EPIP Congress 2015

Is Europe falling behind in data mining?
Copyright's impact on data mining in academic research

Christian HANDKE
ESHCC, Erasmus University Rotterdam, Rotterdam and IViR, University of Amsterdam
handke@eshcc.eur.nl

Lucie GUIBAULT[b]
IViR, University of Amsterdam
L.Guibault@uva.nl

Joan-Josep VALLBÉ
GREL, University of Barcelona
vallbe@ub.edu

- Paul Heald:
- USA allowed, some case in 1991.

Introduction

- Data mining requires use of (copyright) works and data, usually in bulk.
- Copyright should affect the supply of potential input works …
- … and the costs of follow-up use of copyrighted works produced by others.
- We present empirical evidence regarding the net effect of differences in copyright on the output of data mining-related academic research.
- Policy debate in particular in the European Union (EU):
  - In most EU Member States, DM requires prior authorization of rights holders even if the potential user has lawful access to the research articles and databases in question.[1]
  - In many other territories, "the right to read is the right mine".

[1] Directives 1996/9/EC (Databases) and 2001/29/EC (InfoSoc).

Empirical strategy

Dependent variable

- Annual number of research articles on "data mining" per country/country type.

Independent variables

- Categorization of countries according to relevant copyright regulation.
- Rule of law.
  Since law ≠ de facto protection.

Control variables

- Total research output.
  We use the share of "data mining" articles in total research output, thus controlling for size and productivity of research activities in various countries (DM share).
- GDP/capita
- Population size

We estimate a multilevel regression model to account for unobserved, constant country differences.

Data

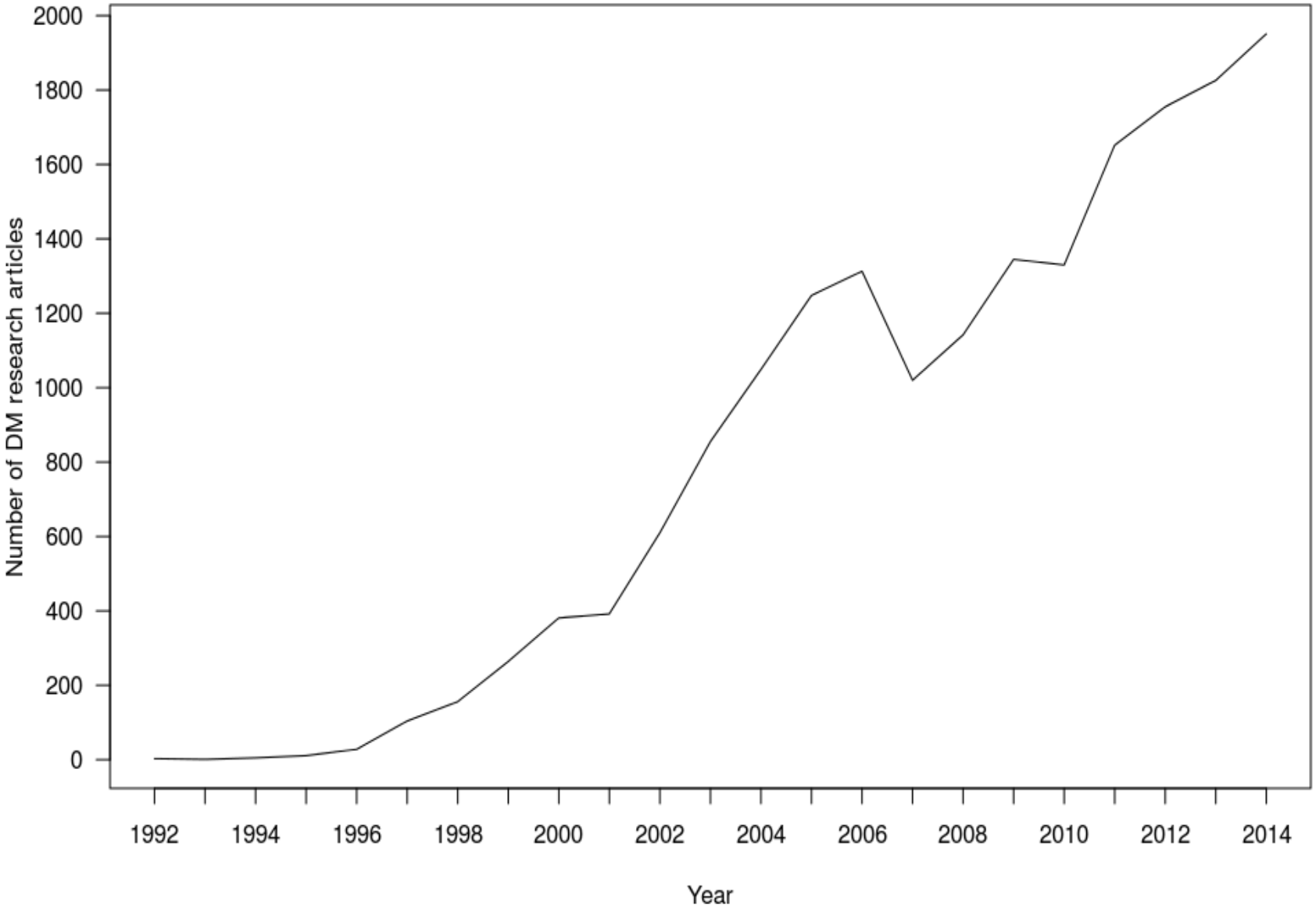
- All data collected from Thomson Reuter's Web of Science (WoS).[1]
- The 42 largest economies by GDP in 2013.
- Yearly data from 1992 to 2014.
  The first "data mining articles" are from 1992.
- That adds up to 966 country-year observations.
- Boolean search of peer-reviewed journal articles with terms:
  Topic = "data mining"; country of residence of author(s); year published

Among the 42 countries covered, searches on WoS brought up:

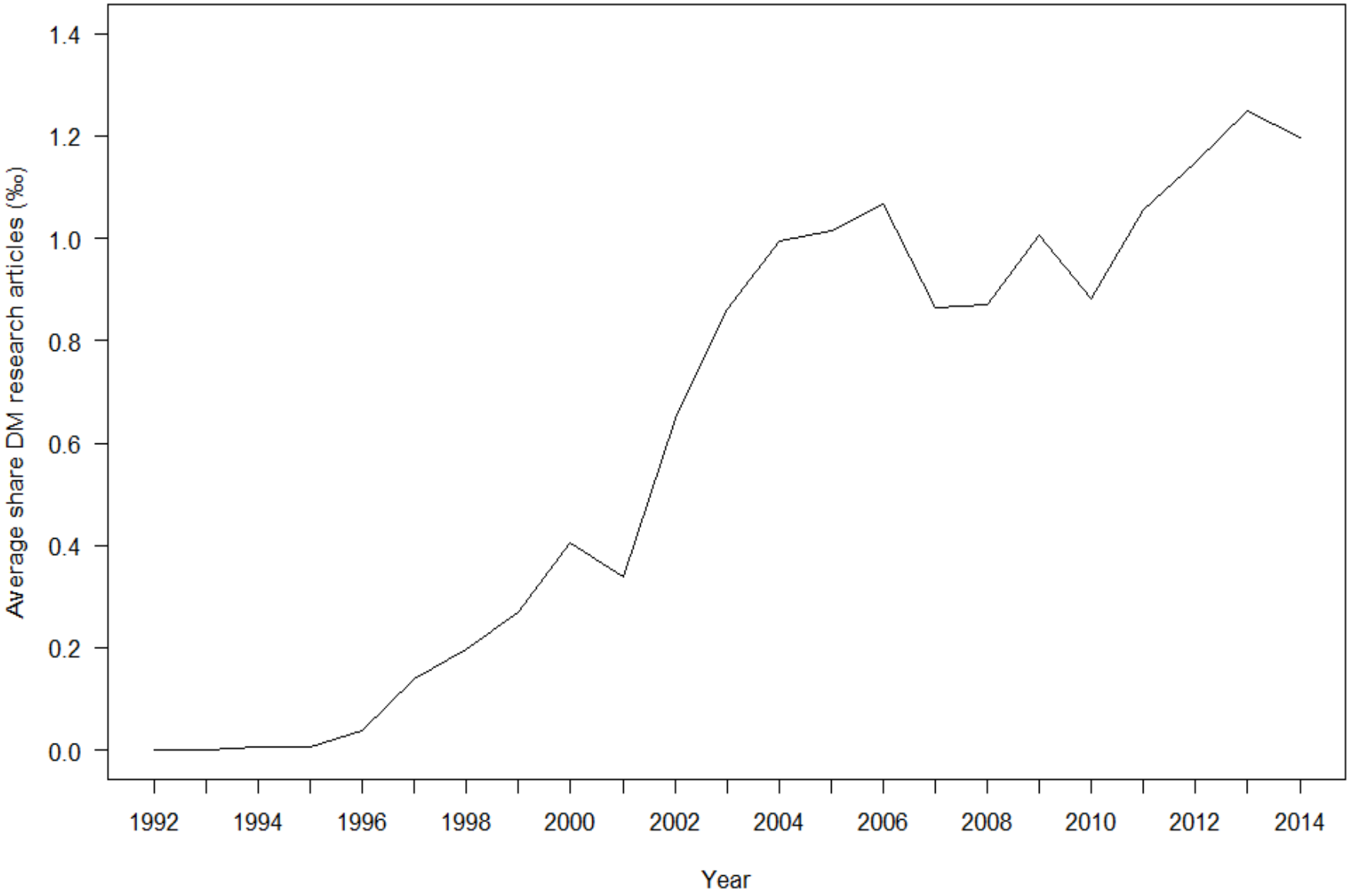
- 18,441 DM-related articles between 1993 and 2014.
- 23,802,650 articles for the entire panel.
- For the entire panel, 0.7‰ of all articles had DM as a topic.

[1] We used the entire WoS Core Collection Database including the so-called Science Citation Index Expanded, Social Science Citation Index and Art & Humanities Citation Index.

Absolute number of DM research articles published per year (42 countries; 1992 to 2014)



Average share of DM research articles in the total number of research articles in ‰ (42 countries; 1992 to 2014)



Country classification

- Countries differ in relevant copyright arrangements
- We classify into four ordinal categories:
  - 'Not allowed'
    Countries with a closed list of exception and limitation, without relevant exception.
    EU/EFTA Member States, Switzerland, Russia, Latin America
  - 'Probably not allowed'
    Countries with a fair use/dealing exception, without relevant case law. *Australia, India, Malaysia, Nigeria, Thailand*
  - 'Probably allowed'
    Countries with fair use/dealing defence, without relevant case law.
    Canada, China, Israel, Korea, Singapore, Taiwan
  - 'Allowed'
    Countries with specific relevant exception or fair use/dealing defense with relevant case law.
    Japan, United Kingdom, United States
- Some status changes between 1992 and 2014.
  United Kingdom (2014), Japan (2010), Canada, China, Israel, Korea, Singapore, Taiwan

Descriptive statistics regarding the DM share in copyright categories

Copyright category	Average	Standard deviation	Number of observations
Not allowed	0.54	0.54	551
Probably not allowed	0.60	0.66	184
Probably allowed	1.76	1.37	59
Allowed	0.70	0.18	6

Results of multilevel regressions with DM share as dependent variable and with varying intercept (random effects).

	Model 1	Model 2	Model 3	Model 4
(Intercept)	0.542***	0.709	3.362***	3.812***
	(0.075)	(0.860)	(0.928)	(0.936)
Copyright [Ref. *Not allowed*]				
Allowed	0.301	0.342	0.288	9.678
	(0.280)	(0.357)	(0.335)	(27.063)
Probably allowed	1.460***	1.453***	1.246***	0.821***
	(0.161)	(0.168)	(0.166)	(0.191)
Probably not allowed	0.073	0.296**	0.183	-0.075
	(0.137)	(0.130)	(0.131)	(0.150)
GDP/capita ($1,000)		0.010**	-0.019***	-0.005
		(0.004)	(0.005)	(0.006)
Population (log)		-0.021	-0.439***	-0.449***
		(0.047)	(0.068)	(0.067)
Rule of Law		-0.001*	-0.001*	-0.392***
		(0.0004)	(0.0004)	(0.103)
Total research output (log)			0.552***	0.525***
			(0.061)	(0.060)
*Definitely allowed**Rule of Law				-6.980
				(20.582)
*Probably allowed**Rule of Law				0.393***
				(0.103)
*Probably not allowed**Rule of Law				0.391***
				(0.103)
R^2	0.165	0.168	0.278	0.308
F	50.079***	18.704***	30.572***	24.657***
	(df = 3; 763)	(df = 6; 557)	(df = 7; 556)	(df = 10; 553)
N	767	564	564	564

*p<0.1; **p<0.05; ***p<0.01

Summary of results

With DM share as the dependent variable and 'not allowed' as the reference category, we find:

- Significant positive coefficients for the category 'probably allowed' and 'probably not allowed' in all specifications.
- Using a multiplicative interaction term, we find that stronger rule of law is associated with significantly lower DM share for countries from the 'not allowed' category.
  (In other categories, rule of law is positively associated with DM research.)
- In particular the combination of strong copyright law and strong rule of law appears to reduce academic researchers' data mining performance.

➔ Overall, there is extensive evidence that DM share is lower in countries with stronger DM-related copyright protection.

Conclusions

- In most EU/EEA Member States, DM-related copyright protection is comparatively strong.
- Our results suggest that the net effect is a weaker performance of domestic academic researchers in this increasingly important type of research.

Is Europe falling behind in data mining?

Copyright's impact on data mining in academic research

Paper available on SSRN:

http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2608513

Contact:

handke@eshcc.eur.nl

L.Guibault@uva.nl

vallbe@ub.edu