

# Investigating Patterns of Technological Innovation

J. Raimbault<sup>1,2\*</sup>, A. Bergeaud<sup>3</sup> and Y. Potiron<sup>4</sup>  
\*`juste.raimbault@parisgeo.cnrs.fr`

<sup>1</sup>UMR CNRS 8504 Géographie-cités

<sup>2</sup>UMR-T IFSTTAR 9403 LVMT

<sup>3</sup>London School of Economics

<sup>4</sup>Faculty of Business and Commerce, Keio University

CSS 2016 - Amsterdam

*Session CFS1*

22th September 2016



## Why study patents ?

→ Applied Epistemology : particular case of the ecology and evolution of knowledge ; propagation of knowledge

→ Economics [Griliches, 1998] : dynamics of innovation ; R&D returns ; innovation at the core of endogenous growth theories [Aghion and Howitt, 1992]

### Examples :

- [Youn et al., 2015] interaction between technological fields ; combinatorial nature of inventions
- [Bruck et al., 2016] citation network analysis to detect emerging research front
- [Gerken and Moehrle, 2012] [Tseng et al., 2007] semantic analysis

## Research Context

Technological classification by Offices does not anticipate future research fields ; retrospective reclassification at some times when classification become obsolete

*However, information contained in patent themselves should contain some kind of endogenous structure, as emerge from microscopic activities of inventors/firms : e.g. semantic dynamics show combinations before classes are established by Offices*

**Research Objective :** *Investigate endogenous patterns of technological innovation contained in particular in semantic content of patents and metadata ; assess complementarity of other informations and potentialities for economic modeling*

# Database Construction

Construction of a Database from US Patent and Trademark Office redbook 1976-2012 (full patent description), which provide raw data but on separate files and different formats

## Data Collection Procedure

- Automatic download of raw data file
- Parsing depending on format : dat or xml (varying schema)
- Uniformisation and storing in MongoDB

→ 4391272 patents with text data ; dated by application date (current state of knowledge)

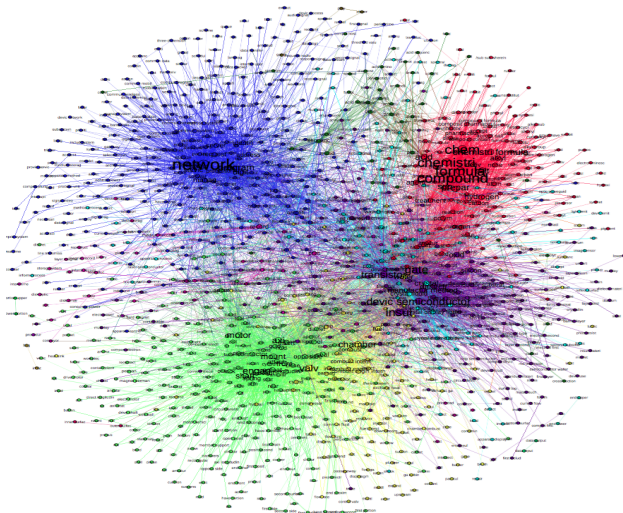
# Extraction of Semantic Features

*Text-mining in python with nltk* [Bird, 2006], method adapted from [Chavalarias and Cointet, 2013]

- Parsing and tokenizing / pos-tagging (word functions) / stemming done nltk built-in pos-tagger
- Selection of potential *n*-grams (with  $1 \leq n \leq 3$ ) : English  $\cap \{NN \cup VBG \cup JJ\}$
- Database insertion for instantaneous utilisation (several days  $\rightarrow$  1min)
- Estimation of *n*-grams relevance, following co-occurrences statistical distribution (*termhood* score as chi-2 score)

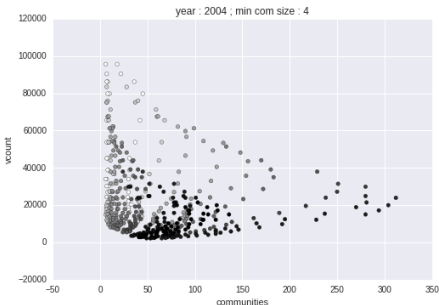
# Semantic Network

*Co-occurrence network of keywords*



# Semantic Classes Construction

Communities in network not well defined (presence of connector words) : need for a filtering, done on technological class disparity (exogenous control) and edge weight ; pareto optimization on (network size, communities)





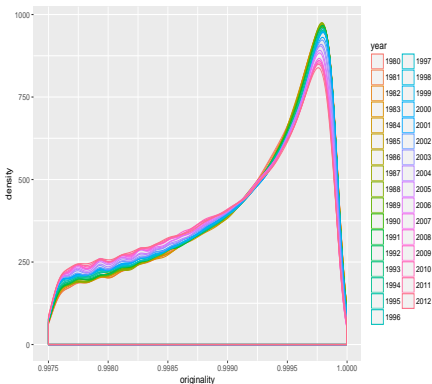
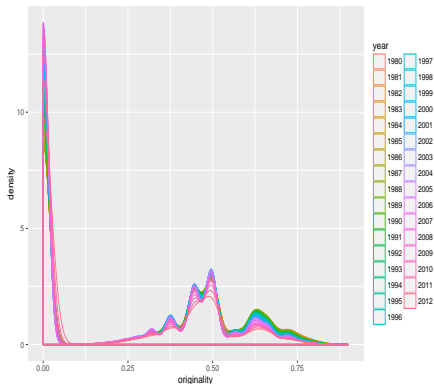
## Classes Examples (2000-2004)

- Memory devices : semiconductor memori devic ; memori cell plural ; memori cell transistor ; layer ferroelectr
- Chemical analysis : time-of-flight mass spectromet ; chromatograph column ; ion trap mass
- Particular steel : martensit ; austenit stainless steel
- Laser : emit laser beam ; vertic caviti surfac ; vcsel
- Sewing : circular knit machin ; stitch ; sew machin ; embroideri
- Lithography : lithograph mask ; project beam radiat ; heat-sensit ; planograph print plate
- Tobacco : cigarett filter ; cigarett pack ; tobacco ; tobacco rod

# Results : Patent-level originalities

Patent originality  $o_i = 1 - \sum_k p_{ik}^2$

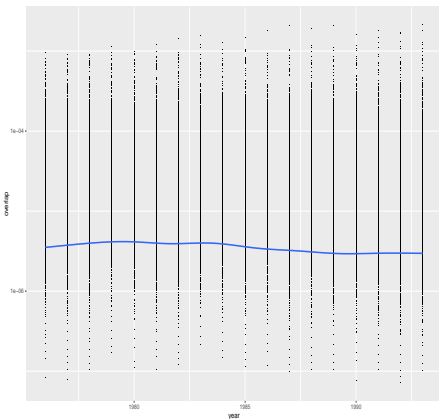
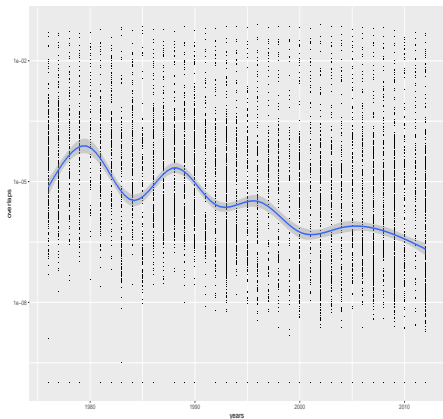
*Decrease (resp. increase) of technological (resp. semantic) originality*  
→ **increase of invention originality in time**



## Results : Communities overlaps

Overlap between classes as a measure of specialization

*Decrease of semantic overlap ; constant technological overlap*  
→ **increase of specialization in time**



# Discussion

## Theoretical and Practical Implications

- Seems to contain relevant information ; full complementarity/relation with other type of information (citations, classification) to be confirmed
- Semantic analysis methodology can be applied to any network whose nodes have textual description, adding a layer to the network

## Further Developments and Potentialities

- Bipartite Network patents/keywords analysis
- Machine Learning to identify innovative patents/emerging research fronts
- Stochastic Block modeling with citation links to statistically validate classifications
- Application for interactive exploration of semantic database

## Conclusion

- Explorative insight into patterns of technological innovation ; semi-big-data treatments (previous literature always restrained to subfields)
- Crucial role of interdisciplinarity and integration of qualitative/quantitative approaches

*Code available at <https://github.com/JusteRaimbault/PatentsMining>*

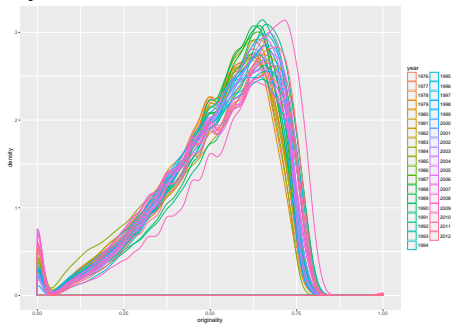
*Data available on Dataverse [Raimbault, 2016], at <http://dx.doi.org/10.7910/DVN/BW3ACK>*

## Reserve Slides

*Reserve slides*

# Sensitivity to Window Size

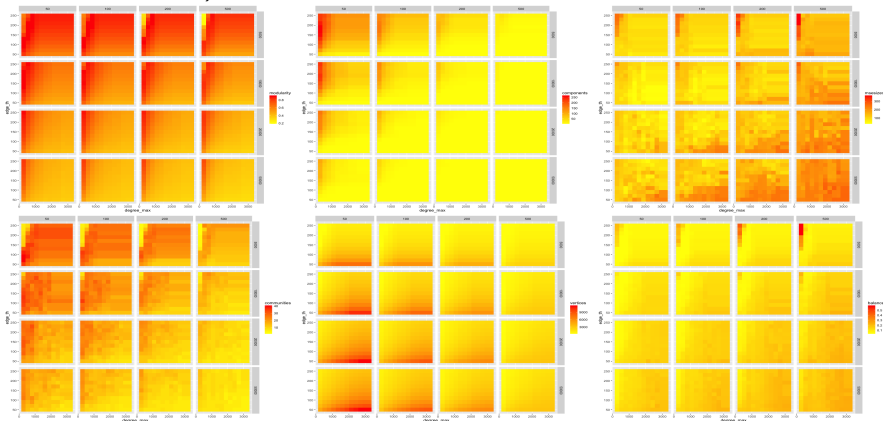
1 year window : more noise in classes



3 year/10 year windows : work in progress

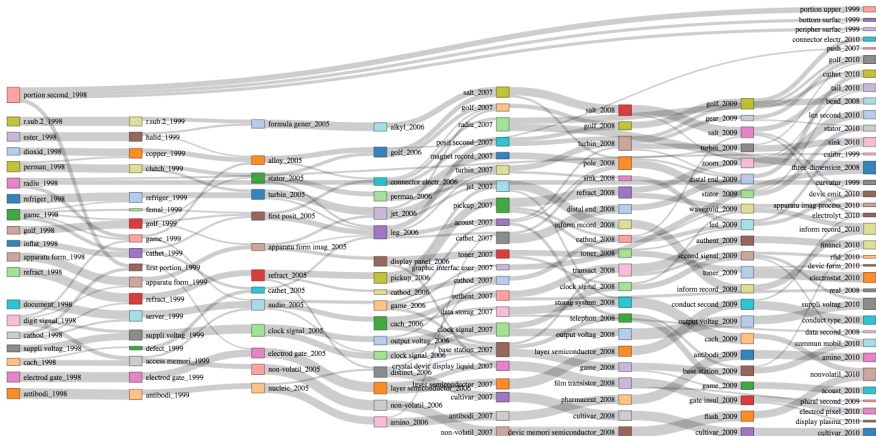
# Graph clustering : alternatives

Alternative pre-clustering filtering : max degree, edge weight, min/max frequency  $\rightarrow$  more difficult sensitivity analysis but more intuitive (*used here for overlaps*)











# Temporal Semantic Flows







## References I

-  Aghion, P. and Howitt, P. (1992).  
A Model of Growth through Creative Destruction.  
*Econometrica*, 60(2):323–51.
-  Bais, S. (2010).  
*In praise of science: curiosity, understanding, and progress*.  
MIT Press.
-  Bird, S. (2006).  
Nltk: the natural language toolkit.  
*In Proceedings of the COLING/ACL on Interactive presentation sessions*, pages 69–72. Association for Computational Linguistics.

## References II

-  Bruck, P., Réthy, I., Szente, J., Tobochnik, J., and Érdi, P. (2016). Recognition of Emerging Technology Trends. Class-selective study of citations in the U.S. Patent Citation Network.  
*ArXiv e-prints.*
-  Chavalarias, D. and Cointet, J.-P. (2013). Phylomemetic patterns in science evolution—the rise and fall of scientific fields.  
*Plos One*, 8(2):e54847.
-  Gerken, J. M. and Moehrle, M. G. (2012). A new instrument for technology monitoring: novelty in patents measured by semantic patent analysis.  
*Scientometrics*, 91(3):645–670.

## References III

-  Griliches, Z. (1998).  
Patent statistics as economic indicators: a survey.  
*In R&D and productivity: the econometric evidence*, pages 287–343.  
University of Chicago Press.
-  Raimbault, Juste; Bergeaud, A. P. Y. (2016).  
Uspto redbook patent data 1976-2012.
-  Tseng, Y.-H., Lin, C.-J., and Lin, Y.-I. (2007).  
Text mining techniques for patent analysis.  
*Information Processing & Management*, 43(5):1216–1247.
-  Youn, H., Strumsky, D., Bettencourt, L. M. A., and Lobo, J. (2015).  
Invention as a combinatorial process: evidence from us patents.  
*Journal of The Royal Society Interface*, 12(106).