Compendium of additional material

This compendium contains additional material to the paper "Copyright's impact on data mining in academic research", under review. The material covers two main issues that could not be included in the submitted draft due to length restrictions, but that deserve further explanation. The first part of this document delivers a detailed explanation of the process followed to build the paper's main independent variable: the level of copyright restriction. Given the highly technical legalistic foundation of this variable, which results in a cross-country, longitudinal measure, the legal technical criteria used to build the measure deserve explanation. A subsection of this first part gives also details about further control variables used in the models. On the other hand, the second part of this material gives further information about the building process of our dependent variable: the share of papers about data mining of a particular country on a particular year. In particular, it focuses on the text search, keyword-based method used to identify and retrieve the papers on data mining produced in the largest 42 economies in the world, between 1992 and 2014.

1. Main independent variable: copyright

Most forms of academic data mining (DM) research entail the reproduction and search of vast amounts of data, followed by the communication of the results to the public. Where the DM process involves the copying, digitization, or reformatting of copyright protected material without the rights owner's permission, these acts may give rise to copyright infringement. This holds even if the user has lawfully accessed the data. In the absence of the rights holder's permission, the mining of copyright protected material can take place lawfully only if copyright law provides for an exception covering this type of activity.

Despite international and regional harmonization efforts,[1] copyright protection is determined at the national level (Goldstein and Hugenholtz, 2012). Different elements of the copyright regime may have an impact on the lawfulness of DM activities, for instance: the definition of the protected subject matter (are compilations of data, a 'work'?); the requirement for protection (do compilations meet the requirement of 'originality'?); and the duration of the protection. The two features of the copyright system that bear the most on DM are the scope of rights granted on compilations of articles and other data, and the exceptions on these rights recognized in the various jurisdictions. The grant of exclusive rights on a compilation or a database determines the extent to which a specific use of that compilation/database is subject to the prior permission of the rights owner. The need to obtain such permission is, however, subject to the existence of exceptions and limitations which generally permit the unatuhorised use of a copyrighted work under specified circumstances.

Depending on the legal system, exceptions and limitations are provided either in statutory law or developed by case law. As we shall see below, less than a handful of countries recognise at this time, an explicit exception allowing DM activities to take place for research purposes without the authorization of the rights holder, either as a result of legislative intervention or of judicial interpretation. By contrast, the exceptions contained in the laws of a majority of countries are too narrow to cover DM activities and consequently reserve such activities exclusively to the rights holder. In many countries, however, the law is unclear: in other words, neither the wording of a statutory exception nor the interpretation by the courts allows a clear determination of the lawfulness of DM activities. Only a ruling by a high court would settle the issue definitely, which has so far hardly lead to any litigation. Our assessment of the state of the copyright rules in each jurisdiction is based on a reading of the current legislative provisions, as well as the scholarly commentaries and the judicial interpretation, when available.[2] In the following, we thus classify countries according to whether DM by academic researchers, who have lawful access to data, is either definitely 'not allowed', 'probably not allowed', 'probably allowed', or definitely 'allowed'.

Among the countries examined, sixteen belong to the European Union (EU) or the European Economic Area (EEA). As such, these countries are bound by the secondary legislation adopted at the European level in the area of copyright law. Directive 2001/29/EC confers rights owners with the exclusive right to reproduce, communicate to the public and distribute their works. Member States must ensure that the reproduction right covers both the direct and indirect, as well as the temporary and permanent making of reproductions of works by any means, in whole or in part. In addition to the protection afforded under European copyright law to original compilations

[1] For instance the Berne Convention of 1971, the 1994 Agreement on Trade-related Aspects of Intellectual Property (TRIPs Agreement), the WIPO Copyright Treaty of 1996, as well as bilateral treaties and European directives.

[2] This assessment could in the future be subject to contrary interpretation by the courts, should a dispute arise in any of these countries.

(e.g. by virtue of their "selection or arrangement"), Directive 1996/9/EC grants protection with respect to non-original databases if they show a substantial investment in the obtaining, verification or presentation of the data. This specific database right grants the maker of a database the exclusive right to prevent the extraction and/or re-utilization of the whole or of a substantial part, evaluated qualitatively and/or quantitatively, of the contents of that database. The rights granted under Directive 2001/29/EC and Directive 1996/9/EC have traditionally received a broad interpretation from the European Court of Justice (ECJ) (Hargreaves et al., 2014; Triaille, 2013).

The Directive 2001/29/EC contains a list of exceptions on these exclusive rights, the most relevant of which in the context of data mining allows EU Member States to provide for exceptions in the case of 'use for the sole purpose of illustration for teaching or scientific research, as long as the source, including the author's name, is indicated, unless this turns out to be impossible, and to the extent justified by the non-commercial purpose to be achieved'. Like the vast majority of other exceptions in the Directive, this exception is optional; Member States may decide whether to implement it or not (Guibault 2010, p.55). As a result, the research exception is unevenly implemented at national level (Triaille 2013, p.403). The same holds for the database right, where Member States are free to adopt an exception allowing the substantial extraction of the content of a database for research purposes, but not the re-utilization. From the text of the exceptions and the traditionally restrictive interpretation given to them by the ECJ, it follows that DM activities in Europe are generally seen as not being covered by any exception and therefore as falling within the scope of exclusive rights (Triaille, 2014). Most EU and EEA Member states are thus classified as 'academic DM *not allowed* without express consent'.

The UK differs from the rest of the EU/EEA, as the legislator adopted a specific copyright exception in the course of 2013, allowing DM activities for non-commercial research purposes to take place without the need to obtain prior authorization from the rights holders. Since 2014, the UK is thus classified as 'allowed'. Before, the situation in the UK was similar to that of the rest of Europe ('not allowed').

Switzerland, is not a member of the EU nor of the EEA and is thus not bound by the European legal framework. The Swiss Copyright Act grants authors of original works a number of exclusive rights including that of reproduction, retransmission and making available. Among the several exceptions listed in the act, none would seem to cover acts of mining for purposes of research, beyond the right to make a copy for private use. Switzerland does not protect databases separately. Nevertheless, DM activities are most likely 'not allowed'. The same holds for Russia and Turkey. In these countries, DM activities are most likely 'not allowed'.

The laws of most countries with a colonial past have been influenced by those of the colonial power, most often a European country. The copyright laws of Latin American countries[3] root in the legislation of Continental Europe. The rights and exceptions recognised are similar to their European counterparts. As a result, acts of DM are most likely 'not allowed' in these countries.

While the modern Japanese Copyright Act was strongly influenced by the German Copyright Act of 1965, the recent development of the Japanese Act pursued its own course, under a greater influence from the USA. Up to 2009, DM activities were 'probably not allowed' in Japan. In that year, Japan is reported to have been the first country in the world to adopt a specific copyright exception allowing the 'analysis of in-copyright works using computers in order to extract statistics and information', and come up with new ideas (Triaille, 2014, p.10). At the current state of our information, Japan is classified as 'allowed' since 2010.[4]

In the countries adhering to the Anglo-American copyright system, there are two different approaches with respect to exceptions on copyright: some countries recognize a 'fair dealing' exception, while others recognize a 'fair use' defense. The countries of the British Commonwealth[5] commonly recognize 'fair dealing' exceptions for different purposes, including for criticism and comment, private use and research. To fall under the fair dealing exception, the purpose of the dealing must qualify as one of the allowable purposes under the copyright act, and the dealing must be fair. The fair dealing exception generally receives a restrictive interpretation, which lets us conclude that DM activities are 'probably not allowed' in most 'fair dealing' countries. Compared to other 'fair dealing' countries, Canada has followed in recent years a more flexible approach: not only has the Supreme Court ruled twice in favour of fair dealing for research purposes, but the Copyright Act was amended in 2012, to expand the allowable fair dealing purposes. Since 2012, DM activities in Canada are 'probably allowed'.

The 'fair dealing' exception differs from the 'fair use' defense primarily in the fact that the latter is characterized by an open-ended list of purposes for which the use of a work may be regarded as fair, marked by the words 'such

[3] Argentina, Brazil, Colombia, Mexico, Venezuela.

[4] The classification of Japan is not straightforward, since this provision excludes its application to databases that are precisely made for data analysis.

[5] Australia, Canada, India, Ireland, Malaysia, Nigeria, Singapore, South Africa, United Kingdom.

as'. The fair use defense was first developed at the beginning of the 20th Century in the USA as a judicial doctrine before being codified in § 107 of the Copyright Act 1976. The assessment of whether a particular use is fair is done by the judge according to four factors: the purpose and character of your use; the nature of the copyrighted work; the amount and substantiality of the portion taken, and the effect of the use upon the potential market. Until the end of 2014, the USA is classified on the basis of the case law as 'probably allowed'.[6] In two recent rulings the Court of Appeals for the Second Circuit has found that text and data mining meets the criteria of fair use and does not, therefore, amount to copyright infringement. This leads us to believe that DM activities now qualify as fair use thereby resulting in a status change of the USA to 'allowed'.

For a long time, the fair use doctrine was a unique feature of the American copyright regime. The copyright acts of some countries, like Israel and the Republic of Korea, contained a list of specific exceptions, which were too narrow to cover acts of DM. However, following the conclusion of a bilateral trade agreement with the USA, both Israel (2008) and the Republic of Korea (2012) introduced a fair use defense in their copyright legislation in addition to a list of specific exceptions. Since the legislative amendment introducing the fair use defense, acts of DM possibly shifted from 'probably not allowed' to 'probably allowed'. Furthermore, Singapore has already changed from 'fair dealing' (i.e. 'probably not allowed') to 'fair use' similar to the USA ('probably allowed') on 1st January 2005 (Tan, 2012; Ghafele and Gibert, 2012).

The People's Republic of China only adopted Berne Convention compliant copyright norms in 2007, upon its accession to the TRIPS Agreement. Before that time, copyright protection on the Chinese territory was below the Berne standard, meaning that the existence and enforcement of copyright rules was not a priority. The Chinese Copyright Act of 2007 recognizes the same basic exclusive rights of reproduction, retransmission and making available. Article 22 of the Act lists the permissible exceptions, including for use of a published work for the purposes of the user's own private study, research or self-entertainment. Literal interpretation of this provision would not permit acts of data mining. However, Geller and Nimmer (2015, CHI 8.72) report that the Supreme People's Court of China issued a policy document at the end of 2011, according to which in circumstances necessary to stimulate technical innovation and commercial development, an act that would neither conflict with the normal use of the work nor unreasonably prejudice the legitimate interest of the author could be deemed "fair use". Such a finding would be conditional to the conformity of the act with the four fair use factors. In other words, the policy document makes it possible to use a work without the permission of the right-holder even if this use is not among those specified under the Copyright Act. This policy document was followed in a 2014 case. Since that time, acts of DM are probably allowed in China, while they were 'probably not allowed' between 2007 and 2012.

Taiwan Copyright Law has a long history, having first been enacted in 1928. Taiwan's modern Copyright Act was adopted in 1992 and contained a list of exceptions, none of which was broad enough to encompass DM activities. In recent years, Taiwan copyright law has been marked by American influence. In 2003 the list of exceptions was complemented by a fair use provision, which must be applied in conjunction with the specific exceptions. Since that time, acts of DM are probably allowed in Taiwan, provided that they meet the four factors used to evaluate fair use in any of its many enumerated circumstances. By contrast, the law of Thailand contains a more restrictive provision according to which DM is 'probably not allowed'.

Although most Muslim countries included in the sample are members of the Berne Convention, finding specific information on the scope of the exceptions in the laws of Iran, Indonesia, Saudi Arabia and the United Arab Emirates proves difficult. They are thus excluded in the data analysis.

See Table 2 for an overview of the country categorization according to DM-related copyright, and Table 3 for the average DM share among the four copyright categories. The largest number of observations is available for the category 'not allowed', and the average DM share for this category is lower than for all other categories. The average DM share for the category 'allowed' is relatively low. This category only contains six observations, so that it is hardly suited for a statistical analysis. Furthermore, these observations come from very recent changes, and the full effect of changing to 'allowed' may transpire over a longer period than covered by our data.

An important remark concerning the legal classification of countries relates to the potential impact of publisher licenses on the number of DM research articles published. In principle contractual agreements between users (e.g. typically research institutions) and rights holders (e.g. publishers of scientific journals) could be a substitute to an exception on copyright to establish the modalities for technical access to the relevant data sets. Especially in countries where DM is 'not allowed', researchers could seek permission directly from the publisher in order to

[6] United States Court of Appeals for the Second Circuit, 13-4829-cv (*Google Books vs. Authors' Guild)* (16.10.2015); United States Court of Appeals for the Second Circuit, June 10, 2014 (*Authors' Guild of America vs. Hathitrust*), No. 12-4547-cv., 755 F.3d 87, 91 (2d Cir. 2014).

engage in DM activities so as to compensate the strict character of the law. In reality, however, licenses are hardly a solution: first, because the licensing practice is only slowly emerging in the academic publishing sector, far from being widespread. Moreover DM license terms are at this time usually more restrictive than permissive, putting limits on what researchers can do with the data sets they wish to mine. Second, because of high transaction costs researchers have difficulty negotiating and concluding case-by-case agreements with publishers (Triaille 2014, p.94).[7] Arguably, if licensing agreements were a proper substitute to a flexible copyright system, the DM research output of countries where DM is 'not allowed' or 'probably not allowed' would be comparable with that of countries where DM is 'probably allowed' or 'allowed'. As we will see below, this is not the case.

Other control variables

Besides the total research output of countries, we use several control variables: (1) GDP per capita as reported by the World Bank World Development Indicators (World Bank, 2015a), with complete data for the 1992-2013 period; (2) country population size, also from official World Bank data (World Bank, 2015a) and complete for the entire time period studied; and (3) the level of rule of law as reported by the Worldwide Governance Indicators Project (WGI) (Kaufmann et al., 1999 & 2010; World Bank, 2015b). This rule of law indicator is one of the six dimensions of governance of the WGI indicators, and is defined as "the extent to which agents have confidence in and abide by the rules of society" (World Bank, 2015b), including the quality of contract enforcement and property rights. We use it as a proxy for the level of enforcement of quasi-property rights such as copyright. Data availability for this indicator begins in 1996 and last estimates are from 2013, but until 2003 we only have estimates for alternate years. To avoid an excessive loss of data, and given the generally low variation of this indicator in industrialized countries, the scores for 1997, 1999, and 2001 for each country were estimated computing the arithmetic mean of the rule of law score in the previous and posterior year. This indicator is normalized to have a mean around zero and standard deviation around 1.

2. Dependent variable: share of papers on data mining

As explained in the paper, our data was collected from Thomson Reuter's Web of Science (WoS), using the entire WoS Core Collection Database including the so-called Science Citation Index Expanded, Social Science Citation Index and Art & Humanities Citation Index.

Our research is concerned with the underlying concept of the search for patterns in large data sets, often from a combination of separate sources of secondary data. The population of articles on this practice is unknown. The aim is to develop a set of search terms, which produce a result reflecting the actual number of articles on this practice as closely as possible. 'Data mining' is a compound expression with a reasonably consistent and well-established definition corresponding to this underlying concept and no alternative meanings in popular use. For our purposes, a parsimonious approach with a single search term 'data mining' turns out to be most adequate.

General problems in searches for key words concern polysemes (where a single term also carries unrelated meanings in some search results) and synonyms (alternative use of expressions with very similar meanings but without referring to the search terms used). The polyseme problem can lead to the erroneous inclusion of unrelated items. The synonym problem can lead to the erroneous omission of relevant items. There is trade-off: first, in practice most expressions are polysemes to various degrees, so that a minimal number of adequate search terms is desirable; second, to minimize the synonym/omission problem, the inclusion of many synonyms and related expressions is desirable.

We considered broader searches with several other search terms. To do so, we checked the ten most widely cited articles and books on data mining on WoS and on Google Scholar as well as the five most popular dictionaries and thesauri on Google for synonyms and closely related terms to "data mining". General thesauri contained no synonyms. Two sources listed multiple, closely related terms (Fayyad et al. 1996; Wikipedia). Table 1 presents the related terms found, as well as the number of articles on WoS for (combinations of) these terms related to the search for patterns in large data sets. We applied two criteria for inclusion of search terms for our main panel: frequency of use and centrality.

'Data mining' and 'machine learning' are the most frequently used terms. Excluding double-counting of articles with authors from several countries, Web of Science contains 14,746 articles featuring 'data mining' as a topic and

[7] Licences for Europe – Working Group on Text and Data Mining (TDM), Report on the 6th Meeting (14 October 2013), available at: https://ec.europa.eu/licences-for-europe-dialogue/en/content/wg4-presentations-6th-meeting-14-october

14,769 articles featuring ‘machine learning’ published between 1992 and 2014. Both these compound expressions feature over five times more frequently than any other single term. ‘Machine learning’ is defined as the “field of scientific study that concentrates on induction algorithms and on other algorithms that can be said to ‘learn’” (Kohavi and Provost 1998). Data mining (or the related term ‘knowledge discovery in databases’ (KDD), which is popular in computer science) is often an aspect of machine learning. As other terms contribute relatively few items to our search results but may each introduce an unknown set of erroneous polyseme items, we exclude these ‘further related terms’.

Table 1. Popular DM related terms and overlap between them

1992 to 2014 / topics (collected 2 October 2015)	No. of articles for term	Articles featuring terms and **‘data mining’**		Articles featuring terms and **‘machine learning’**		Articles featuring terms and any of the **‘further related terms’**	
Most popular terms		Absolute	% share of articles also featuring ‘data mining’	Absolute	% share of articles also featuring ‘machine learning’	Absolute	% share of all articles on specific term also featuring any of the ‘further related terms’
Data mining	14746	--	--	1321	8.94	1531	18.06
Machine learning	14769	1321	8.94	--	--	579	6.83
Further related terms							
Knowledge discovery	2732	1220	44.66	275	10.07	143	5.23(1)
(Knowledge discovery in databases)	282	160	56.74	39	13.83	9	3.19(2)
Analytics	2512	124	4.94	80	3.18	187	7.44(3)
Information extraction	1934	68	3.52	179	9.26	81	4.19(3)
Big data	927	88	9.49	56	6.04	167	18.02(3)
Knowledge extraction	519	94	18.11	39	7.51	60	11.56(3)
Information discovery	169	11	6.51	3	1.78	12	7.10(3)
Data archaeology	10	2	20.00	1	10.00	1	10.00(3)
Information harvesting	5	1	20.00	2	40.00	1	20.00(3)
data pattern processing	0	0	0	0	0	0	0
Separate articles on further related terms(4)	8475						
Separate articles on all terms(4)	36032						

(1) Excluding term itself and ‘knowledge discovery in databases’.
(2) Excluding term itself and ‘knowledge discovery’.
(3) Excluding term itself.
(4) Excluding double-counting for articles featuring several terms in abstract.

Centrality concerns the extent to which search results for one search term overlaps with search results for other expressions related to the underlying concept. Just under 9% of all articles on WoS on ‘data mining’ also contain ‘machine learning’. This is lower than the weighted average overlap between ‘data mining’ and the ‘further related terms’. Of all articles on further related terms, 18.06% also feature ‘data mining’. The corresponding score for ‘machine learning’ is 6.83%. ‘Data mining’ has greater overlap with 7 out of 10 of the further related terms, with one tie. We thus conclude that ‘data mining’ is the more suitable search term for research on search for patterns in large, secondary databases and that the inclusion of 'machine learning' is unlikely to generate more valid results for our purposes. Future research could expand our investigation by addressing and combining different search terms, or the particularly laborious content analysis of complete articles.

We checked the validity of search results for each country (year) by comparing the sum of individual search results with the total number of search results for the entire time period (set of countries). Data were collected over three separate time periods: (1) 20 July 2014 to 23 July 2014; (2) 10 April 2015 and 11 April 2015; and (3) 6 and 7 May 2015. With the same search settings, most results were stable within a narrow range but updates to the database sometimes had an effect also for earlier years. We report and use the data collected in May 2015. To establish whether these data contain a bias in favour of countries where English language articles account for a particularly large share of research articles, we checked for translations and synonyms of 'data mining' in Chinese, Spanish, French and German. We found virtually no additional articles in this manner. Apparently, much of the high quality research output featured on WoS is published in English or at least contains an English summary with the relevant key words.

There is apparently some double-counting with multiple authors. The total number of search results for the 42 countries with 'Topic' defined as "data mining" and 'Year Published' set to "1992-2014" brings up 14,746 separate articles.

Around the year 2007, there is a remarkable dent in a long-term upward trend of the number of DM articles and their share in total research output. We found that this is almost entirely explained by the reclassification of two regular publications, *Lectures Notes in Computer Science* and *Lecture Notes in Artificial Intelligence*. These two periodicals contained a quickly growing number of DM publications between 1996 and 2006 (each with more than eight hundred DM articles in 1996 to 2006 and ca. two hundred DM articles each in 2006). By 2007, papers in both these periodicals were no longer classified as "articles" by WoS but as "proceedings", perhaps for concern with the impact of these papers (average citation <4 for both periodicals). This reclassification explains virtually the entire level change in the number and share of DM publications in 2007, as well as the strong growth between 2002 and 2006 compared to the other years covered.