Do national borders slow down knowledge diffusion within new technological fields? The case of big data in Europe*

Tatiana Kiseleva†, Ali Palali and Bas Straathof

CPB Netherlands Bureau for Economic Policy Analysis, The Hague

July 6, 2016

Abstract

Big data technologies enhance the storage, processing, and analysis of large data sets and can be applied economy-wide. Despite this potential, only one percent of big data patents come from Europe. This paper investigates the diffusion of big data technologies across national borders by using speed of big data patent citations. Using mixed proportional hazard models with fixed effects and censoring correction we compare big data patents to non-big data ICT patents that have been filed at the USPTO. We find that big data patents are cited slower compared to other ICT patents. This delay fades as big data technologies mature. National borders do not systematically affect the diffusion of big data technology, also for regions which host little big data innovation like Europe.

Keywords: Big data patents, technology diffusion, patent citation, fixed effects duration model.

JEL classification: F23, O33

*We are thankful to Pierre Koning for the comments on an earlier draft of the paper. We are also thankful to Matteo Ramina for assisting us through out the whole process.

†Corresponding author. Email: t.kiseleva@cpb.nl

1 Introduction

The progress and diffusion of information and communication technologies (ICT) caused the world's capacity for computations to double every fourteen to eighteen months - which is similar to Moore's law for chip performance (Hilbert and López (2011)). This has prompted the development of new methods and technologies for data management and analysis - 'big data' technologies. Big data technologies can be categorized as general purpose technologies (GPTs, Bresnahan and Trajtenberg (1995)) because they enhance the use of ICT in general. Big data technologies can also facilitate innovation in other technological fields directly by providing new ways of analyzing data. In either way, access to big data technologies can contribute to a country's overall innovation activity.

If a country's amount of research and development (R&D) determines its capacity to absorb foreign knowledge (Cohen and Levinthal (1990)), then knowledge will diffuse more slowly to countries where R&D activity is modest. This effect might be stronger for new technological fields – like big data – than for established technological fields as researchers need time to familiarize themselves with novel concepts, mechanisms, applications, etc. before they can contribute to a new field.

The distribution of innovative activity is more uneven for big data than for other technologies. The United Kingdom Intellectual Property Office (UK Intellectual Property Office (2014)) reported that inventors listed on big data patent applications worldwide are predominantly located in the United States (46%) and China (29%). The share of European inventors is about 6% – about half the European share for all patent applications.[1] This small share raises the question whether European researchers are lagging behind researchers in other countries in the field of big data.

This paper examines the role of national borders on the diffusion of knowledge within new technological fields. In particular, we study whether European inventors are slower in applying big data technologies than inventors from other countries and in comparison to other ICT-technologies. It is well known that new technologies diffuse slower than established ones (Hall and Khan (2003); Atkeson and Kehoe (2007)). Seminal work by Jaffe and Trajtenberg (1999) shows that patents are more likely to be cited by patents from the same country and that domestic patents are cited earlier than patents from other countries. A more recent study (Griffith et al. (2011)) shows that border effects seem to have decreased over time and are now almost absent.

To our knowledge, it has not been studied before whether the effect of national borders is stronger for new technologies than it is for established ones. We compare border effects for big data patent citations with those for citations of established ICT technologies. Using survival analysis techniques we show that big data technologies indeed diffuse slower than already established ICT technologies. We do not find evidence of differential border effects for citations of big data patents. This result also holds for Europe which hosts little big data innovation.

Several empirical studies have analyzed technology diffusion using proxies for the number of users such as number of firms that introduced a new technology (Mansfield (1961)), consumption per capita (Comin and Hobijn (2004)) and demand for skilled labor (Bresnahan et al. (1999)). These types of proxies measure *adoption of technology*, i.e. how widespread a technology is

[1] The WIPO reports that 13% of patent applications world wide originated from Europe in 2014 (WIPO (2015)) One explanation for the small share of European inventors is that 'programs for computers' are not patentable according to Article 52 of the European Patent Convention.

among end users. Other studies have focused on technology diffusion using patent citations as a proxy (Jaffe and Trajtenberg (1999); Thompson and Fox-Kean (2005); Thompson (2006); Griffith et al. (2011)), thus analyzing *knowledge diffusion*, i.e. application of a technology for further innovation. In this paper we investigate the diffusion of big data technologies across national borders by using the speed of big data patent citations as a proxy for technology diffusion.[2].

In our empirical analysis we compare the effect of national borders on the diffusion of big data technologies to the border effect of other ICT technologies, controlling for the effects of cross-firm and cross-technology citations. A single patent can have inventors from different countries and regions. We consider all inventor locations in our definition of cross-border citations: a citation is considered 'cross-border' if all locations of the inventor of the citing patent are different from all those of the cited patent. Empirical analysis is performed through mixed proportional hazard models with correlated fixed effects and censoring correction to account for endogeneity and sample selection. Following Griffith et al. (2014) we control for technological distance between patents and for joint ownership of the patents, i.e. whether the cited and citing patent belong to the same firm.

Due to the relative novelty of the term a unique definition of a 'big data' technology does not exist. We have used two sources to identify 'big data' patents. The first definition is provided by the UK Intellectual Property Office (UKIPO) in their report 'Eight great technologies: big data' from 2014. The second definition has been compiled by Thomson Reuters(TR) at our request. A definition of a 'big data' patent consists of a list of International Patent Classification (IPC) and Corporate Patent Classification (CPC) codes, and a list of keywords[3]. These lists are then used to make a search for 'big data' patents. Using two definitions of 'big data' technologies we create two sets of big data patents. The UKIPO query selects around six thousand patents. The TR definition selects around 44 thousand patents[4]. Among others both sets contain patents for parallel computing methods, data processing methods and equipment, digital computing methods and equipment. We use the TR set of patents for our core analysis and the UKIPO set for sensitivity analysis.

We find that citations of big data patents are slower compared to other ICT patents: the delay is nine percent for the whole sample and twelve percent for the early years of big data. This confirms the hypothesis that new technologies diffuse slower than already established technologies, and that the delay fades over time. Moreover, our analysis shows that the duration of citations of big data patents within national borders do not differ significantly from the duration of cross-border citation. From this result we conclude that big data technologies diffuse within and across borders in a similar way. Even Europe, which has few big data patents, does not seem to experience delays in knowledge diffusion caused by national borders. We also find that cross-technology and cross-firms citations are significantly slower which is consistent with the existent literature (Griffith et al. (2014); Jaffe and Trajtenberg (1999)). Finally, the results of various sensitivity analysis show that our findings are robust.

The paper is organized as follows. Section 2 briefly presents the history of big data tech-

[2]It is well known that patents do not capture all inventive activities as not all inventions get patented. For example, in the ICT sector about 47% of innovations got patented in Japan (Nagaoka et al. (2010))

[3]Full description of IPC and CPC codes and keywords can be found in Appendix A and B.

[4]Such a striking difference in the number of patents can be explained by the novelty of the term 'big data'. There is no standardised definition of 'big data technologies' yet. UKIPO and Thomson Reuters have compiled a list of IPC/CPC codes and keywords that in *their* opinion capture the term 'big data technologies' best. Manual editing of selected sets of patents makes the difference between the UKIPO and TR definitions even bigger.

nologies. Section 3 describes our modeling strategy. Section 4 provides a detailed description of data that we use. In Section 5 we describe our main findings. Finally, Section 6 concludes.

2 A brief history of big data technologies

The term 'big data' has first appeared in a NASA article (Cox and Ellsworth (1997)) which argued that enormous growth of data volume was becoming an issue for current information technologies. Though computational capacity has been increasing with 58% a year, the volume of information have shown higher rates of increase (Hilbert and López (2011); LEF (2011)). The shortage of storage and computational capacity compared to the amount of data that had to be processed was noticeable in many economic sectors (McAfee et al. (2012)).

In 2004 Google designed and built a new data processing infrastructure MapReduce, which provided reliable and scalable storage and allowed computations to be split among large numbers of servers and carried out in parallel (Dean and Ghemawat (2008)). In 2006 Hadoop was created on the basis of MapReduce. Hadoop is a 100% open source way to store and process big data (Olson (2010)). Figure 1 demonstrates fast rising interest to 'big data' and Hadoop among internet users around the globe. The figure suggests that 2007 is a start of a 'big data' revolution. And given a wide specter of big data applications - from business analytics to health care - it is a revolution of a yet another general purpose technology.

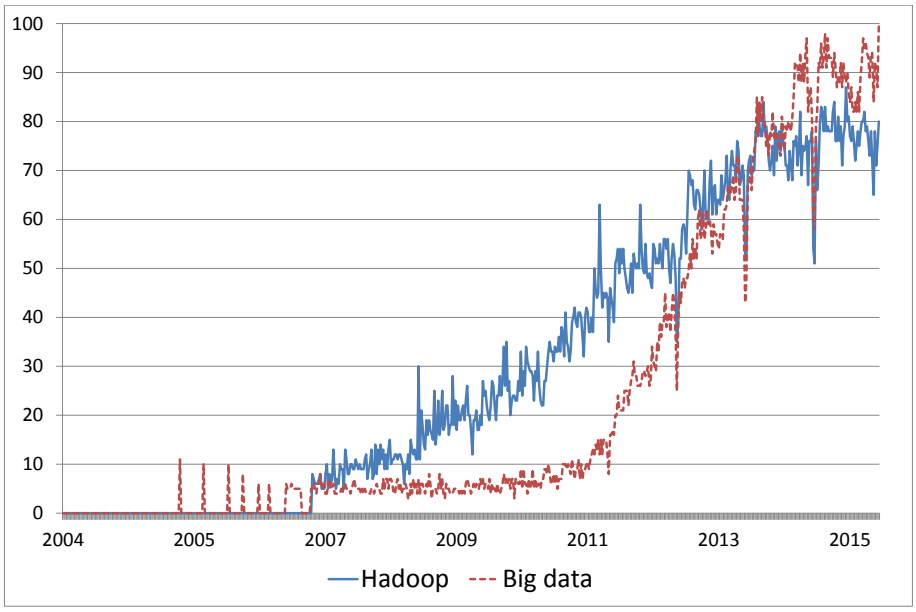

Figure 1: Trends of the search "Hadoop" and "Big data" in Google Search. The values are indexed with the highest number of searches =100 (achieved in December 2015 for the search term 'Big data'). Source: Google Trends.

These days the use of big data technologies generates significant financial value across economic sectors. It is estimated to generate 300 billion dollars in US health care and 250 billion euro per year in Europe public sector administration(Manyika et al. (2011)). The statistics on citation of big data patents shows that big data technologies are used in almost all economic

sectors. Table 1 shows that 'Machinery&equipment', 'Publishing&printing' and 'ICT&other services' are the most intensive economic sectors in terms of innovation using big data technologies.

Table 1: Diffusion of big data technologies in economic sectors illustrated by the number of patents, citing big data patents, filed to USPTO by companies in different economic sectors in 2002-2015. Source: Tomson Reuters.

Sector	Patents	Sector	Patents
Machinery&Equipment	63 442	Construction	580
Publishing&Printing	26 847	Insurance companies	556
ICT &Other services	23 531	Transport	297
Wholesale&Retail trade	2 328	Metals&Metal products	199
Post &Telecommunications	2 304	Food&Beverages&Tobacco	117
Banks	1 819	Gas&Water&Electricity	90
Education&Health	1 249	Wood&Cork&Paper	70
Chemicals	706	Textiles	63
Public administration&Defense	700	Hotels&Restaurants	28

3 Modelling strategy

We consider two sets of patents: *cited* patents $l = 1, \ldots, L$ and *citing* patents $k = 1, \ldots, K$. A patent from the set of cited patents can potentially be cited by one or more citing patents. If a patent $l \in [1, L]$ is cited by a patent $k \in [1, K]$ we compute the number of days t_{lk} between the application dates of patent l and patent k. The variable t_{lk} measures how fast the knowledge contained in patent l has been transfered to patent k. In other words, the variable t_{lk} is a proxy for speed of knowledge diffusion; it is also called *diffusion lag* in the literature.

There are many factors that influence the speed of patent citations. These factors include the unobserved characteristics of the cited patent V_l which are of crucial importance. For example, higher quality patents may be cited faster than lower quality patents. The observed characteristics of the pair cited-citing patent X_{lk} also influence the diffusion lag. For example, knowledge diffuses faster within one technological field than across the fields(Griffith et al. (2014)). Similarly, a firm cites its own patents faster than patents of others firms (Griffith et al. (2014)).

In this paper we are interested in cross-border effects on patent citations. We examine whether and to what extend do national borders slow down knowledge diffusion. It has been shown in (ref) that patents are cited faster by patents from the same country, and Griffith et al. (2014) shows that the delay decays over time. In this paper we focus on the diffusion lag for the new technological field - big data technologies. We hypothesize that knowledge diffuses slower within new fields compared to existent ones. This may happen, for example, due to low number of innovators working in the field which complicates information exchange between them. To test the hypothesis we compare big data technologies to the control group of other (non big data) ICT technologies[5]. We control for the effects of the technological fields by using dummies BD_{lk} which are equal 1 if both cited l and citing k patent are big data patents. Thus BD_{lk} is a dummy for with-in-field knowledge diffusion. To control for cross-border effect we use dummies CB_{lk} which are equal 1 if countries of all inventors of the cited patent l differ from countries of all inventors of the citing patent k. In this respect our paper is different from the rest of

[5]By using ICT technologies as a control group we eliminate the effects of institutional differences between different patent offices on the diffusion lag. See Section 4 for more details.

literature which mostly uses the country of the first inventor only. Using the countries of all inventors allows us to measure the cross border effect more precisely.

We consider a multiple spell version of the mixed proportional hazard model. The hazard rate of the patent l cited by the patent k on the t_{lk}^{th} day after application conditional on $V_l = v_l$, $X_{lk} = x_{lk}$, BD_{lk} and CB_{lk} is given by

$$\theta(t_{lk}|CB_{lk}, BD_{lk}, x_{lk}, v_l) = \lambda_l(t_{lk}|v_l) \exp(\alpha CB_{lk} + \delta BD_{lk} + \gamma CB_{lk} * BD_{lk} + x'_{lk}\beta), \quad (1)$$

where $\lambda_l(t_{lk}|v_l)$ is a cited-patent-specific hazard function. The function $\lambda_l(t_{lk}|v_l)$ is left unspecified and can vary across cited patents. Thus the model allows for unobserved heterogeneity in the hazard functions of cited patents.

The coefficient δ in (1) measures the diffusion lag within the field of big data technologies. A significant negative δ would confirm out hypothesis that new technologies (namely big data technologies) diffuse slower compared to existent ones (namely ICT). The coefficient α measures the effect of national borders on the speed of patent citations, and the coefficient γ measures the additional 'home-bias' effect for citations between big data patents. A significant negative γ would mean that new technologies travel even slower across national borders than established technologies.

v_l is for the unobserved patent characteristics of the cited patents. In our data set most of the patents are cited for multiple times by different patents. Using the information from multiple citations, we can allow for correlation between observed characteristics X_{lk} and unobserved characteristics v_l through fixed effects. This is crucial to investigate patent citations. One of the important unobserved patent characteristics is patent quality. Controlling for patent quality is of high importance as patent quality can be directly related to citation durations and can be systematically different across countries and across technologies due to differences in instutitions and legal conditions, etc. This means that if patent quality is uncontrolled for, then the results can be severely biased.

We allow for correlation between observed characteristics X_{lk}, which are constant within each spell but vary across spells, and unobserved characteristics V_l, on which we do not impose any assumption. Moreover, following Griffith et al. (2014) we impose the conditional independence assumption - the citation durations t_{lk_1} and t_{lk_2} are independent of each other conditional on X_{lk_1}, X_{lk_2} and V_l. This implies that one citation does not lead to another citation.

Under the conditional independence assumption we can estimate the coefficients $\alpha, \beta, \gamma, \delta$ using the conditional likelihood approach of Ridder and Tunali (1999). The intuition behind this approach is as follows. Assume for simplicity that there are only two potentially citing patents ($K = 2$). The conditional probability that the observed first citation of patent l is first is given by[6]

$$\begin{aligned} & Pr[T_{l1} \leq T_{l2} | T_{l1} = t_{t1}, Y_{l1} = y_{l1}, Y_{l2} = y_{l2}, V_l = v_l] \\ =& \frac{\lambda_l(t_{l1}|v_l) \exp(y'_{l1}\beta^*)}{\lambda_l(t_{l1}|v_l) \exp(y'_{l1}\beta^*) + \lambda_l(t_{l1}|v_l) \exp(y'_{l2}\beta^*)} \\ =& \frac{\exp(y'_{l1}\beta^*)}{\exp(y'_{l1}\beta^*) + \exp(y'_{l2}\beta^*)}. \end{aligned} \quad (2)$$

This implies that the probability does not depend on $\lambda_l(t_{lk}|v_l)$ or v_l as both are canceled out.

[6]For simplicity we introduce $y'_{lk}\beta^* = \alpha CB_{lk} + \delta BD_{lk} + \gamma CB_{lk} * BD_{lk} + x'_{lk}\beta$.

Therefore the coefficients $\alpha, \beta, \gamma, \delta$ can be estimated without specification of the base line hazard function $\lambda_l(t_{lk}|v_l)$ and at the same time taking the fixed effects v_l into account. Intuitively each patent contribute to the conditional likelihood by several times depending on the number of citations that this patent receives.

A usual problem with this types of models is censoring. Patents that have been cited one time only or have not been cited at all do not contribute to the analysis. This may cause two selection problems. The first one is that our data set is biased towards higher quality patents, as lower quality patents are likely to be cited less than two times. The second problem is that our data biased towards older patents. Young patents have less citations on average compared to older patents. To correct for the selection bias we use a modified version of the conditional likelihood estimator developed in Griffith et al. (2011). Specifically, all observations are weighted with an inverse censoring probability[7]. Assuming that censoring probability is independent of the durations of citations and observables, weighting corrects for the selection bias. Asymptotic properties of the fixed effects model with inverse censoring probability weights can be found in Griffith et al. (2011).

4 Data

We use data from three different sources: PATSTAT, Thomson Reuters and Orbis. PATSTAT is the Woldwide Patent Statistical Database of the European Patent Office, which contains bibliographic patent data such as application dates, IPC codes[8], inventor information, citations, etc. Thomson Reuters data base contains not only bibliographical information of patents but also data on technological classes of innovations. Orbis - a worldwide database collected by Bureau van Dijk - provides firm specific information, such as number of employees, number of patents, operating revenue etc, for over 200 milions firms around the globe. We use Orbis to obtain information about firms that apply for patents.

In our analysis we only use patents applied at the USA Patent Office (USPTO). The reason for that is patents filed to one patent office are easy to compare, but patents filed to different patent offices are difficult to compare due to differences in citation practices, novelty requirements, etc. Inventions that are patented at USPTO are protected in the USA only, but do not necessarily have a US inventor. A foreign firm may file its inventions to USPTO if it expects the invention to enter the US technology market or to be used for further innovation by US inventors. Thus our data set contains not only US firms and inventors, but also foreing ones that apply to USPTO. However, considering only USPTO patents may create a selection problem. US inventors are more likely to apply for a patent at USPTO than foreign ones, and thus our data set can be biased toward US inventors. To address this problem in the analysis of big data patents we introduce a control group of non-big-data ICT patents filed to USPTO. If US inventors are more likely to apply for big data patents at USPTO, then they are also more likely to apply for other type of ICT patents. Thus comparing big data patents with non-big-data ICT patents we estimate the difference in the diffusion lags between the two groups.

To select big data patents we have made an inquiry at Thomson Reuters. The field of big data technologies is relatively new, and the standardized definition of big data technologies does not yet exist. That is why we need to employ the expertise of Thomson Reuters to create the

[7] For a detailed derivation of the weighted conditional likelihood function see Griffith et al. (2011).

[8] International Patent Classification Codes are symbols used for classification of patents according to different areas of technology.

correct search inquiry for big data patents[9]. The inquiry results in 86961 big data patents, from which 44000 have been applied at USPTO. To check whether our results are sensitive to the definition of big data technologies, we perform sensitivity analysis with the selection of big data patents provided by UK Intellectual Property Office in their report from 2014.

To create the control group we use the list of classification codes of ICT patents from OECD (2010). We select all ICT patents from PATSTAT database which results in over 3 million patents. We then draw a random sample of 44000 patents among non-big-data ICT patents and merge it with the set of big data patents.[10] This results in a set of 88000 patents where half of them are big data patents and the other half are non-big-data ICT patents. As the next step we use PATSTAT to obtain information on the citations of the resulting group of 88000 patents. Finally, we merge the two sets of patents, cited and citing, with Orbis database to link patents to the firm specific information of the patent owners.

Table 2 gives the full list of variables used in the analysis. The dependent variable is citation duration (t_{lk}). We use the location of all inventors as a regressor. We divide countries of inventors into three groups: EU, USA and OTHERS. The group EU consists of all European countries, the group OTHERS contains all countries other than EU and USA. A dummy (EU_l to USA_k) is equal to 1 if all inventors of the cited patent l are from EU and all inventors of the citing patent k are from USA . We also include technological distance between cited and citing patents as a regressor. We compute it as follows. We consider two sets $A = \{a_1, a_2, \ldots, a_n\}$ and $B = \{b_1, b_2, \ldots, b_m\}$, where A are IPC codes of the cited patents and B are IPC codes of the citing patent. Then we compute the share of IPC codes that the patents have in common. It measures the technological similarity between the two patents. To obtain the technological distance between them we subtract this number from 1. The technological distance is thus given by

$$Tech.distance = 1 - \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} I_{a_i=b_j}}{n \times m}, \tag{3}$$

where $I_{a_i=b_j}$ is an indicator of the event $a_i = b_j$. The variable *Tech.distance* takes values from 0 (all IPC codes of both patents equal) to 1 (the patents have no IPC codes in common). We also use firm-level information about the owner of the patent[11], such as number of patents that a firm owns, its annual revenue and the number of employees. Additionally we add a dummy *within-firm* to the regression, it is equal to 1 if the cited and citing patents belong to one firm.

4.1 Descriptive statistics

Figure 2 presents the number of patents in both groups - big data patents and non-big-data ICT patents - per application year. There is no big data patents with the application date before 1997. For that reason we pick only patents applied between 1997 and 2014 when compiling the control group of non-big-data ICT patents. Both groups have a peak in 2007. After that year the number of big data patents gradually decreases. The reason could be the financial crisis that hit the world then. The considerable drop in the number of all patents after 2011 can be due to administrative delays in assigning applications number to patents. Due to the same reason we do not have very young patents in our data set, as they have not yet been assigned with application numbers.

[9] Detailed information on the search inquiry of Thompson Reuters can be found in the Appendix.

[10] Note that we also perform a propensity score matching to construct our control sample instead using random draws. This does not affect our conclusion.

[11] We use the data of the Global Ultimate Owner.

Table 2: Description of variables used in empirical analysis.

Dep. variable	Description
Citation duration	Number of days elapsed from the application date of the cited patent until the application date of the citing patent.
Dummies	
BD	1 if both cited and citing patents are big data patents
CB	1 if locations of all inventors of cited patent are different from those of citing patent
EU to USA	1 if cited patent from EU, citing patent from USA
EU to OTHERS	1 if cited patent from EU, citing patent from OTHERS
EU to USA/OTHERS	1 if cited patent from EU, citing patent from USA/OTHERS
USA to EU	1 if cited patent from USA, citing patent from EU
USA to OTHERS	1 if cited patent from USA, citing patent from OTHERS
USA to EU/OTHERS	1 if cited patent from USA, citing patent from EU/OTHERS
OTHERS to EU	1 if cited patent from OTHERS, citing patent from EU
OTHERS to USA	1 if cited patent from OTHERS, citing patent from USA
OTHERS to EU/USA	1 if cited patent from OTHERS, citing patent from EU/USA
EU/USA to OTHERS	1 if cited patent are from EU/USA, citing patent from OTHERS
USA/OTHERS to EU	1 if cited patent from USA/OTHERS, citing patent from EU
EU/OTHERS to USA	1 if cited patent from EU/OTHERS, citing patent from USA
Sector dummies	19 sectors in which the citing firm is operating
Within firm	1 if the firm-owner for cited and citing patents is the same
Regressors	
Tech. distance	Percentage of IPC codes common in cited and citing patents
Nr. patents	Number of patents applied by the citing firm in total
Revenue	Operating revenue of the firm which applied for the citing patent
Nr. employees	Number of employees working at citing firm

Table 3 gives the sample statistics for the variables used in the analysis. On average it takes 1435 days for a big data patent to be cited for the first time, which is 127 faster compared to other ICT patents. It can happen due to the fact that big data patents are on average of higher quality due to novelty of the field[12]. Or it can be due to the fact that big data patents are technologically more similar to each other ($Tech.distance = 0.778$) than other ICT patents ($Tech.distance = 0.894$), and within-field citations arrive faster than across-field citations (ref). Only 34% citations of big data patents come from other big data patents. The other 66% come from other fields. This is an indication that big data technologies are widely applied in other technological fields.

Figure 3 shows the percentage of patents by the number of citations they receive. Almost 8% of big data patents and 4% of non big data patents do not receive any citations. 22% of big data patents and 15% of non big data patents receive only 1 citation. The percentages gradually decrease as the number of citations increase. Finally, around 4% of both type of patents receive more than 30 citations in total.

Table 4 displays the percentage of cited and citing patents based on locations of inventors. Most of the cited and citing big data patents have all their inventors located in the USA, 72% and 65% correspondingly. European inventors are responsible for only 3% of big data cited patents and 5% of citing patents. The figures are similar for non-big-data ICT patents and consistent with the figure from the recent report of UKIPO on big data innovation (UK Intellectual Property Office (2014)). This indicates that there is little innovation activity in the ICT sector in Europe. However it does not necessarily mean that there is a lag in application

[12]In a new field most inventions are often fundamental inventions with higher impact. In an established field most inventions are incremental with a lower impact.

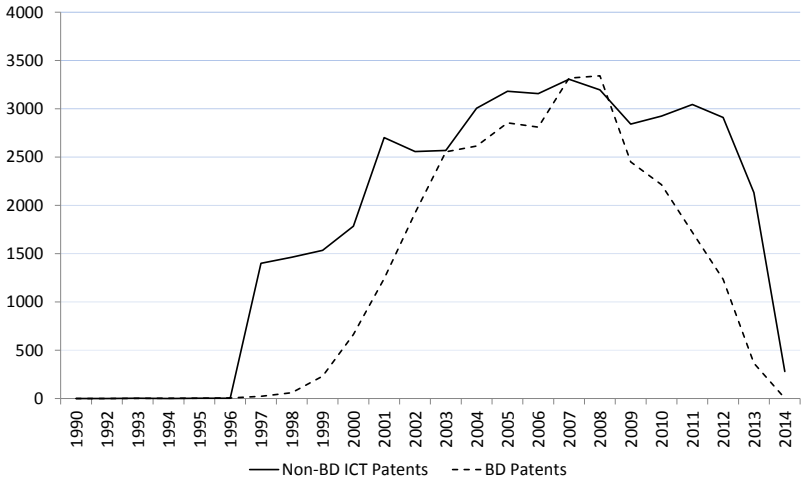

Figure 2: Number of big data and non-big data ICT patents per application year.

of big data and other ICT technologies by European inventors. To investigate this we analyze whether European inventors are slower in citing big data patents than inventors from other countries.

Finally, Figure 4 shows the hazard rates and cumulative probability of being cited for big data and non big data ICT patents and for domestic and cross border citations. For both groups of patents the cumulative probability is higher for domestic citations in comparison to cross border citations. This implies that patents are cited faster by inventors from the same country compared to inventors from other countries.

5 Results

Table 5 shows the parameter estimates for three different model specifications. For the estimations we use the first 10 citations of patents[13]. We find that technological distance between citing and cited patents is significantly negative in all three specifications. This implies that patents of technologies that are similar cite each other faster than patents of more distant technologies. Moreover, firms tend to cite their own patents faster than patents from other firms. Both results are consistent with the literature.

Let us first discuss the results obtained through a standard Cox model without fixed effects, which are reported in column (1). The parameter estimate for the cross border citation (CB) is negative and statistically significant. This means that the hazard rate, i.e. probability of being cited after a certain number of days after application, is lower for cross border citations. Therefore inventors from the same country as the country of the inventors cite a patent faster than foreign inventors. The parameter estimate for the interaction effect ($CB * BD$) suggests that big data patents receive cross border citations faster than non-big data ICT patents. Put

[13]In the sensitivity analysis we estimate the model with less and more citations in order to check the robustness of the results.

Table 3: Sample statistics of variables used in empirical analysis.

	Big data patents				Non-big data ICT patents			
Variables	Mean	St.Dev	Min	Max	Mean	St.Dev	Min	Max
1st citation (days)	1435.837	743.841	12	7804	1562.584	796.232	10	5739
2nd citation (days)	1802.883	770.256	37	8077	1959.480	834.750	79	5981
10th citation (days)	2724.329	801.341	501	5984	3002.459	921.369	594	6060
CrossBorder	0.205	0.403	0	1	0.290	0.454	0	1
EU to USA	0.018	0.131	0	1	0.028	0.166	0	1
EU to OTHERS	0.004	0.066	0	1	0.015	0.120	0	1
EU to USA/OTHERS	0.003	0.056	0	1	0.004	0.063	0	1
USA to EU	0.027	0.162	0	1	0.028	0.164	0	1
USA to OTHERS	0.075	0.263	0	1	0.084	0.277	0	1
USA to EU/OTHERS	0.006	0.076	0	1	0.005	0.070	0	1
OTHERS to EU	0.004	0.066	0	1	0.015	0.122	0	1
OTHERS to USA	0.050	0.217	0	1	0.091	0.288	0	1
OTHERS to EU/USA	0.004	0.065	0	1	0.007	0.085	0	1
EU/USA to OTHERS	0.006	0.075	0	1	0.006	0.079	0	1
USA/OTHERS to EU	0.004	0.064	0	1	0.003	0.053	0	1
EU/OTHERS to USA	0.004	0.063	0	1	0.004	0.063	0	1
BigData citation	0.340	0.474	0	1				
Tech. distance	0.778	0.317	0	1	0.894	0.209	0	1
Within firms	0.328	0.470	0	1	0.315	0.465	0	1
Number of patents	52468.6	79749.96	0	418131	61460.24	100096.5	0	418131
Revenue(in 100000)	440	401.000	0	0.042	400	457	-0.02	4210
Number of employees	128980.8	120062.9	0	488824	103589.3	104300.9	0	488824

Table 4: Share of patents according to the location of inventors.

Locations of the inventors	Cited patents		Citing patents	
	Non-BD ICT	BD	Non-BD ICT	BD
USA	0.55	0.72	0.54	0.65
EU	0.06	0.03	0.06	0.05
OTHER	0.28	0.10	0.26	0.13
USA & EU	0.03	0.05	0.04	0.05
USA & OTHERS	0.06	0.09	0.08	0.11
EU & OTHERS	0.01	0.01	0.01	0.01
EU & USA & OTHERS	0.01	0.01	0.01	0.01

differently, big data technologies travel faster across national borders than other ICT technologies. It can possibly be explained by the novelty of the field of big data technologies, where many innovation are fundamental and thus get cited faster than incremental innovations.

Table 5: Parameter estimates of the baseline specifications.

	(1) Cox		(2) Fixed effects		(3) Fixed effect Cens.	
CB	-0.092***	(0.005)	0.006	(0.007)	0.007	(0.007)
BD	0.004	(0.006)	-0.092***	(0.009)	-0.094***	(0.010)
CB*BD	0.048***	(0.014)	-0.004	(0.018)	-0.008	(0.019)
Tech. distance	-0.312***	(0.008)	-0.309***	(0.012)	-0.314***	(0.013)
Within firm	0.183***	(0.004)	0.226***	(0.006)	0.236***	(0.006)
N	309271		309271		309271	

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

The Cox model does not account of unobserved heterogeneity. Which implies that unobserved patent quality might bias the results. High quality patents are more likely to receive cross

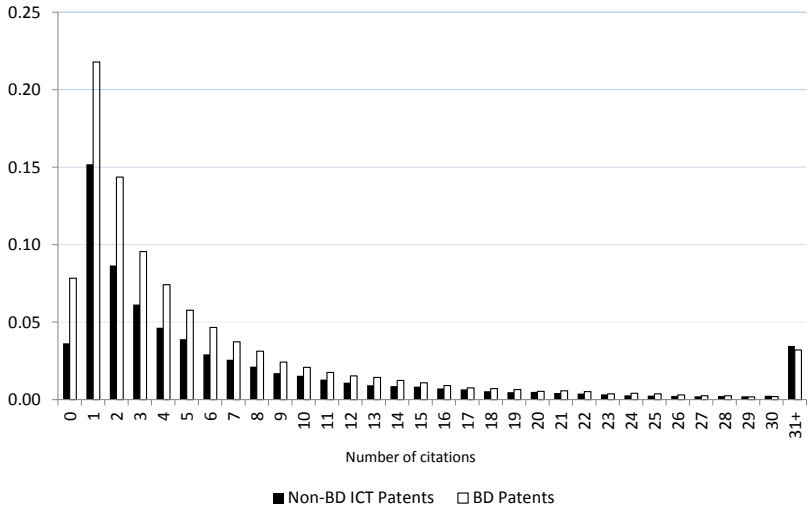

Figure 3: Number of big data and non-big data ICT patents per number of citations.

border citations. Moreover quality might be correlated with the location of patents. In column (2) we control for the unobserved patent quality through fixed effects. The significant negative estimate of BD indicates that BD to BD citations happen slower compared to BD to non-BD ICT citations conditional on patent quality and cross border effects.[14] In other words, non-BD ICT patents cite BD patents faster than BD patents cite BD patents. Which might signal a big number of fundamental BD inventions, that are used widely in other ICT sectors. Moreover, the cross border effect (CB) and the interaction effect ($CB * BD$) disappear. These results show that unobserved patent quality is indeed an important factor influencing citation durations.

Finally, column (3) in Table 5 reports the results when sample selection due to censoring is taken into account. The results hardly change. The coefficient estimate for BD to BD citations remains as -0.09 and significant, indicating that big data to big data citations happen approximately 9% slower compared to BD to non-BD ICT citations.

We now further explore the cross border effects in more detail by dividing the cross border dummy variable into 12 different categories for the three regions (USA, EU and OTHERS). Table 6 presents the parameter estimates of these variables together with their interactions with the (BD) variable. Even though our preferred model is the model with fixed effects and censoring, we present the results for the Cox model and the fixed effects model as well for completeness. As the title of the paper suggests we focus on the cross border effects for the Europe. European inventors seems to lag behind the local inventors in citing patents from USA and OTHERS. EU is 5% slower to cite USA patents, and 9% slower to cite OTHERS patents. Whereas EU is 13% faster than local inventors in citing patents from USA/OTHERS.

Let us look at the interaction effects of big data technologies with country dummies. None

[14]Note that the interpretation of the estimated coefficient of BD variable is different in columns (1) and (2). In column (1) the reference group is all citations excluding BD to BD citations (i.e. a citation of a big data patent by a big data patents). In column (2) the reference group is BD to non-BD ICT citations. When we separately control for non-bBD ICT to BD citations and non-BD ICT to non-BD ICT citations in column (1), the coefficient estimate of the dummy BD becomes similar to those in column (2) and (3).

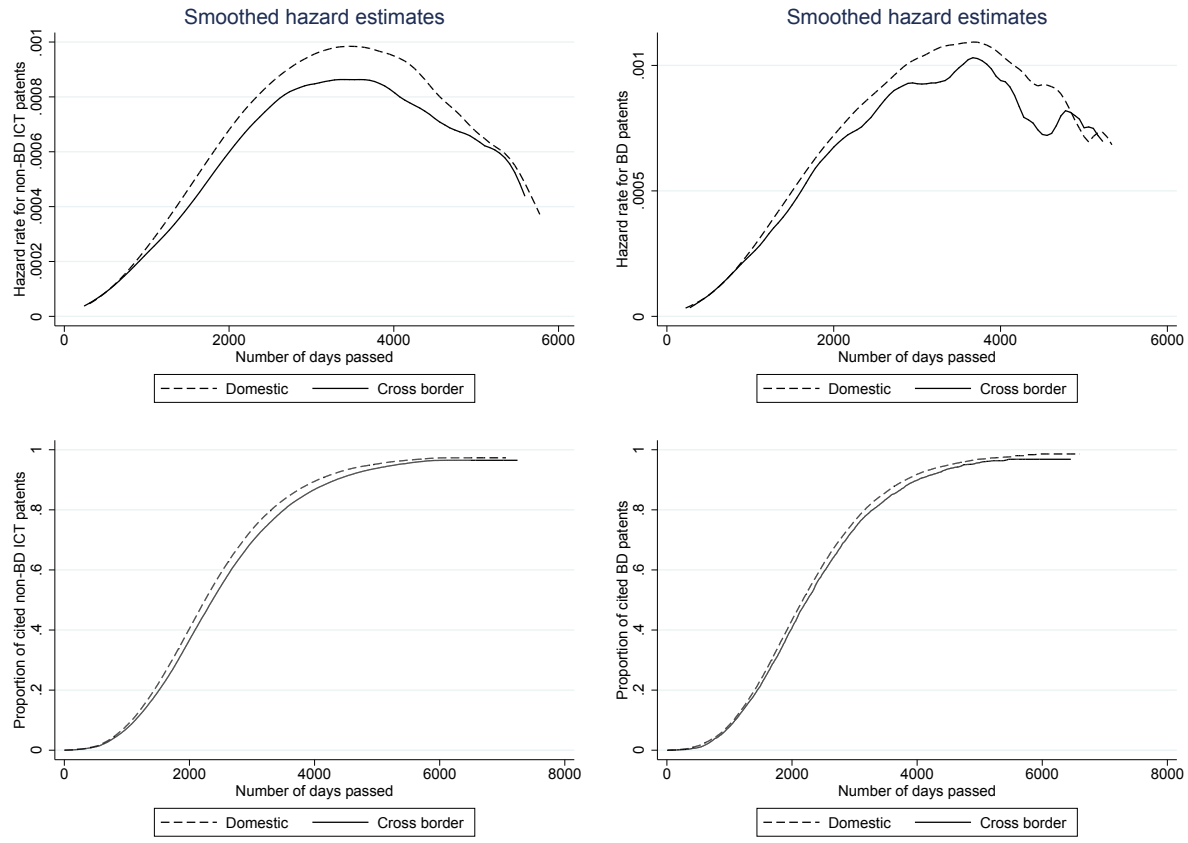

Figure 4: Cumulative probability of being cited for non big data ICT patents and big data patents for domestic and cross border citations.

of the interaction effects is significant at a 5% significance level for our preferred model (column (3)). Moreover excluding the coefficients of From US to EU and From US to OTHERS, the rest of the interaction effects are jointly insignificant.[15] Therefore it is hard to draw clear conclusions related to citations of big data patents by big data patents from the three regions. The results suggest that big data patents are not different than other ICT patents when it comes to cross border citations. And even though we find that big data patents are cited slower as a new technology, we show that this effect is the same within borders and across borders.

5.1 Sensitivity analysis

In this subsection we run a few sensitivity analysis estimations to check whether our results are sensitive to the model specifications. First, we explore whether firm characteristics of the applicant affect the results. Then we focus on the number of citations and the application period of the patents. Finally, we check whether the results would hold if we use another definition of big data technologies.

Table 7 presents the first set of sensitivity analysis using firm characteristics of the citing patents. In all four columns our preferred specification with 12 cross border variables is replicated by adding firm specific variables one by one. In column (1) we control for the total number of patents applied by the firm. This variable serves as a proxy for innovativeness of the firm.

[15]Including these two coefficients, all of the interactions effects are jointly significant at only 10-percent significance level

Table 6: Parameter estimates for the disentangled cross border effects.

	(1) Cox		(2) Fixed effects		(3) Fixed effect Cens.	
EU to USA	-0.171***	(0.013)	-0.018	(0.029)	-0.028	(0.031)
EU to OTHERS	-0.178***	(0.020)	-0.026	(0.035)	-0.049	(0.037)
EU to USA/OTHERS	-0.119***	(0.034)	-0.076	(0.051)	-0.090	(0.054)
USA to EU	-0.184***	(0.012)	-0.046**	(0.016)	-0.053**	(0.017)
USA to OTHERS	-0.015	(0.008)	0.033***	(0.010)	0.039***	(0.011)
USA to EU/OTHERS	-0.208***	(0.027)	-0.183***	(0.035)	-0.199***	(0.037)
OTHERS to EU	-0.251***	(0.020)	-0.068**	(0.026)	-0.092**	(0.028)
OTHERS to USA	-0.081***	(0.008)	0.029*	(0.014)	0.022	(0.015)
OTHERS to EU/USA	-0.109***	(0.027)	-0.063	(0.036)	-0.076*	(0.038)
EU/USA to OTHERS	0.073**	(0.028)	0.112**	(0.037)	0.146***	(0.040)
OTHERS/USA to EU	0.004	(0.036)	0.106*	(0.043)	0.130**	(0.046)
EU/OTHERS to USA	0.004	(0.032)	-0.075	(0.059)	-0.023	(0.063)
BD	0.004	(0.006)	-0.094***	(0.009)	-0.095***	(0.010)
Interaction Effects:						
EU to USA	0.153***	(0.042)	0.226***	(0.062)	0.176**	(0.068)
EU to OTHERS	0.033	(0.098)	0.128	(0.124)	0.163	(0.132)
EU to USA/OTHERS	0.062	(0.095)	-0.060	(0.122)	-0.113	(0.133)
USA to EU	0.170***	(0.040)	0.036	(0.049)	0.028	(0.053)
USA to OTHERS	-0.047*	(0.023)	-0.071*	(0.028)	-0.067*	(0.030)
USA to EU/OTHERS	0.141*	(0.067)	0.099	(0.087)	0.070	(0.094)
OTHERS to EU	0.199*	(0.094)	0.296*	(0.116)	0.324**	(0.124)
OTHERS to USA	0.057*	(0.024)	0.026	(0.034)	0.030	(0.036)
OTHERS to EU/USA	-0.226***	(0.068)	-0.265*	(0.105)	-0.226*	(0.112)
EU/USA to OTHERS	-0.105	(0.081)	-0.195*	(0.097)	-0.240*	(0.106)
OTHERS/USA to EU	-0.020	(0.104)	-0.194	(0.144)	-0.230	(0.156)
EU/OTHERS to USA	0.387***	(0.076)	0.109	(0.116)	0.101	(0.117)
Tech. distance	-0.311***	(0.008)	-0.310***	(0.012)	-0.314***	(0.013)
Within firm	0.185***	(0.004)	0.226***	(0.006)	0.236***	(0.006)
N	309271		309271		309271	

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

As expected the coefficient estimate is positive and statistically significant. More innovative firms, i.e. firms with higher number of patents, cite faster than those with lower number of patents. Column (2) additionally controls for operating revenue of the firm. There is no evidence for revenue effects. Column (3) adds the total number of employees working at the firm. These three factors - number of employees together with total number of patents and operating revenue - capture the effects of firm size and efficiency. Operating revenue has a positive and significant effect on hazard rates once we control for the total number of employees. However, total number of employees has a negative effect on hazard rates. Therefore, firms with higher number of employees cite slower than those with lower number of employees, conditional on total number of patents and operating revenue. The reason can be that conditional on the number of patents and operating revenue, firms with higher number of employees are actually less efficient. Finally, column 4 adds dummy variables for the sectors in which citing firms are operating. This strengthens the effects of operating revenue and the total number of employees. In all 4 columns, the effect of big data to big data citations and interaction effects remain robust.

Table 8 presents a second set of sensitivity analysis using different sample designs. We perform this analysis in order to check if our results are sensitive to the design of our sample. In the main analysis we use first 10 citations of the patents. In column (1) and (2) we explore whether our results would hold if we use the first 5 citations or the first 15 citations correspondingly. In

Table 7: Sensitivity to adding firm specific information.

	(1)		(2)		(3)		(4)	
EU to USA	0.003	(0.049)	-0.012	(0.051)	-0.007	(0.054)	-0.002	(0.054)
EU to OTHERS	0.059	(0.055)	0.014	(0.058)	-0.034	(0.067)	-0.014	(0.068)
EU to USA/OTHERS	-0.069	(0.077)	-0.095	(0.080)	-0.141	(0.086)	-0.140	(0.086)
USA to EU	0.015	(0.027)	0.024	(0.029)	0.019	(0.030)	0.012	(0.030)
USA to OTHERS	0.043*	(0.017)	0.062***	(0.017)	0.059**	(0.022)	0.084***	(0.023)
USA to EU/OTHERS	-0.105	(0.055)	-0.070	(0.058)	-0.038	(0.060)	-0.039	(0.061)
OTHERS to EU	-0.141**	(0.043)	-0.147**	(0.046)	-0.095	(0.050)	-0.126*	(0.051)
OTHERS to USA	0.063**	(0.022)	0.046*	(0.023)	0.103***	(0.027)	0.081**	(0.027)
OTHERS to EU/USA	-0.056	(0.056)	-0.014	(0.058)	0.005	(0.062)	-0.022	(0.062)
EU/USA to OTHERS	0.087	(0.058)	0.097	(0.060)	0.198**	(0.075)	0.230**	(0.075)
OTHERS/USA to EU	0.150*	(0.071)	0.208**	(0.076)	0.227**	(0.079)	0.222**	(0.079)
EU/OTHERS to USA	-0.004	(0.094)	-0.010	(0.097)	0.053	(0.105)	0.084	(0.106)
BD	-0.071***	(0.014)	-0.067***	(0.014)	-0.055***	(0.015)	-0.054***	(0.015)
Interaction Effects:								
EU to USA	0.190*	(0.095)	0.155	(0.097)	0.101	(0.101)	0.099	(0.101)
EU to OTHERS	0.178	(0.172)	0.210	(0.175)	0.257	(0.191)	0.280	(0.191)
EU to USA/OTHERS	-0.120	(0.168)	-0.191	(0.175)	-0.137	(0.178)	-0.145	(0.178)
USA to EU	0.122	(0.080)	0.146	(0.082)	0.227**	(0.085)	0.224**	(0.085)
USA to OTHERS	-0.148***	(0.042)	-0.166***	(0.043)	-0.193***	(0.053)	-0.187***	(0.054)
USA to EU/OTHERS	-0.108	(0.130)	-0.141	(0.130)	-0.307*	(0.143)	-0.307*	(0.144)
OTHERS to EU	0.308	(0.186)	0.323	(0.191)	0.400	(0.209)	0.379	(0.209)
OTHERS to USA	0.019	(0.051)	0.015	(0.052)	-0.066	(0.056)	-0.069	(0.056)
OTHERS to EU/USA	-0.285	(0.149)	-0.243	(0.151)	-0.265	(0.154)	-0.270	(0.154)
EU/USA to OTHERS	-0.138	(0.139)	-0.180	(0.145)	-0.303	(0.185)	-0.312	(0.186)
OTHERS/USA to EU	0.104	(0.244)	0.056	(0.243)	0.023	(0.249)	0.035	(0.249)
EU/OTHERS to USA	0.033	(0.174)	-0.009	(0.175)	-0.029	(0.177)	-0.036	(0.179)
Tech. distance	-0.245***	(0.018)	-0.240***	(0.018)	-0.226***	(0.019)	-0.230***	(0.019)
Within firm	0.255***	(0.015)	0.259***	(0.015)	0.256***	(0.017)	0.252***	(0.017)
Firm Characteristics								
Nr. patents	0.031***	(0.005)	0.027***	(0.006)	0.039***	(0.008)	0.025**	(0.008)
Revenue			-0.009	(0.011)	0.051*	(0.021)	0.083***	(0.023)
Nr. employees					-0.031***	(0.007)	-0.046***	(0.009)
Sector dummies							Yes	
N	176389		169166		151539		151358	

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Number of patents and number of employees are in 100 thousand, revenue is in 100 million.

both specifications the results are similar to the baseline results.

In columns (3) and (4) we explore whether our results are sensitive to the age of patents in the sample. Griffith et al. (2011) show that cross border effect seems to be decreasing over time. Therefore our results might change if we estimate the model for older patents. In column (3) we restrict the sample to patents applied before 01 January 2008, in column (4) – before 01 January 2006. In both specifications our results remain robust. The coefficients for the country dummy are consistent with the findings of Griffith et al. (2014), the cross border effect is decreasing over time. Moreover, knowledge transfer within the field of big data seems to speed up with time.

Furthermore, in Table 9 we present the results of estimations with an alternative definition for big data patents. In these estimations we used the list of big data patents obtained from the UKIPO instead of Thomson Reuters.[16] Since the UKIPO list is more restrictive in identifying big data patents, sample size decreases considerably. However, as the results in both columns show our main findings remain the same.

[16]Details on search inquiry used by the UKIPO to identify big data patents are given in Appendix A.

Table 8: Sensitivity to the changes in the sample design.

	(1) First 5		(2) First 15		(3) Before 2008		(4) Before 2006	
EU to USA	-0.116**	(0.040)	-0.085**	(0.029)	-0.001	(0.032)	0.036	(0.035)
EU to OTHERS	-0.101*	(0.048)	-0.102**	(0.034)	-0.095*	(0.039)	-0.096*	(0.042)
EU to USA/OTHERS	-0.128	(0.070)	-0.211***	(0.048)	-0.021	(0.055)	0.008	(0.062)
USA to EU	-0.077***	(0.023)	-0.030*	(0.015)	-0.098***	(0.018)	-0.149***	(0.019)
USA to OTHERS	0.050***	(0.014)	0.060***	(0.010)	0.007	(0.011)	-0.031*	(0.013)
USA to EU/OTHERS	-0.079	(0.052)	-0.187***	(0.033)	-0.183***	(0.039)	-0.318***	(0.045)
OTHERS to EU	-0.097**	(0.036)	-0.079**	(0.026)	-0.105***	(0.029)	-0.105***	(0.032)
OTHERS to USA	0.006	(0.019)	0.022	(0.013)	0.046**	(0.015)	0.075***	(0.017)
OTHERS to EU/USA	0.000	(0.050)	0.001	(0.034)	-0.065	(0.040)	-0.023	(0.043)
EU/USA to OTHERS	0.200***	(0.052)	0.150***	(0.035)	0.059	(0.043)	0.094*	(0.047)
USA/OTHERS to EU	0.084	(0.057)	0.092*	(0.041)	0.102*	(0.049)	0.092	(0.056)
EU/OTHERS to USA	-0.062	(0.082)	-0.047	(0.055)	-0.012	(0.065)	-0.089	(0.076)
BD	-0.089***	(0.013)	-0.095***	(0.009)	-0.124***	(0.011)	-0.142***	(0.012)
Interaction Effects:								
EU to USA	0.314***	(0.081)	0.110	(0.060)	0.248***	(0.071)	0.273***	(0.080)
EU to OTHERS	0.140	(0.164)	0.086	(0.126)	0.225	(0.142)	0.039	(0.182)
EU to USA/OTHERS	-0.074	(0.162)	-0.014	(0.121)	-0.028	(0.140)	-0.095	(0.163)
USA to EU	0.071	(0.069)	0.011	(0.048)	0.031	(0.057)	0.105	(0.063)
USA to OTHERS	-0.033	(0.038)	-0.083**	(0.027)	-0.079*	(0.032)	-0.085*	(0.037)
USA to EU/OTHERS	-0.275	(0.144)	-0.004	(0.083)	0.031	(0.097)	0.095	(0.111)
OTHERS to EU	0.539***	(0.149)	0.548***	(0.106)	0.246	(0.141)	0.194	(0.156)
OTHERS to USA	0.007	(0.047)	-0.017	(0.033)	0.027	(0.038)	0.029	(0.044)
OTHERS to EU/USA	-0.260	(0.148)	-0.286**	(0.098)	-0.244*	(0.117)	-0.413**	(0.135)
EU/USA to OTHERS	-0.352**	(0.128)	-0.071	(0.091)	-0.106	(0.114)	-0.349*	(0.145)
USA/OTHERS to EU	-0.031	(0.193)	-0.167	(0.146)	-0.060	(0.157)	0.135	(0.166)
EU/OTHERS to USA	0.237	(0.153)	0.148	(0.102)	0.116	(0.125)	0.172	(0.159)
Tech. distance	-0.272***	(0.016)	-0.321***	(0.012)	-0.345***	(0.014)	-0.412***	(0.016)
Within firm	0.224***	(0.008)	0.243***	(0.006)	0.245***	(0.007)	0.243***	(0.007)
N	215565		361526		255883		202552	

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Finally, we performed several other sensitivity analysis to investigate the robustness of our estimates. In one of such analysis we investigated the sensitivity of our results to the choice of control sample. In the previous analysis the control sample is constructed by choosing a random sample of all ICT patents. In the sensitivity analysis we choose a similar sized control sample among all ICT patents by using propensity score matching. We preform the matching on location of inventors, application date and total number of citations that a patent receives. In doing so we contruct a control sample that is similar to BD patents in terms of aforementioned observable characteristics. The results are reported in Tables 10 and 11 in Appendix C. Even though there are small differences between a few estimates, our conclusion remains the same. In another analysis, we investigated the sensitivity of our results to the firm sizes. In our data when we look at the patents with available firm information, around 85-percent of the patents are produced by top 25-percent of the firms in terms of operating revenue or total number of employees. In order to check the robustness of our results to this heterogeneity, we performed the fixed effects models with censoring by dividing our sample into 2 groups depending on firm sizes measured by the operating revenue and then by the number of employees: lowest 75% of the firms and top 25% of the firms. The parameter estimates are reported in Tables 12 and 13 in Appendix C.

Table 9: Sensitivity to the changes in the definition of big data patents

	(1)		(2)	
	Fixed effect	Cens.	Fixed effect	Cens.
CB	-0.053	(0.064)		
BD	-0.085**	(0.028)	-0.085**	(0.028)
CB*BD	0.120	(0.067)		
EU to USA			-0.412	(0.232)
EU to OTHERS			0.544	(0.416)
EU to USA/OTHERS			-0.986**	(0.303)
USA to EU			-0.261	(0.154)
USA to OTHERS			0.173	(0.107)
USA to EU/OTHERS			-0.086	(0.190)
OTHERS to EU			0.332	(0.325)
OTHERS to USA			-0.167	(0.139)
OTHERS to EU/USA			1.337*	(0.570)
EU/USA to OTHERS			0.346	(0.349)
OTHERS/USA to EU			-0.667*	(0.332)
EU/OTHERS to USA			0.031	(0.347)
Interactions with BD:				
EU to USA			0.260	(0.222)
EU to OTHERS			-0.509	(0.424)
EU to USA/OTHERS			0.702*	(0.339)
USA to EU			0.344*	(0.162)
USA to OTHERS			-0.036	(0.111)
USA to EU/OTHERS			-0.031	(0.223)
OTHERS to EU			-0.348	(0.348)
OTHERS to USA			0.120	(0.137)
OTHERS to EU/USA			-1.518**	(0.587)
EU/USA to OTHERS			-0.054	(0.381)
OTHERS/USA to EU			0.727*	(0.344)
EU/OTHERS to USA			-0.128	(0.356)
Tech. distance	-0.116**	(0.037)	-0.115**	(0.037)
Within firm	0.239***	(0.020)	0.239***	(0.020)
N	41525		41525	

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

6 Conclusions

This paper investigates the role of national borders on diffusion of new technologies. In particular we focus on whether European inventors are slower in applying big data technologies than inventors from other countries. We measure technological diffusion using patent citations. We compare the speed of citations of big data patents to that of other ICT technologies, controlling for cross-border, cross-firm and cross-technology effects. In the empirical analysis we use fixed effects duration models with inverse censoring probability weights. This enables us to control for observable as well as unobservable factors that can affect patent citation as well as sample selection problems due to censoring.

We find that citations of big data patents are slower compared to other ICT patents. This confirms that big data technologies, as a new technology, indeed diffuse slower than already established ICT technologies. However there is no evidence for systematic differences between domestic and cross-border diffusion, also for regions with little innovative activity like Europe. The sensitivity analysis show that this conclusion is robust.

Our results suggests that Europe is not lagging behind other countries in absorption and use of knowledge on big data technologies, which we argue to be general purpose technologies.

European inventors apply new big data technologies for further innovation as fast as inventors from other regions do. This means that Europe keeps up with the technological advancement of general purpose technologies, which is essential for its future economic growth.

A UKIPO query of big data patents

To identify big data patents the following keywords in combination with each other and with the IPC and CPC codes have been used.

Keywords: big data, Hadoop, Yarn, Aster, Datameer, FICO Blaze, Vertica, Platfora, Splunk, MapReduce, open data, data warehous*, informatic*, data mine?, data mining, simulate*, model*, analy*, artificial intelligence, neural network*, distributed*, (cluster*, cloud*, grid?) [within 3 words of] (based, comput*, server?, process*, software, application), croudsourc*, crowd sourc*, massively parallel process*, massively parallel software, massively parallel database?, distributed process*, distributed server?, distributed quer*, distributed database?, massive data

CPC codes:	**Description**
G06F 17/30*#	Digital computing or data processing equipment or methods, specially adapted for specific functions > Information retrieval; Database structure there for
G06F 19/70*#	Digital computing or data processing equipment or methods, specially adapted for specific functions > Chemo-informatics, i.e. data processing methods or systems for the retrieval, analysis, visualisation or storage of physiochemical or structural data of chemical compounds
G06F 19/30*#	Digital computing or data processing equipment or methods, specially adapted for specific functions > Medical informatics, i.e. computer-based analysis or dissemination of patient or disease data
G06F 19/10*#	Digital computing or data processing equipment or methods, specially adapted for specific functions > Bioinformatics, i.e. methods or systems for genetic or protein-related data processing in molecular biology
G06Q 10/063*#	Resources, workflows, human or project management, e.g. organising, planning, scheduling or allocating time, human or machine resources; Enterprise planning; Organisational models > Operations research or analysis
G06Q 30/02*#	Commerce, e.g. shopping or e-commerce > Marketing, e.g. market research and analysis, surveying, promotions, advertising, buyer profiling, customer management or rewards; Price estimation or determination
G06F 17/50*#	Computer aided design
G06N*#	Computer systems based on specific computational models

* Subsidiary subgroups also included, # Combined with selected keywords

IPC codes:	**Description**
G06F 17/30*#	Digital computing or data processing equipment or methods, specially adapted for specific functions > Information retrieval; Database structure there for
G06F 19/10*#	Digital computing or data processing equipment or methods, specially adapted for specific functions > Bioinformatics, i.e. methods or systems for genetic or protein-related data processing in molecular biology
G06Q 30/02*#	Commerce, e.g. shopping or e-commerce > Marketing, e.g. market research and analysis, surveying, promotions, advertising, buyer profiling, customer management or rewards; Price estimation or determination
G06F 17/50*#	Computer aided design
G06N*#	Computer systems based on specific computational models

* Subsidiary subgroups also included, # Combined with selected keywords

B Thomson Reuters query of big data patents

Big data patents have been identified by the team of experts from Thomson Reuters by searching combinations of the following keywords, classification codes and manually editing the data.
Keywords: big data, large OR massive OR huge OR enormous data set, DataStax, Marklogic, Accumulo, etc.

IPC/CPC codes:	**Description**
G06F	Electric digital data processing
G06N	Computer systems based on specific computational models [7]
G06Q	Data processing systems or methods, specially adapted for administrative, commercial, financial, managerial, supervisory or forecasting purposes; systems or methods specially adapted for administrative, commercial, financial, managerial, supervisory or forecasting purposes, not otherwise provided for [2006.01]

DWPI codes:	**Description**
T01-E	Data processing
T01-D	Data conversion
T01-J	Data processing systems

US classification:	**Description**
700000	Data processing: generic control systems or specific applications (7 child classes)
701000	Data processing: vehicles, navigation, and relative location (3 child classes)
702000	Data processing: measuring, calibrating, or testing (6 child classes)
703000	Data processing: structural design, modeling, simulation, and emulation (11 child classes)
704000	Data processing: speech signal processing, linguistics, language translation, and audio compression/decompression (11 child classes)
705000	Data processing: financial, business practice, management, or cost/price determination (7 child classes)
706000	Data processing: artificial intelligence (12 child classes)
707000	Data processing: database and file management or data structures (5 child classes)
709000	Electrical computers and digital processing systems: multicomputer data transferring (22 child classes)
715000	Data processing: presentation processing of document, operator interface processing, and screen saver display processing (15 child classes)
716000	Data processing: design and analysis of circuit or semiconductor mask (3 child classes)
717000	Data processing: software development, installation, and management

C Sensitivity analysis

Table 10: Sensitivity to the control sample: base line model.

	(1) Cox		(2) Fixed effects		(3) Fixed effects Cens.	
CB	-0.039***	(0.004)	-0.025***	(0.006)	-0.018*	(0.007)
BD	0.008	(0.006)	-0.110***	(0.009)	-0.111***	(0.011)
CB*BD	0.025**	(0.013)	-0.025	(0.018)	-0.010	(0.022)
Tech. distance	-0.421***	(0.008)	-0.373***	(0.011)	-0.387***	(0.014)
Within firm	0.130***	(0.004)	0.271***	(0.006)	0.269***	(0.007)
N	403962		403962		403962	

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 11: Sensitivity to the control sample: the model with disentangled border effects.

	(1) Cox		(2) Fixed effects		(3) Fixed effects Cens.	
EU to USA	-0.133***	(0.016)	-0.135***	(0.029)	-0.165***	(0.035)
EU to OTHERS	-0.223***	(0.022)	-0.170***	(0.035)	-0.205***	(0.042)
EU to USA/OTHERS	-0.135***	(0.039)	-0.231***	(0.055)	-0.163*	(0.065)
USA to EU	0.027	(0.018)	0.048*	(0.024)	-0.075*	(0.030)
USA to OTHERS	0.172***	(0.012)	0.069***	(0.016)	0.096***	(0.019)
USA to EU/OTHERS	-0.016	(0.042)	-0.071	(0.051)	-0.081	(0.062)
OTHERS to EU	-0.166***	(0.012)	-0.145***	(0.015)	-0.171***	(0.018)
OTHERS to USA	-0.038***	(0.006)	-0.019*	(0.008)	-0.019*	(0.010)
OTHERS to EU/USA	-0.081***	(0.017)	-0.082***	(0.022)	-0.079**	(0.026)
EU/USA to OTHERS	0.237***	(0.040)	0.238***	(0.051)	0.284***	(0.064)
OTHERS to USA/EU	0.012	(0.037)	0.123**	(0.046)	0.195***	(0.055)
EU/OTHERS to USA	0.158***	(0.040)	0.005	(0.065)	0.097	(0.077)
BD	0.016**	(0.006)	-0.098***	(0.009)	-0.095***	(0.011)
Interaction effects:						
EU to USA	0.071	(0.043)	0.229***	(0.062)	0.166*	(0.078)
EU to OTHERS	0.063	(0.097)	0.182	(0.124)	0.234	(0.149)
EU to USA/OTHER	0.061	(0.096)	-0.020	(0.123)	-0.227	(0.153)
USA to EU	-0.065	(0.042)	-0.054	(0.052)	-0.118	(0.065)
USA to OTHER	-0.241***	(0.025)	-0.091**	(0.030)	-0.065*	(0.036)
USA to EU/OTHERS	-0.054	(0.074)	-0.019	(0.094)	-0.098	(0.118)
OTHERS to EU	0.064	(0.099)	0.390***	(0.114)	0.393**	(0.137)
OTHERS to USA	-0.005	(0.023)	0.041	(0.034)	0.026	(0.039)
OTHERS to EU/OTHERS	-0.253***	(0.065)	-0.217*	(0.104)	-0.196	(0.125)
EU/USA to OTHERS	-0.292***	(0.085)	-0.293**	(0.102)	-0.381*	(0.145)
OTHERS to USA/EU	-0.144	(0.121)	-0.275	(0.148)	-0.365*	(0.176)
EU/OTHERS to USA	0.166*	(0.084)	0.102	(0.116)	0.071	(0.122)
Tech. distance	-0.416***	(0.008)	-0.373***	(0.011)	-0.387***	(0.014)
Within firm	0.143***	(0.004)	0.272***	(0.006)	0.271***	(0.007)
N	403962		403962		403962	

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 12: Sensitivity to firm sizes I: Operating revenue.

	(1)		(2)	
EU to USA	0.204	(0.128)	-0.049	(0.032)
EU to OTHERS	-0.192	(0.155)	-0.041	(0.039)
EU to USA/OTHERS	0.383	(0.221)	-0.117*	(0.056)
USA to EU	-0.141**	(0.053)	-0.042*	(0.018)
USA to OTHERS	-0.037	(0.032)	0.048***	(0.011)
USA to EU/OTHERS	0.089	(0.117)	-0.233***	(0.039)
OTHERS to EU	-0.054	(0.097)	-0.095**	(0.029)
OTHERS to USA	-0.059	(0.054)	0.027	(0.015)
OTHERS to EU/OTHERS	-0.260	(0.139)	-0.061	(0.040)
EU/USA to OTHERS	-0.029	(0.151)	0.153***	(0.041)
OTHERS/USA to EU	-0.074	(0.167)	0.165***	(0.048)
EU/OTHERS to USA	0.498*	(0.205)	-0.054	(0.066)
BD	-0.171***	(0.029)	-0.085***	(0.011)
Interaction effects:				
EU to USA	0.401	(0.226)	0.123	(0.072)
EU to OTHER	0.718*	(0.346)	0.088	(0.143)
EU to USA/OTHERS	-0.518	(0.531)	-0.134	(0.138)
USA to EU	0.366*	(0.174)	-0.007	(0.056)
USA to OTHERS	0.167*	(0.073)	-0.104**	(0.033)
USA to EU/OTHERS	0.022	(0.241)	0.063	(0.102)
OTHERS to EU	0.328	(0.530)	0.324*	(0.128)
OTHERS to USA	-0.050	(0.140)	0.037	(0.038)
OTHERS to EU/OTHERS	-0.341	(0.591)	-0.201	(0.114)
EU/USA to OTHERS	-0.042	(0.335)	-0.221*	(0.110)
OTHERS/USA to EU	-0.032	(0.670)	-0.301	(0.161)
EU/OTHERS to USA	0.266	(0.508)	0.082	(0.120)
Tech. distance	-0.375***	(0.037)	-0.312***	(0.013)
Within firm	0.105***	(0.026)	0.243***	(0.006)
N	30843		278428	
Number of firms	1461		487	
Percentile	<75		>75	

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 13: Sensitivity to firm sizes II. Number of employees.

	(1)		(2)	
EU to USA	0.179	(0.107)	-0.053	(0.033)
EU to OTHERS	-0.072	(0.123)	-0.052	(0.039)
EU to USA/OTHERS	0.304	(0.201)	-0.117*	(0.056)
USA to EU	-0.106	(0.056)	-0.046**	(0.018)
USA to OTHERS	-0.018	(0.032)	0.045***	(0.011)
USA to EU/OTHERS	0.046	(0.121)	-0.227***	(0.039)
OTHERS to EU	0.186	(0.107)	-0.109***	(0.029)
OTHERS to USA	0.038	(0.063)	0.020	(0.015)
OTHERS to EU/OTHERS	0.188	(0.142)	-0.085*	(0.040)
EU/USA to OTHERS	-0.120	(0.124)	0.169***	(0.042)
OTHERS/USA to EU	-0.172	(0.155)	0.174***	(0.048)
EU/OTHERS to USA	-0.098	(0.324)	-0.011	(0.064)
BD	-0.162***	(0.028)	-0.086***	(0.011)
Interaction effects:				
EU to USA	0.639*	(0.269)	0.125	(0.070)
EU to OTHER	0.852*	(0.344)	0.103	(0.140)
EU to USA/OTHERS	-1.939	(1.031)	-0.079	(0.136)
USA to EU	0.226	(0.176)	0.007	(0.056)
USA to OTHERS	0.142*	(0.069)	-0.107**	(0.033)
USA to EU/OTHERS	0.163	(0.241)	0.031	(0.102)
OTHERS to EU	0.110	(0.437)	0.335**	(0.129)
OTHERS to USA	-0.154	(0.164)	0.042	(0.037)
OTHERS to EU/OTHERS	-1.069	(0.673)	-0.167	(0.114)
EU/USA to OTHERS	-0.054	(0.221)	-0.213	(0.119)
OTHERS/USA to EU	0.022	(0.503)	-0.309	(0.163)
EU/OTHERS to USA	1.287	(0.999)	0.058	(0.118)
Tech. distance	-0.390***	(0.037)	-0.311***	(0.013)
Within firm	0.090***	(0.026)	0.245***	(0.006)
N	30727		278544	
Number of firms	1461		487	
Percentile	<75		>75	

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

References

Atkeson, A. and P. J. Kehoe (2007). Modeling the transition to a new economy: lessons from two technological revolutions. *The American Economic Review 97*(1), 64–88.

Bresnahan, T. F., E. Brynjolfsson, and L. M. Hitt (1999). Information technology, workplace organization and the demand for skilled labor: Firm-level evidence. Technical report, National Bureau of Economic Research.

Bresnahan, T. F. and M. Trajtenberg (1995). General purpose technologies: Engines of growth? *Journal of Econometrics 65*(1), 83–108.

Cohen, W. M. and D. A. Levinthal (1990). Absorptive capacity: A new perspective on learning and innovation. *Administrative Science Quarterly 35*(1), 128–152.

Comin, D. and B. Hobijn (2004). Cross-country technology adoption: making the theories face the facts. *Journal of Monetary Economics 51*(1), 39–83.

Cox, M. and D. Ellsworth (1997). Application-controlled demand paging for out-of-core visualization. In *Proceedings of the 8th Conference on Visualization'97*, pp. 235–ff. IEEE Computer Society Press.

Dean, J. and S. Ghemawat (2008). Mapreduce: simplified data processing on large clusters. *Communications of the ACM 51*(1), 107–113.

Griffith, R., S. Lee, and B. Straathof (2014). Recombinant innovation and the boundaries of the firm. *CEMMAP working paper: CWP40/14*.

Griffith, R., S. Lee, and J. Van Reenen (2011). Is distance dying at last? Falling home bias in fixed effects model of patent citations. *Quantitative Economics 2*, 211–249.

Hall, B. H. and B. Khan (2003). Adoption of new technology. Technical report, National Bureau of Economic Research.

Hilbert, M. and P. López (2011). The worlds technological capacity to store, communicate, and compute information. *Science 332*(6025), 60–65.

Jaffe, A. B. and M. Trajtenberg (1999). International knowledge flows: evidence from patent citations. *Economics of Innovation and New Technology 8*(1-2), 105–136.

LEF (2011). Data revolution. Technical report, The Leading Edge Forum.

Mansfield, E. (1961). Technical change and the rate of imitation. *Econometrica 29*(4), 741–766.

Manyika, J., M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers (2011). Big Data: The next frontier for innovaion, competition, and productivity. *McKinsey Global Institute*.

McAfee, A., E. Brynjolfsson, T. H. Davenport, D. Patil, and D. Barton (2012). Big data: The management revolution. *Harvard Business Review 90*(10), 61–67.

Nagaoka, S., K. Motohashi, and A. Goto (2010). Patent statistics as an innovation indicator. *Handbook of the Economics of Innovation 2*, 1083–1127.

OECD (2010). *OECD Information Technology Outlook 2010*. OECD Publishing.

Olson, M. (2010). Hadoop: Scalable, flexible data storage and analysis. *IQT Quart 1*(3), 14–18.

Ridder, G. and I. Tunali (1999). Stratified partial likelihood estimation. *Journal of Econometrics 92*(2), 193–232.

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.

Thompson, P. and M. Fox-Kean (2005). Patent Citations and the Geography of Knowledge Spillovers: A Reassessment: Reply. *American Economic Review 95*(1), 465–466.

UK Intellectual Property Office (2014). Eight great technologies: Big data.

WIPO (2015). WIPO IP Facts and Figures, Economics & Statistics Series.