

Classifying Patent Based on their Semantic Content

An empirical exploration into patent text mining

Antonin Bergeaud Yoann Potiron Juste Raimbault

Banque de France, Keio University and ISC-PIF

Computational and Financial Econometrics
Session Text Mining
14 December 2018

Plan

1 Introduction

2 Background

3 Methods

4 Results

5 Conclusion

Motivation

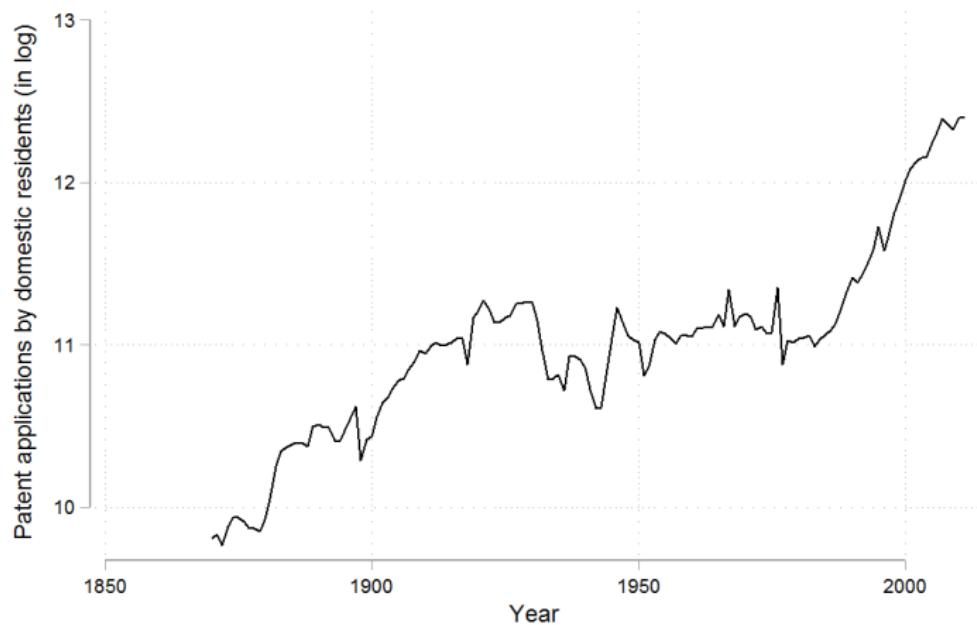


Figure 1: Log number of patent applications per capita at the USPTO. Source : USPTO

Motivation

What are patents?

- Property rights to an invention (novelty, non-obviousness and industrial applicability). Set the state of the art of a technology.
- Classification system in technology fields (or classes).
- Concept of “Person having ordinary skill in the art”: an invention must be sufficiently disclosed in the description of the patent.
- Different realities between countries/IP offices (USPTO, EPO, JPO, CNIPA...).

In practice, patent data are available in the form of a relational database.

Motivation



US 20170278033A1

(19) **United States**

(12) **Patent Application Publication**
Van Wonterghem

(10) **Pub. No.: US 2017/0278033 A1**
(43) **Pub. Date: Sep. 28, 2017**

(54) **DETERMINING COMPLEMENTARITY**

(71) Applicant: **Geert Arthur Edith Van Wonterghem,**
Zwijndrecht (BE)

(72) Inventor: **Geert Arthur Edith Van Wonterghem,**
Zwijndrecht (BE)

(21) Appl. No.: **15/506,341**

(22) PCT Filed: **Aug. 27, 2015**

(86) PCT No.: **PCT/BE2015/000039**

§ 371 (c)(1),

(2) Date: **Feb. 24, 2017**

(30) **Foreign Application Priority Data**

Aug. 29, 2014 (BE) 2014/0675

Publication Classification

(51) **Int. Cl.**
G06Q 10/06 (2006.01)
G06F 17/30 (2006.01)

(52) **U.S. CL.**
CPC **G06Q 10/0637** (2013.01); **G06F 17/3007**
(2013.01); **G06F 17/30011** (2013.01)

(57) **ABSTRACT**

Method for determining an indication of complementarity between two entities, wherein a word database is compiled for each entity in which word clusters related to the entity are entered, wherein at least two areas are distinguished in the database for each entity, these being—areas of activity of the entity; and—areas of capacity of the entity, wherein an algorithm is used to calculate a semantic similarity between the areas of activity and between the areas of capacity of the two entities; wherein the method produces a positive indication of complementarity when the first and second semantic similarity lie respectively above and below a threshold value.

Figure 2: Example of patent front page. Source : USPTO

Why study patents ?

- Although imperfect, patents are the most commonly used measure of innovation in economics.
- Applied Epistemology : particular case of the ecology and evolution of knowledge; diffusion of knowledge.

Patents and text-mining:

- Main usage: matching between patents and firms/inventors
- Used in the search of prior-art/litigation
- Technology mapping

A large scale semantic insight into USPTO database

Proposed approach

Complement existing patent office classifications using patent semantic content

Why?: more endogenous, flexible and informative (?). And comparison with *technological classification* to detect outliers.

Takeaway results

- *An endogenous semantic classification is constructed for the full USPTO abstracts and titles, 1976-2012*
- *Information carried in the semantic classification is complementary to the technological classification*

An example of classification outcome



US006833896B2

(12) **United States Patent**
Lee

(10) **Patent No.:** US 6,833,896 B2
(45) **Date of Patent:** Dec. 21, 2004

(54) **METHOD OF FORMING AN ARRAY
SUBSTRATE FOR AN IN-PLANE
SWITCHING LIQUID CRYSTAL DISPLAY
DEVICE HAVING AN ALIGNMENT FILM
FORMED DIRECTLY ON A THIN FILM
TRANSISTOR**

6,407,791 B1 * 6/2002 Suzuki et al. 349/129
6,642,972 B2 * 11/2003 Yoo et al. 349/40
2001/0040648 A1 * 11/2001 Ono et al. 349/43
2002/0093600 A1 * 7/2002 Choi 349/43

FOREIGN PATENT DOCUMENTS

(75) Inventor: **Yun-Bok Lee**, Seoul (KR)
(73) Assignee: **LG.Philips LCD Co., Ltd.**, Seoul (KR)

JP	09-005764	1/1997
JP	09-073101	3/1997
JP	09-101538	4/1997
JP	09-105908	4/1997

Technological classes: 349 (Liquid Crystal Cells, Elements and Systems), 257 (Active solid-state devices)

Prior art keywords (Google Patents): layer, method, gate, electrode, substrate

Most central keywords (our method): substrat, semiconductor, insul, gate, transistor

Plan

1 Introduction

2 Background

3 Methods

4 Results

5 Conclusion

Descriptive statistics

Summary statistics for the USPTO database between 1976 and 2013

- 4,666,365 utility patents from 1976 to 2013
- 70,000 in 1976 to 278,000 in 2013
- 38,756,292 citation links (84% of within-class citations)
- 270,877 patents with no citations 5 years next to publication

Does this mean that innovation is 4 time larger en 2013 than in 1976...

Not necessary. Although definitely more specialized.

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

→ 4,666,365 patents with text data (abstract); dated by application date (current state of knowledge, differs from the granted date which implies review processes and an exogenous intervention)

Plan

1 Introduction

2 Background

3 Methods

4 Results

5 Conclusion

Extracting relevant n-grams

Text-mining in python with nltk [Bird, 2006], method adapted from [Chavalarias and Cointet, 2013]. Advantage over LDA: scalability and flexibility

- Parsing and tokenizing / pos-tagging (word functions) / stemming with nltk
- Selection of potential *n-grams* (with $1 \leq n \leq 3$) with the rule
 $\cap\{NN \cup VBG \cup JJ\}$
- Estimation of *n-grams* relevance, following co-occurrences statistical distribution (*termhood* score as chi-2 score), filtering on relevance to construct co-occurrence network
- Filtering of network edge (parameter θ_w), with an additional exogenous control by technological class keyword concentration to filter nodes (parameter θ_c)

Plan

1 Introduction

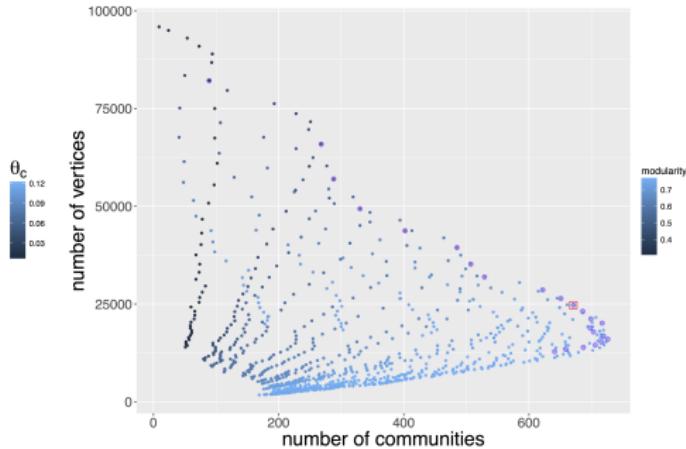
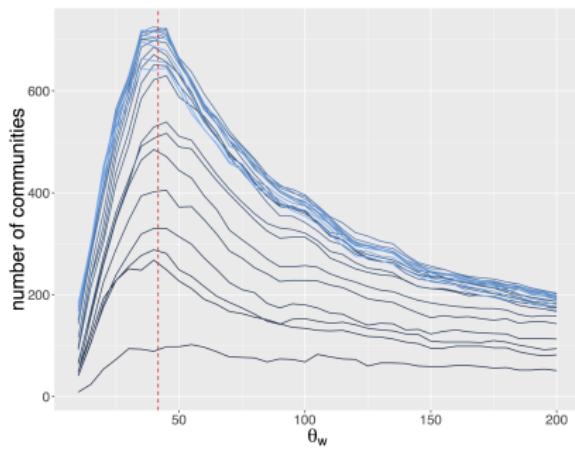
2 Background

3 Methods

4 Results

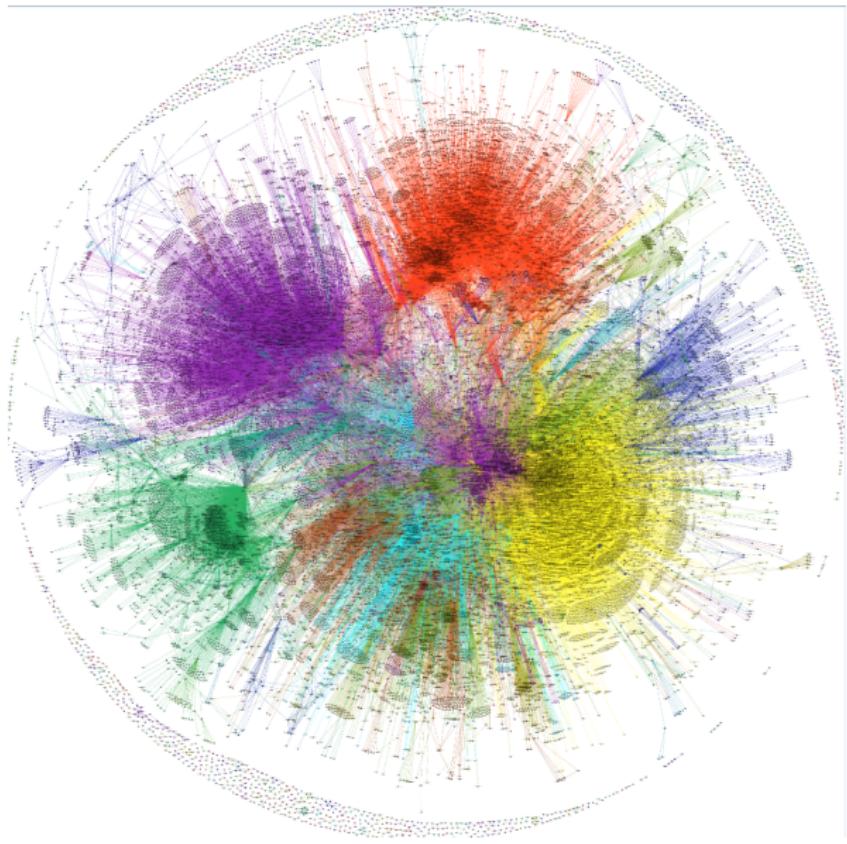
5 Conclusion

Optimizing network structure



Pareto optimization on cutoff parameters with the objectives of modularity, size and number of communities

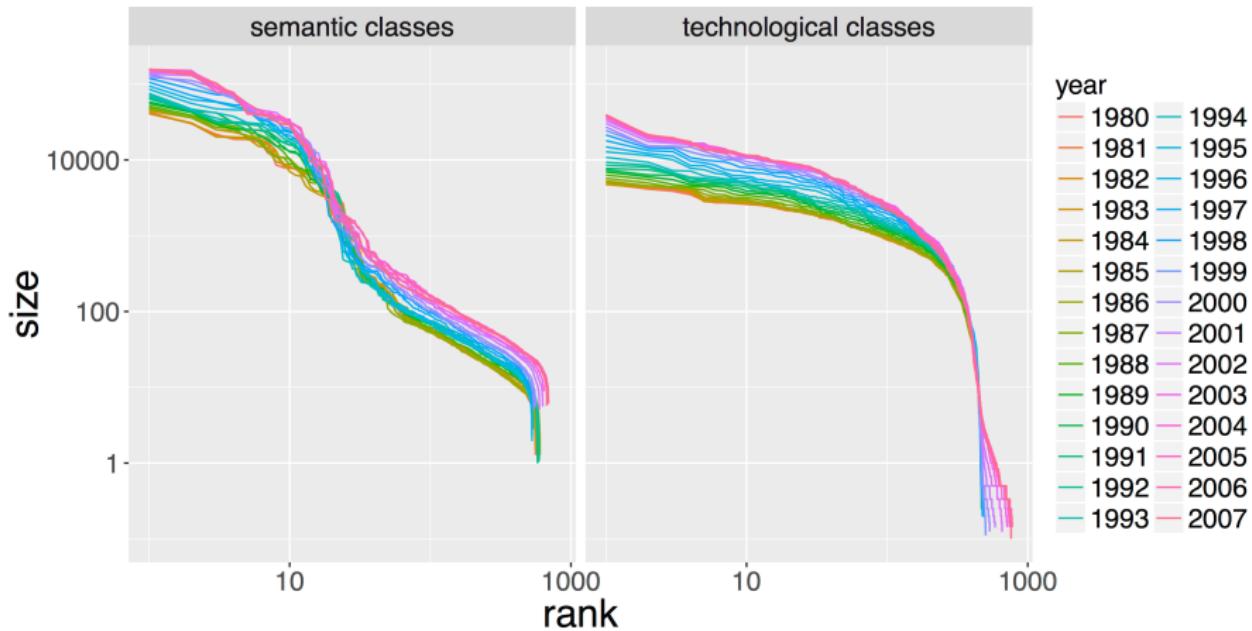
Semantic network visualization (example for 2000-2004)



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

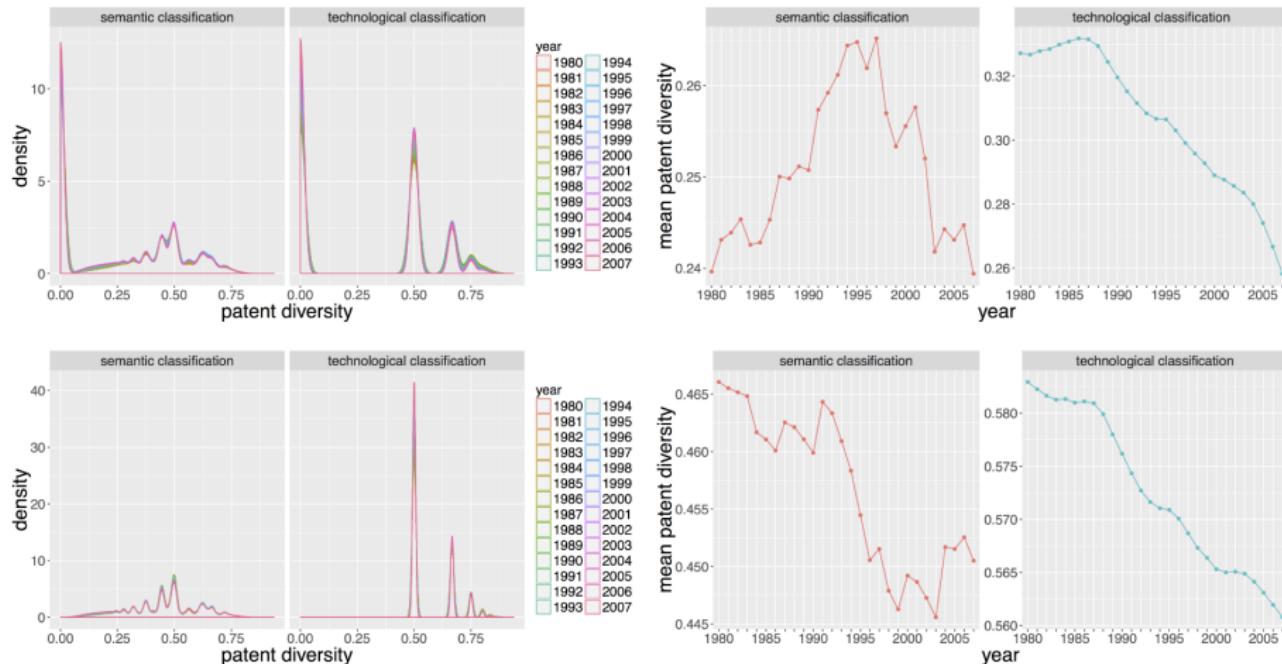
Hierarchical class structure



Fat-tail distribution of class size for both classification, closer to a power-law for semantic classes

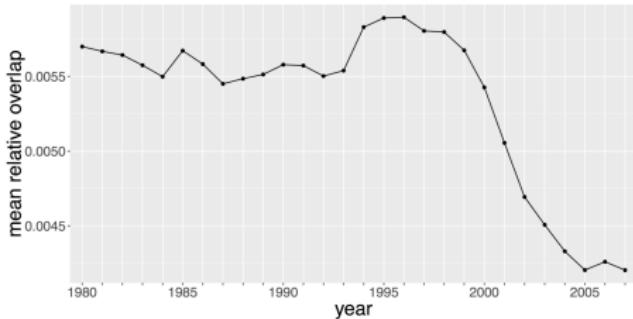
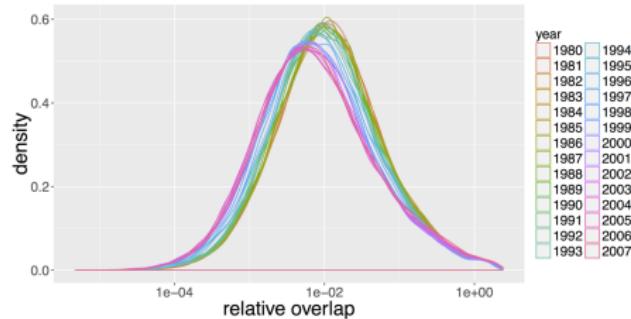
Evolution of patent diversities

Back



General increase in average invention specialization seen both for semantic and technological; semantic regime shift in 1996

Inter-classification overlaps

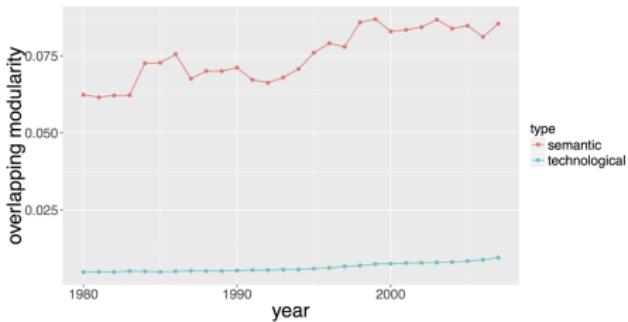
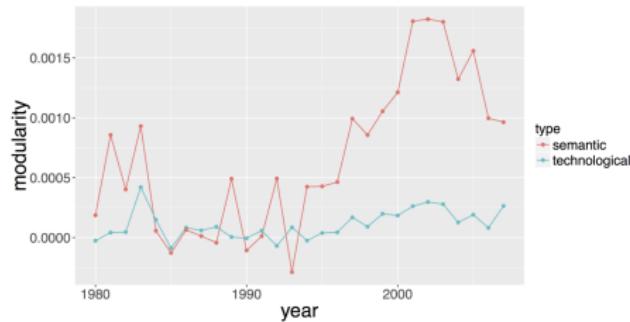


Constant then decreasing average overlap between technological and semantic classes; confirms the change in nature of inventions around 1996.

→ an impact of new information technologies (regarding knowledge production: from expert and contextualized to automatized search in references e.g.) ? Linked to structural changes in economy (increase in firm concentration since 90s) ?

Modularity of classifications

Definition



Semantic classification is more significant regarding the structure of the citation network, both for single-class and multi-class modularities (confirmed statistically using a simple Stochastic Block Model).

Summary of outcomes

- A clean redbook
- Semantic and technological classification for each patent
- Network based measure of centrality for each patent

Most important insight: semantic classification does carry some information beyond what the technological classification provides.

Plan

1 Introduction

2 Background

3 Methods

4 Results

5 Conclusion

Possible developments

Direct

- Look at the correlation between patent quality indicators and centrality indices.
- Use patent-firm matching to study a *firm-level* semantic network.
- Extending the analysis to other patent offices (EPO/JPO).

Possible developments

Direct

- Look at the correlation between patent quality indicators and centrality indices.
- Use patent-firm matching to study a *firm-level* semantic network.
- Extending the analysis to other patent offices (EPO/JPO).

Economic

- Linking semantic measures to values of firms
- Can semantic proximity help understand M&A?
- Firms trajectories in relation to their life-cycle

Possible developments

Direct

- Look at the correlation between patent quality indicators and centrality indices.
- Use patent-firm matching to study a *firm-level* semantic network.
- Extending the analysis to other patent offices (EPO/JPO).

Economic

- Linking semantic measures to values of firms
- Can semantic proximity help understand M&A?
- Firms trajectories in relation to their life-cycle

Technology

- Measure of complementarity between technology?
- Possible application to detection of emerging research fronts
- An interactive exploration of semantic content?

Other projects

At the interface of geography and economics

- Quantifying the diffusion of innovation in urban systems, and its co-evolution with the socio-economic structure

Possible developments in other disciplines

- Agent-based modeling of interactions between technologies
- Full-text mining and history of technology

Conclusion

- A quantitative epistemology insight into the evolution of technology
- Towards reflexive approaches in science and technology to ease technological transfer / foster the co-evolution between the two ?

Paper: Bergeaud, A., Potiron, Y., & Raimbault, J. (2017). Classifying patents based on their semantic content. PloS one, 12(4), e0176310.

Open repository at

<https://github.com/JusteRaimbault/PatentsMining>

Raw database at <http://dx.doi.org/10.7910/DVN/BW3ACK>

Semantic classification database at <http://dx.doi.org/10.7910/DVN/ZULMOY>

Acknowledgments: thanks to ISC-PIF for access to the computation infrastructure.

Reserve slides

Reserve slides

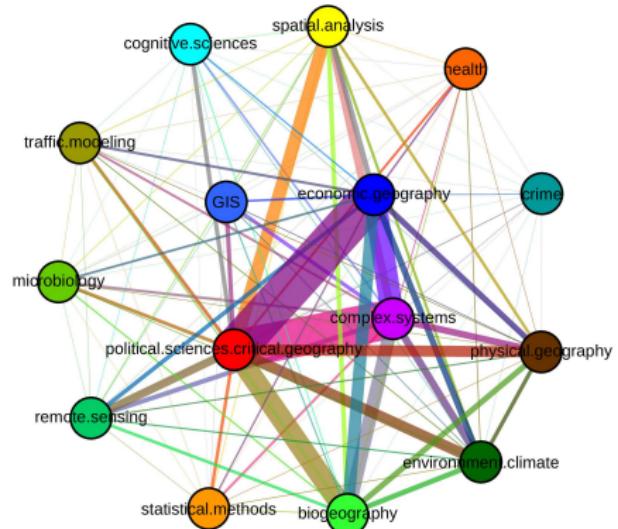
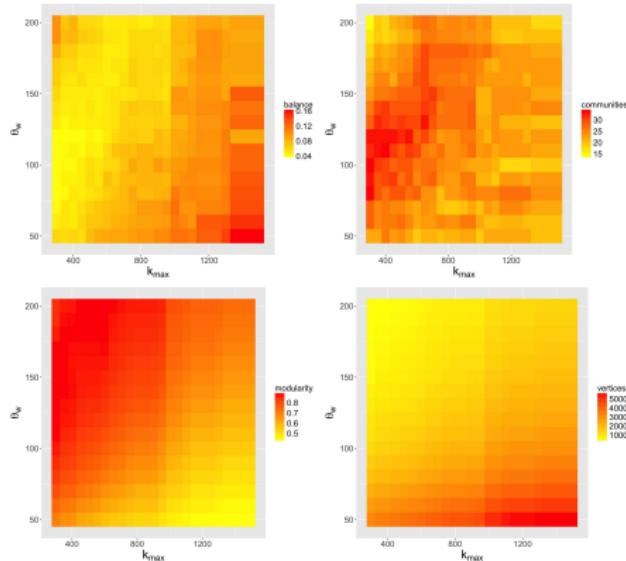
Previous work on patents and text-mining

- [Griliches, 1998]: patent as an economic indicator
- [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 (remains limited to specific fields and time windows)

Technical aspects

- Temporal complexity: $\mathcal{O}(N_P)$ for keyword extraction and co-occurrences (constant I_{max}^2); parallelization on a 60 cores server for a “reasonable” computation time.
- Memory complexity: co-occurrence matrices in $\mathcal{O}(K_W^2)$; necessitates around 600Go RAM when parallelized.
- Database management: MongoDB (nosql suited to big data).

An application of the method to a scientific corpus



*The method has been applied to a scientific corpus; adaptation of optimization procedures as the topological structure is different.
[Raimbault, 2017]*

Estimation of n-gram relevance

- Termhood-based selection of a fixed number of keywords
 $K_W = 100,000$ (*difficulty to select terms in general*) Definition
- Estimation on temporal moving windows $[t - T_w; t]$ (fixed to $T_w = 4y$ after sensitivity analysis)
- Filtering of network edge (parameter θ_w), with an additional exogenous control by technological class keyword concentration to filter nodes (parameter θ_c)
- Low sensibility to the length of the time window for the network structure Sensitivity analysis

Filtration measures

Back

1. *Unithood: n-gram specific frequency-based filtration (4 · K_W keywords)*

$$u_i = f_i \cdot \log(1 + l_i)$$

2. *Termhood: chi-square statistic for the uniformity of co-occurrences*

$$t_i = \sum_{j \neq i} \frac{(M_{ij} - \sum_k M_{ik} \sum_k M_{jk})^2}{\sum_k M_{ik} \sum_k M_{jk}}. \quad (1)$$

3. *Edge weight filtration with a threshold of $\theta_w = \theta_w^{(0)} \cdot N_P$*
4. *Node filtration with a threshold θ_c on technological class concentration given by*

$$c_{tech}(s) = \sum_{j=1}^{N^{(tec)}} \frac{k_j(s)^2}{(\sum_i k_i(s))^2}$$

Patent measures

Back

Diversity

$$D_i^{(z)} = 1 - \sum_{j=1}^{N^{(z)}} p_{ij}^2, \text{ with } z \in \{tec, sem\}$$

Originality and Generality

$$O_i^{(z)} = 1 - \sum_{j=1}^{N^{(z)}} \left(\frac{\sum_{i' \in I_i} p_{i'j}}{\sum_{k=1}^{N^{(z)}} \sum_{i' \in I_i} p_{i'k}} \right)^2 \quad \text{and} \quad G_i^{(z)} = 1 - \sum_{j=1}^{N^{(z)}} \left(\frac{\sum_{i' \in \tilde{I}_i} p_{i'j}}{\sum_{k=1}^{N^{(z)}} \sum_{i' \in \tilde{I}_i} p_{i'k}} \right)^2$$

Definition of modularities

Back

Directed simple modularity:

$$Q_d^{(z)} = \frac{1}{N_P} \sum_{1 \leq i, j \leq N_P} \left[A_{ij} - \frac{k_i^{in} k_j^{out}}{N_P} \right] \delta(c_i, c_j),$$

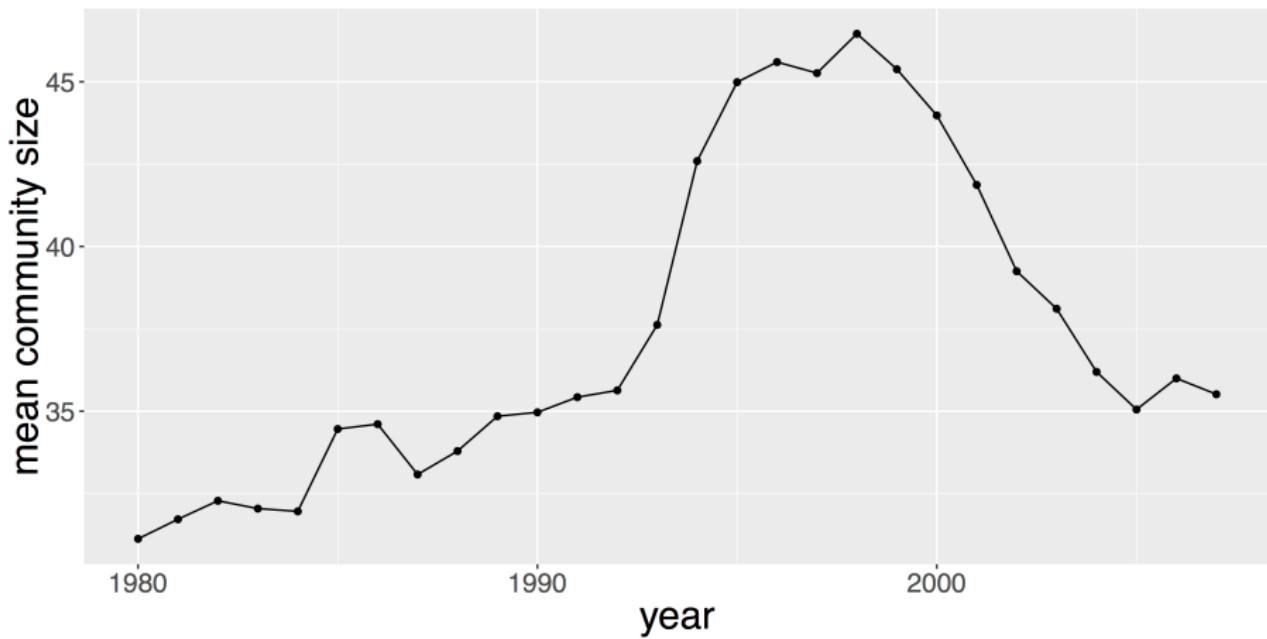
Multi-class modularity:

$$Q_{ov}^{(z)} = \frac{1}{N_P} \sum_{c=1}^{N^{(z)}} \sum_{1 \leq i, j \leq N_P} \left[F(p_{ic}, p_{jc}) A_{ij} - \frac{\beta_{i,c}^{out} k_i^{out} \beta_{j,c}^{in} k_j^{in}}{N_P} \right],$$

where

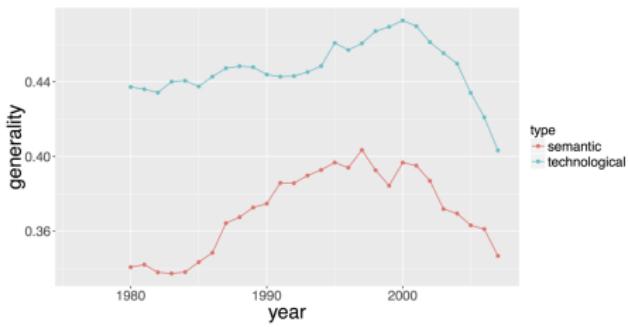
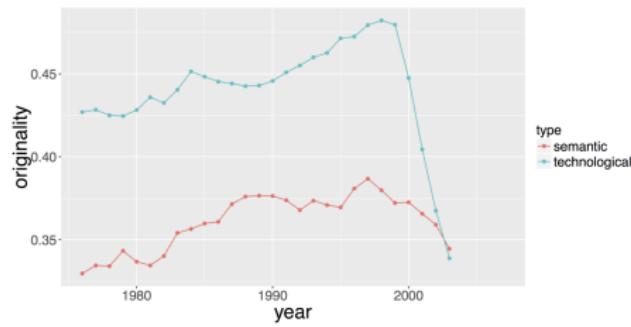
$$\beta_{i,c}^{out} = \frac{1}{N_P} \sum_j F(p_{ic}, p_{jc}) \text{ and } \beta_{j,c}^{in} = \frac{1}{N_P} \sum_i F(p_{ic}, p_{jc}).$$

Size of classes



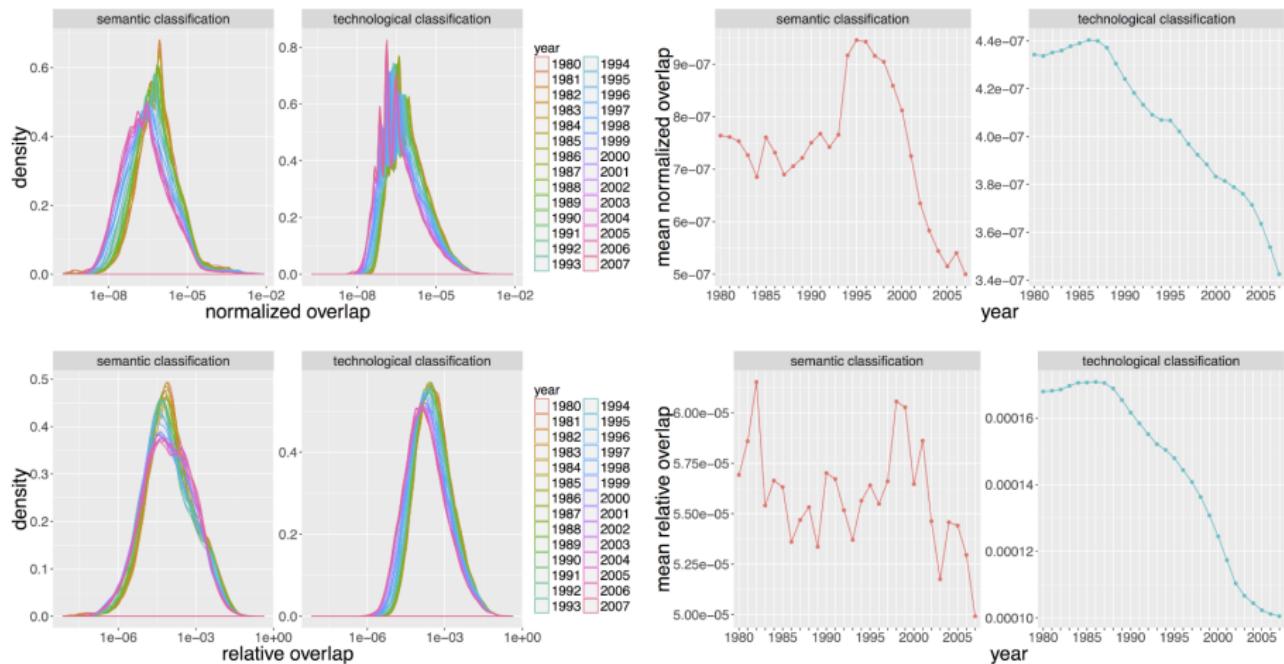
A peak in average class size around 1998

Originality and generality measures



Systematically lower originality and generality (citation-based measures) for the semantic classification (consistent with the higher modularity shown thereafter).

Interaction between classes: intra-classification overlaps



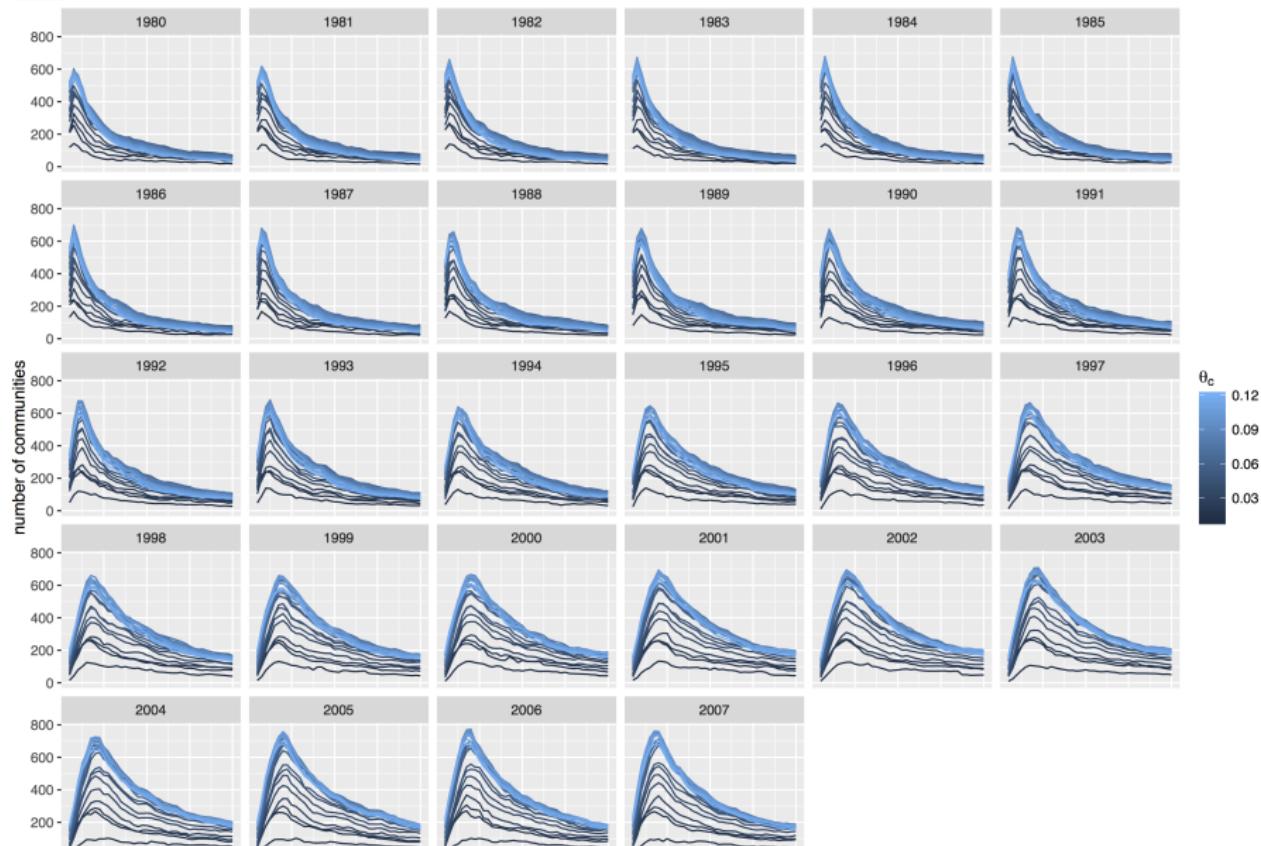
Increased technological specialization; qualitative regime shift confirmed in 1996 for the semantic classification

Stochastic Block Modeling to evaluate consistency of classifications

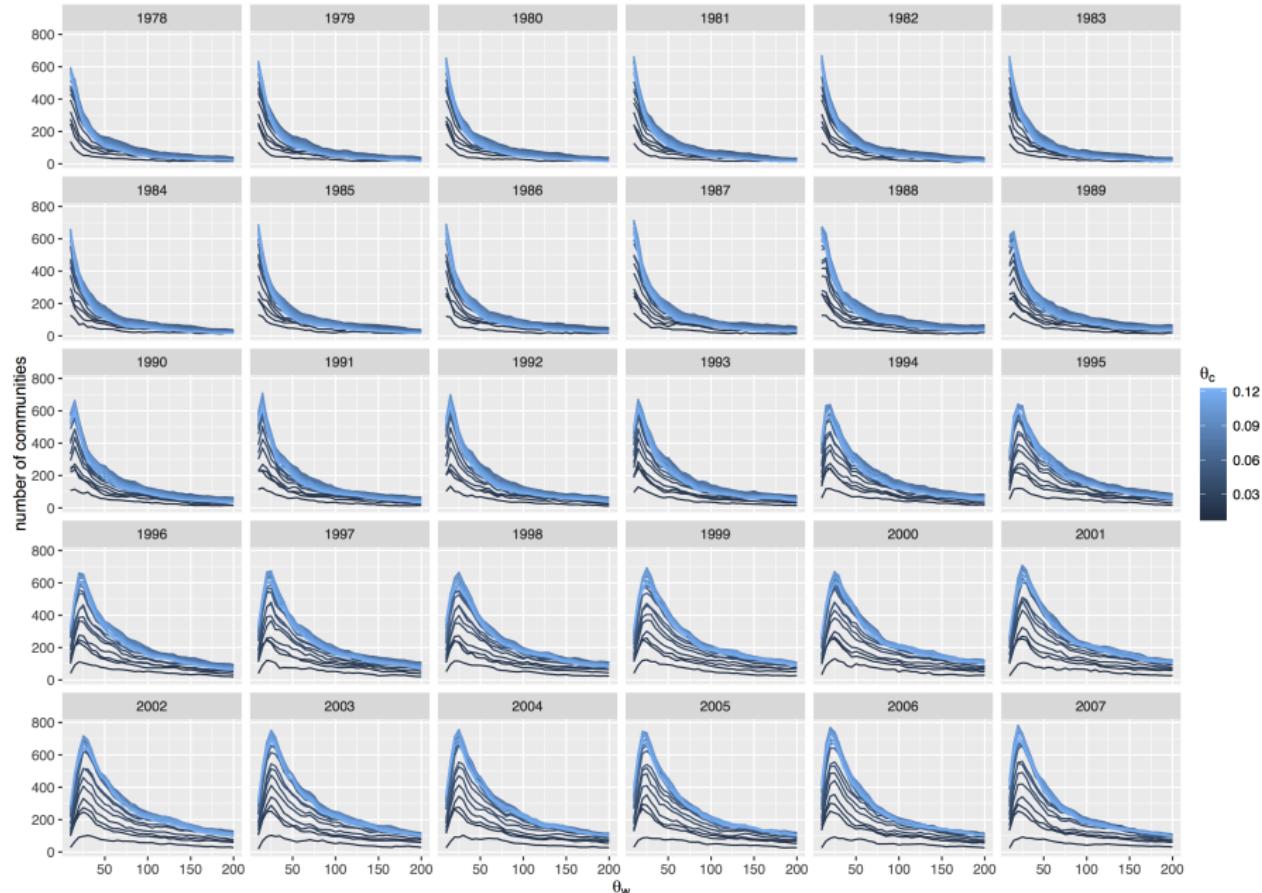
- We complete the analysis by developing a statistical model aiming at quantifying performance of both technological and semantic classification systems
- Intuitively, we look at within class citations proportion (for both technological and semantic approaches)
- Question: is the semantic classification better at predicting future within class citations?
→ short answer is **yes**.

Network sensitivity analysis

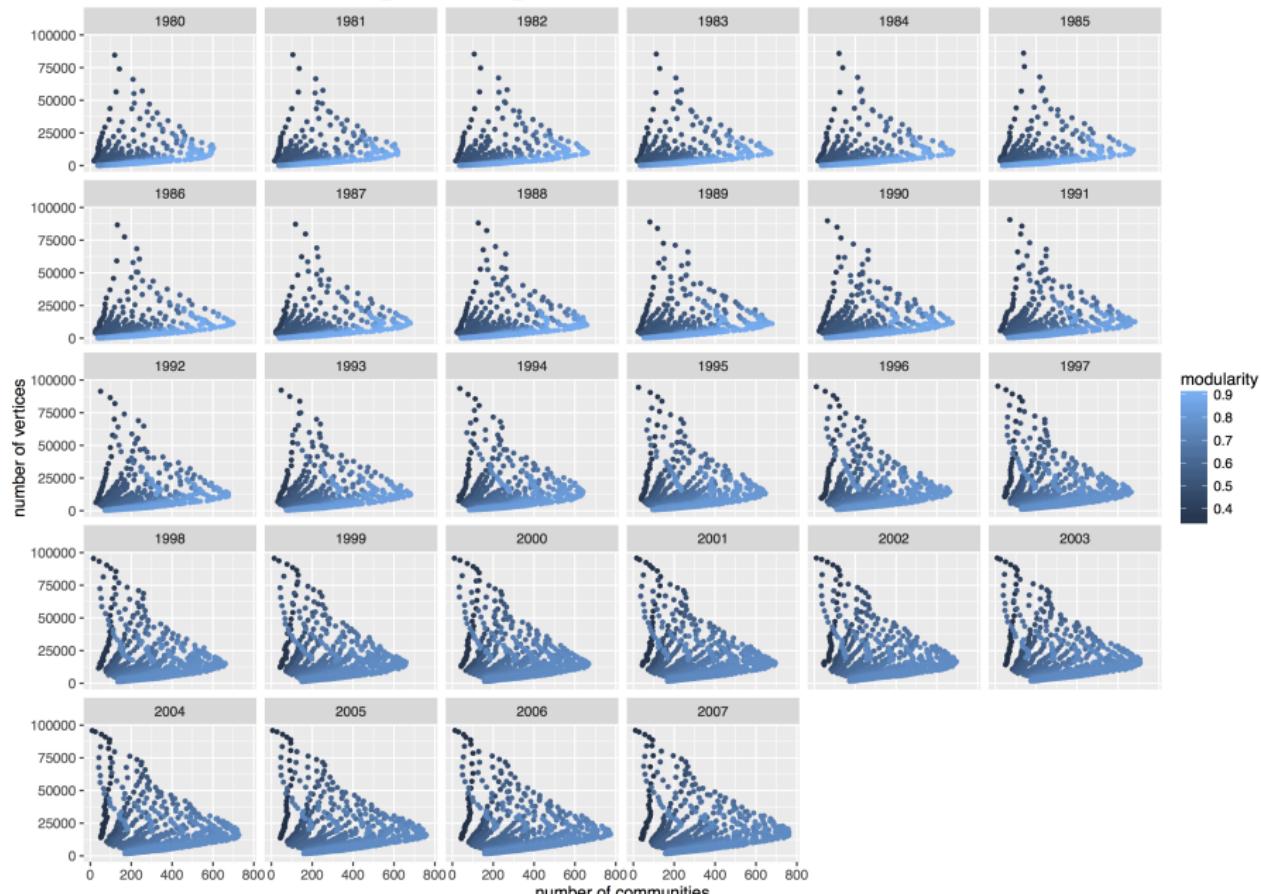
Back



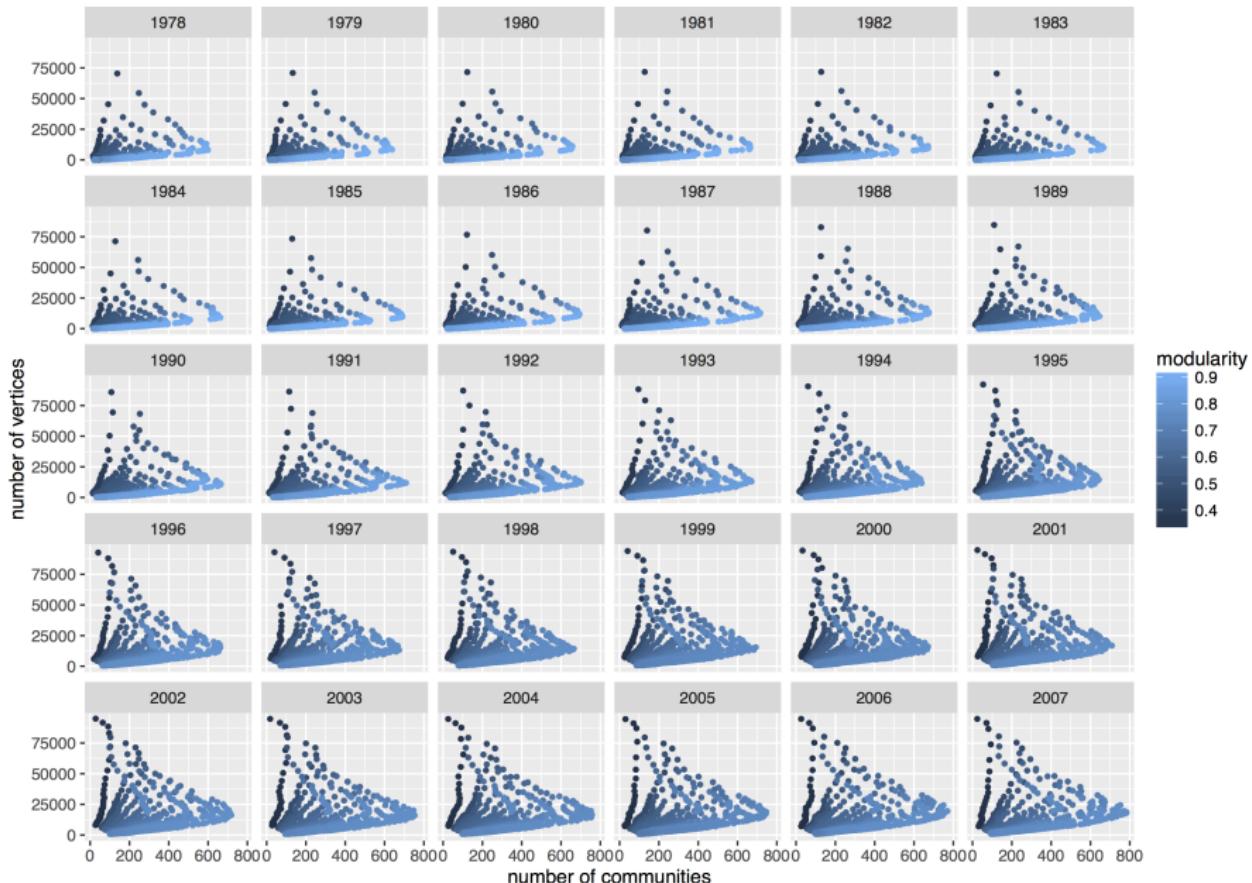
Network sensitivity analysis ($T_W = 2$)



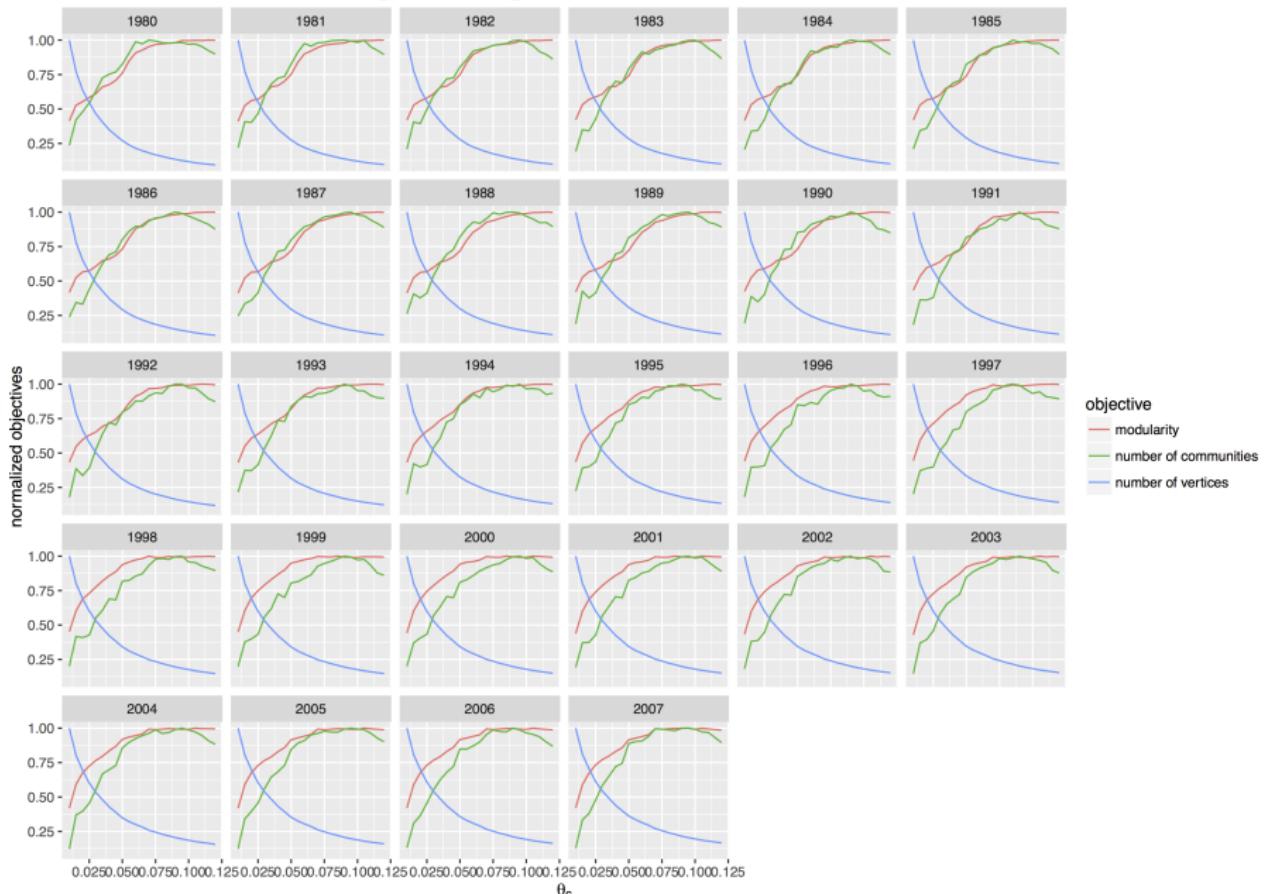
Network sensitivity analysis



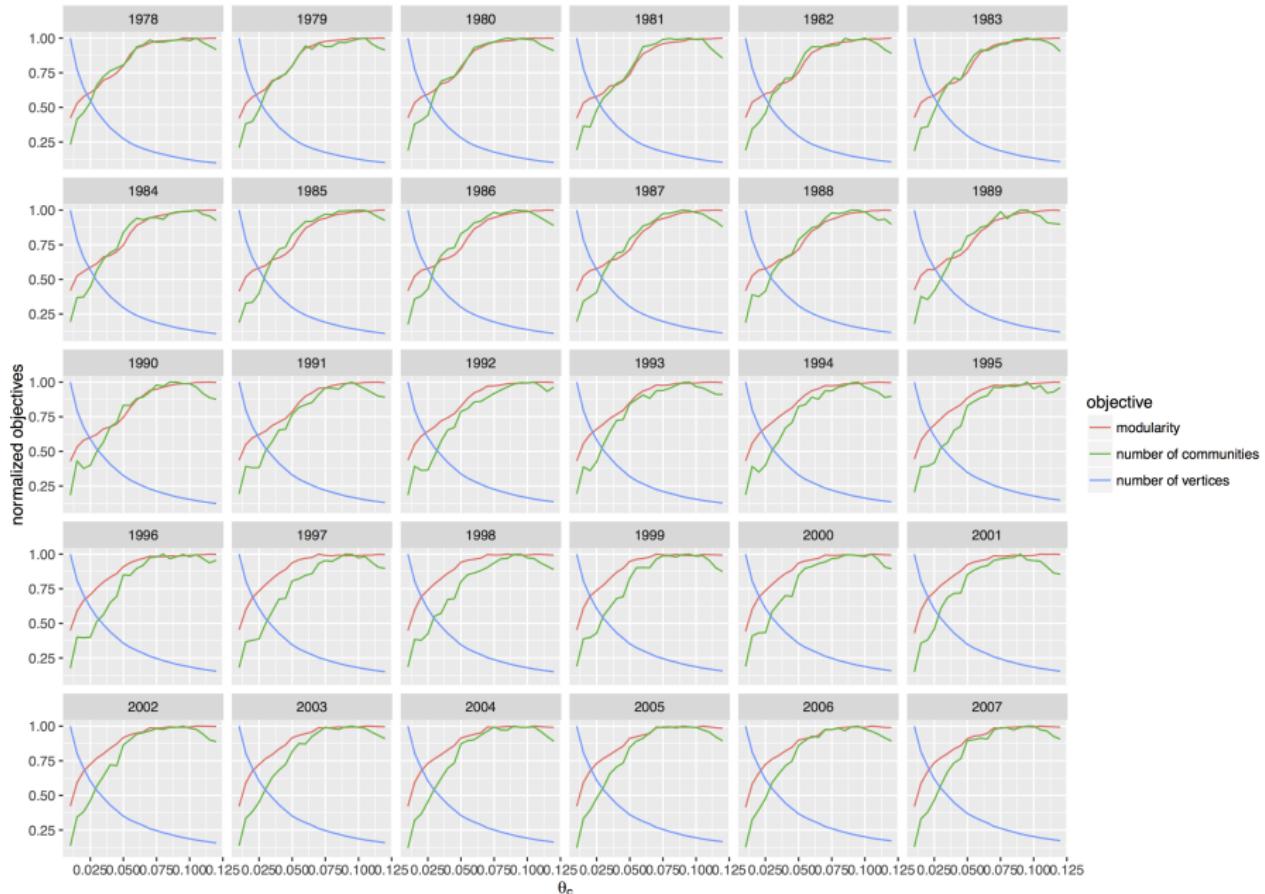
Network sensitivity analysis ($T_W = 2$)



Network sensitivity analysis



Network sensitivity analysis ($T_W = 2$)



References I

-  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.
-  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*.
-  Chavaliarias, D. and Cointet, J.-P. (2013). Phylogenetic patterns in science evolution—the rise and fall of scientific fields. *Plos One*, 8(2):e54847.

References II

-  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.
-  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, J. (2017).
Exploration of an interdisciplinary scientific landscape.
arXiv preprint arXiv:1712.00805.
-  Tseng, Y.-H., Lin, C.-J., and Lin, Y.-I. (2007).
Text mining techniques for patent analysis.
Information Processing & Management, 43(5):1216–1247.

References III

-  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).