	13.51
1996	5,635	12.13	33	6.45
1997	7,097	10.12	56	11.68
1998	6,974	8.30	70	8.84
1999	7,742	7.35	58	5.47
2000	7,798	5.46	65	3.52
TOTAL	**62,485**	**11.38**	**435**	**9.46**

As can be expected, patent-paper pairs are largely related to academic patenting; 52% of biotechnology patents that are linked to a publication have at least one academic patentee, compared to 18% for all non-paired biotechnology patents. Patents with at least one government or non-profit patentee are also overrepresented in the set of patents closely related to publications (23% for paired patents versus 10% for non-paired patents).

Multivariate Analysis

In order to assess whether observed differences are statistically significant, we performed a negative binomial regression with the number of forward patent-to-patent citations as dependent variable and a dummy variable indicating whether a patent is or is not part of a patent-paper pair as independent variable. We use all 88,248 biotechnology patents having a substantial abstract (566 patents that are part of a patent-paper pair and 87,682 patents that are not part of a patent-paper pair).

We further control for the patent system (EPO or USPTO), the number of IPC codes (technological specialization), the presence of academia as patentee, the number of backward scientific non-patent citations, the number of backward patent citations, the number of forward publication citations (citations from Web of Science publications to the particular patent), the number of inventors and the number of patentees. We also included dummy variables for all 11 biotechnology IPC subclasses (4 digits) present in our selection of biotechnology patents (see Appendix A for all IPC-codes as used in the OECD biotechnology definition). Again we include a time variable (1 for the first year, 1991, up to 18 for the last year, 2008) and a squared time variable to include the evolution over time.

Table 5 reports the results of the regression analysis of forward patent citations of patents. Patents being part of a patent-paper pairs have more forward publication citations (variable Pair Y/N), but the difference is not significant. USPTO patents have more citations than EPO patents. All other controlling variables have a significant and positive impact, except for the number of patentees, which has a negative but not significant impact and time, which displays a decreasing, curvilinear relationship with patent citations.

Table 5. Results of negative binomial regression - Number of forward patent citations of patents (corrected for DOCDB patent family members, both at cited and citing side) (1991-2008)

Parameter Estimates

Parameter	*B*	*Std. Error*	*95% Wald Confidence Interval*		*Hypothesis Test*		
			Lower	*Upper*	*Wald Chi-Square*	*df*	*Sig.*
(Intercept)	2.300	.0197	2.262	2.339	13585.101	1	.000
Pair (Y/N)	.058	.0460	-.032	.148	1.599	1	.206
Patent system							
EPO	-.193	.0140	-.221	-.166	190.450	1	.000
USPTO	0	.	.	.	.	.	.
Subfield (IPC subclass dummies) (n=11)	Included						
Number of IPC codes	.043	.0008	.042	.045	3265.759	1	.000
Has university patentee (Y/N)	.036	.0097	.017	.055	13.462	1	.000
Number of backward scientific non-patent citations	.003	.0002	.003	.003	248.658	1	.000

Number of backward patent citations	.016	.0003	.016	.017	3370.269	1	.000
Number of forward publication citations from WOS publications	.141	.0052	.131	.152	744.627	1	.000
Number of inventors	.018	.0019	.014	.022	92.591	1	.000
Number of patentees	-.010	.0074	-.025	.004	1.945	1	.163
Time	-.063	.0042	-.071	-.055	228.478	1	.000
Time²	-.007	.0003	-.008	-.007	766.940	1	.000

After removing outliers, i.e. all patents with a forward citation count larger than the mean plus three times the standard deviation, similar results are obtained as the ones reported in Table 5. Finally, when we limit the time period to all patents applied for between 1991 and 2000 – in order to allow all patents to have at least 10 years of forward patent citations – patent-paper pairs have less forward patent citations, but also this difference is not significant (both when including and excluding outliers). Overall, we observe no significant difference in terms of (forward) patent citations when comparing patents that are associated with a scientific publication with their solitary counterparts.

Discussion and (intermediate) conclusions

In this paper, we have applied a text mining methodology to examine the possible presence of anti-commons effects in biotechnology research. Inspired by previous work undertaken by Murray, Stern and others, we analyse citation flows stemming from patent-paper pairs present within the field of biotechnology. The delineation of the biotechnology domain was based on the use and the refined application of existing classification schemes. An elaborate text mining scheme was developed and implemented in order to identify and validate the patent-paper pairs. A total of 584 pairs were ultimately included in the citation analysis. The necessary validation and control strategies were introduced and executed. After taking into account these controls and studying the citation patterns of the documents included in the patent-paper pairs, we were not able to detect a significant anti-commons effect on the basis of the 584 pairs identified. On the contrary, scientific papers belonging to a patent-paper pair receive significantly more *scientific* citations than their counterparts for which no patent document has been identified. This difference remains outspoken (and significant) after taking into account the granted nature of implied patent documents. As such, our findings do not reveal the presence of anti-commons effects once scientific findings become translated into intellectual property rights (in this case, patents). In terms of technological citations, we observed no difference between patents belonging to a patent-paper pair and patent documents that are not associated directly with a scientific publication. As such, no additional impact – on future technological developments - is observed when patent documents are situated in the vicinity of science.

These findings add to the current stock of insights on the interaction between patenting and publication behaviour. Through the design and application of text mining techniques on a broad set of data, we intended to take the current insights a step further. Extensive validation efforts were undertaken in order to confirm the results obtained.

These results definitely are an invitation to further examine the joint effects of patenting and publishing activities by scientists. The first point of attention that arises is the one of generalization towards other fields of 'techno-scientific' economical activity. Can we substantiate the current findings in technology domains such as materials or in other fields? The second point relates to corroborating and consolidating the robustness of the text mining methodology that was deployed, as well as a further, independent, confirmation of the optimal identification algorithm. The third point pertains to the continuous cross-validation of the

results obtained with our method with the results obtained by sets of patent-paper pairs that have been constructed manually by experts.

References

Archibugi, D. (1992). Patenting as an indicator of technological innovation: a review. *Science and Public Policy*, 19, 357-368.

Argyres, N. S. & Liebeskind, J. P. (1998). Privatizing the intellectual commons: universities and the commercialization of biotechnology. *Journal of Economic Behavior and Organization*, 35(4), 427-454.

Basberg, B. L. (1987). Patents and the measurement of technological change: A survey of literature. *Research Policy*, 16(2-4), 131-141.

Boitani A. & Ciciotti, E. (1990) Patents as Indicators of Innovative Performances at the Regional Level. In R. Cappellin and P. Nijkamp (Eds.), *The Spatial Context of Technological Developmen*. Aldershot: Avebury

Branscomb, L.M., Kodama, F. & Florida, R. (1999). *Industrializing Knowledge: University-Industry Linkages in Japan and the United States.* London: MIT Press.

Breschi, S., Lissoni, F. & Montobbio, F. (2007). The scientific productivity of academic inventors: New evidence from Italian data. *Economics of Innovation and New Technology*, 16(2), 101-118.

Czarnitzki, D., Glänzel, W. & Hussinger, K. (2007). Patent and publication activities of German professors: An empirical assessment of their co-activity. *Research Evaluation*,164, 311-319.

David, P. A. (2000). The digital technology boomerang: new intellectual property rights threaten global open science. Working Papers 00016, Stanford University, Department of Economics.

Dosi, G. (2000). *Innovation, Organization and Economic Dynamics.* Cheltenham: Edward Elgar Publishers.

Etzkowitz, H., Webster, A., & Healey, P. (1998). Capitalizing Knowledge: New Intersections of Industry and Academia Albany, State University of New York press.

Fabrizio, K. R. & Di Minin A. (2008). Commercialization the laboratory: Faculty patenting and the open science environment. *Research Policy*, 37(5), 914-931.

Freeman, C. (1987). *Technology Policy and Economic Performance.* London: Pinter.

Freeman, C. (1994). The economics of technical change. *Cambridge Journal of Economics* 18, 463-514.

Heller, M. A. (1998). The Tragedy of the Anticommons. *Harvard Law Review*, 111, 621-688.

Heller, M. A. & Eisenberg, R. S. (1998). Can Patents Deter Innovation? The Anticommons in Biomedical Research. *Science*, 280,698-701.

Henderson, R., Jaffe, A. B. & Trajtenberg, M. (1998). Universities as a source of commercial technology: A detailed analysis of university patenting. *Review of Economics and Statistics*, 80(1), 119-127.

Krimsky (2003). *Science and the Private Interest.* Rowman-Littlefield Publishing Co.

Lissoni, F., Llerena, P., McKelvey, M. & Sanditov, B. (2008). Academic patenting in Europe: New evidence from the KEINS database. *Research Evaluation*, 17(2), 87-102.

Lundvall, B. A. (1992). *National Systems of Innovation: Towards a Theory of Innovation and Interactive Learning.* London: Pinter Publishers.

Magerman, T., Van Looy, B. & Song, X. (2010). Exploring the feasibility and accuracy of Latent Semantic Analysis based text mining techniques to detect similarity between patent documents and scientific publications. *Scientometrics*, 82(2), 289-306.

Magerman, T., Van Looy, B. & Baesens, B. (2011). Assessment of LSA/SVD text mining algorithms to map patent and scientific publication documents. (forthcoming).

Mansfield, E. (1995). Academic research underlying industrial innovations: sources, characteristics, and financing. *The Review of Economics and Statistics*, 77(1), 55-56.

Mansfield, E. & Lee, J.Y. (1996). The modern university: contributor to industrial innovation and recipient of industrial support. *Research Policy*, 25, 1047-1058.

Meyer, M., Sinilainen, T. & Utecht, J. T. (2003). Towards hybrid Triple Helix indicators: A study of university-related patents and a survey of academic inventors. *Scientometrics*, 58(2), 321-350.

Mowery, D. C. & Nelson, R. R. (1999). *Sources of Industrial Leadership*. Cambridge: Cambridge University Press.
Murray, F. & Scott, S. (2007). Do formal intellectual property rights hinder the free flow of scientific knowledge? An empirical test of the anti-commons hypothesis. *Journal of Economic Behavior and Organization*, 63, 648-687.
Nelson, R. R. & Rosenberg, N. (1993). *Technical Innovation and National Systems*. In R. R. Nelson (Ed.), *National Innovation Systems. A comparative Analysis*: New York: Oxford University Press, Inc.
Nelson, R. R. (1993). *National Innovation Systems: A Comparative Analysis.* New York: Oxford University Press Inc.
OECD (2005). A Framework for Biotechnology Statistics, 29-32, Paris: OECD.
Rosenberg, N. (1998). Chemical engineering as a general purpose technology. In E. Helpman (Ed.) *General Purpose Technologies and Economic Growth.* MIT Press.
Salton, G., Wong, A. & Yang, C. S. (1975). A Vector Space Model for Automatic Indexing. *Communications of the ACM*, 18(11), 613-620.
Stephan, P., Gurmu, S., Sumell, A. & Black, G. (2007). Who's patenting in the university? Evidence from the survey of doctorate recipients. *Economics of Innovation and New Technology*, 16(2), 71-99.
Thursby, J. G. & Thursby, M. C. (2002). Who is selling the Ivory Tower? Sources for growth in university licensing. *Management Science*, 48(1), 90-104.
Trajtenberg, M. (1990). A Penny for your quotes: Patent citations and the value of innovations. *Rand Journal of Economics*, 21(1), 172-187.