PLOS ONE

Classifying Patents Based on their Semantic Content

--Manuscript Draft--

Manuscript Number:	PONE-D-16-51032R1
Article Type:	Research Article
Full Title:	Classifying Patents Based on their Semantic Content
Short Title:	Classifying Patents Based on their Semantic Content
Corresponding Author:	Juste Raimbault Universite Paris Diderot Paris, FRANCE
Keywords:	Technological Innovation; Patent; Classification; Text-mining; Semantic analysis
Abstract:	In this paper, we extend some usual techniques of classification resulting from a large-scale data-mining and network approach. This new technology, which in particular is designed to be suitable to big data, is used to construct an open consolidated database from raw data on 4 million patents taken from the US patent office from 1976 onward. To build the pattern network, not only do we look at each patent title, but we also examine their full abstract and extract the relevant keywords accordingly. We refer to this classification as semantic approach in contrast with the more common technological approach which consists in taking the topology when considering US Patent office technological classes. Moreover, we document that both approaches have highly different topological measures and strong statistical evidence that they feature a different model. This suggests that our method is a useful tool to extract endogenous information.
Order of Authors:	Antonin Bergeaud
	Yoann Potiron
	Juste Raimbault
Opposed Reviewers:	
Response to Reviewers:	Dear Editor, Thank you for considering our manuscript “Classifying Patents Based on their Semantic Content” for possible publication in PLOS ONE. We are also very grateful to your suggestions and comments. This will undoubtedly be of great value to the paper. We have read carefully your suggestions and comments, and have updated the paper accordingly. We provide you now the point-by-point response to the Editor and referees’ reports. We deal first with the Editor’s comments. 1. We made adjustments to fully meet PLOS ONE requirements. All figures were converted and assessed using the PACE tool. 2. "Authors used co-occurrence of keywords to construct a patent network. Is this a new way? Or at least a discussion of the advantages should be provided." → The use of co-occurrences to construct a semantic network has already been used, and is the best way to extract the endogenous semantic structure. We added a discussion on this point. 3. "Authors introduced a measure to correct the network topology. But how to properly determine the threshold? Why choose the value 0.06 ?" → Thank you for pointing out this issue. We have been more specific about it now, adding three sentences. Indeed, the explanation at the end of Section 3.3 was sloppy. 4. "The research background include complex network analysis, community detection and data analysis, some recent progress in these areas should be reviewed" → We included some of the references suggested in Section 3.5, as time series complex network analysis is indeed an interesting potential development.

	5. "All the parameters should be clearly explained." → Thank you for also pointing this out. We have been clearer on the definition of Kw, k and ti (Section 3.2), θ , θ and $\theta(0)$ (Section 3.3). We also provide you responses to Reviewer #3. Reviewers #1 and #2 did not ask anything specific but their comments were taken into account in the adjustments we made. 1. "Maybe a minor typo in Page 24 - "availability of these data" –> "this data" or "these datasets"". → We corrected accordingly. 2. "I also did not see a caption for the figures and it is quite hard to read the text in the figures due to their current small size. New readers are attracted to tables and figures, and thus it is useful to have descriptive captions - the current caption for Table 1 does not describe the variables being used (I know the text does it) - a brief description of theta being the likelihood estimates (or similar) would be useful." → We have made the according changes. All captions inside figures were magnified as large as possible to ensure readability. 3. "I like the authors' approach overall, but would also recommend adding some discussion on how a semantic approach enables information integration and reuse - possibly with how their dataset /ontology can be linked to others already existing in Linked Open Data. If such linking already exists, it should be shown - otherwise, this is a strong direction for future work." → This is indeed a very good suggestion and we add ideas of interesting potential developments by joining our database with existing open databases. We added accordingly a part to the discussion.
Additional Information:	
Question	**Response**
Financial Disclosure Please describe all sources of funding that have supported your work. **This information is required for submission and will be published with your article, should it be accepted.** A complete funding statement should do the following: Include **grant numbers and the URLs** of any funder's website. Use the full name, not acronyms, of funding institutions, and use initials to identify authors who received the funding. **Describe the role** of any sponsors or funders in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. If the funders had **no role** in any of the above, include this sentence at the end of your statement: "*The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.*" However, if the study was **unfunded**, please provide a statement that clearly indicates this, for example: "*The author(s) received no specific funding for this work.*"	Antonin Bergeaud received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors for this research. Yoann Potiron received private funding from Keio University. Juste Raimbault received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors for this research.

* typeset	
Competing Interests You are responsible for recognizing and disclosing on behalf of all authors any competing interest that could be perceived to bias their work, acknowledging all financial support and any other relevant financial or non-financial competing interests. Do any authors of this manuscript have competing interests (as described in the PLOS Policy on Declaration and Evaluation of Competing Interests)? **If yes**, please provide details about any and all competing interests in the box below. Your response should begin with this statement: *I have read the journal's policy and the authors of this manuscript have the following competing interests:* **If no** authors have any competing interests to declare, please enter this statement in the box: "*The authors have declared that no competing interests exist.*" * typeset	The authors have declared that no competing interests exist.
Ethics Statement You must provide an ethics statement if your study involved human participants, specimens or tissue samples, or vertebrate animals, embryos or tissues. All information entered here should **also be included in the Methods section** of your manuscript. Please write "N/A" if your study does not require an ethics statement. **Human Subject Research (involved human participants and/or tissue)** All research involving human participants must have been approved by the authors'	N/A

Institutional Review Board (IRB) or an equivalent committee, and all clinical investigation must have been conducted according to the principles expressed in the Declaration of Helsinki. Informed consent, written or oral, should also have been obtained from the participants. If no consent was given, the reason must be explained (e.g. the data were analyzed anonymously) and reported. The form of consent (written/oral), or reason for lack of consent, should be indicated in the Methods section of your manuscript. Please enter the name of the IRB or Ethics Committee that approved this study in the space below. Include the approval number and/or a statement indicating approval of this research. **Animal Research (involved vertebrate animals, embryos or tissues)** All animal work must have been conducted according to relevant national and international guidelines. If your study involved non-human primates, you must provide details regarding animal welfare and steps taken to ameliorate suffering; this is in accordance with the recommendations of the Weatherall report, "The use of non-human primates in research." The relevant guidelines followed and the committee that approved the study should be identified in the ethics statement. If anesthesia, euthanasia or any kind of animal sacrifice is part of the study, please include briefly in your statement which substances and/or methods were applied. Please enter the name of your Institutional Animal Care and Use Committee (IACUC) or other relevant ethics board, and indicate whether they approved this research or granted a formal waiver of ethical approval. Also include an approval number if one was obtained. **Field Permit** Please indicate the name of the institution	

or the relevant body that granted permission.	
Data Availability PLOS journals require authors to make all data underlying the findings described in their manuscript fully available, without restriction and from the time of publication, with only rare exceptions to address legal and ethical concerns (see the PLOS Data Policy and FAQ for further details). When submitting a manuscript, authors must provide a Data Availability Statement that describes where the data underlying their manuscript can be found. Your answers to the following constitute your statement about data availability and will be included with the article in the event of publication. **Please note that simply stating 'data available on request from the author' is not acceptable.** ***If, however, your data are only available upon request from the author(s), you must answer "No" to the first question below, and explain your exceptional situation in the text box provided.*** Do the authors confirm that all data underlying the findings described in their manuscript are fully available without restriction?	Yes - all data are fully available without restriction
Please describe where your data may be found, writing in full sentences. **Your answers should be entered into the box below and will be published in the form you provide them, if your manuscript is accepted.** If you are copying our sample text below, please ensure you replace any instances of XXX with the appropriate details. If your data are all contained within the paper and/or Supporting Information files, please state this in your answer below. For example, "All relevant data are within the paper and its Supporting Information files." If your data are held or will be held in a public repository, include URLs, accession numbers or DOIs. For example, "All XXX files are available from the XXX database (accession number(s) XXX, XXX)." If this information will only be available after acceptance, please indicate this by ticking the box below. If neither of these applies but you are able to provide details of access elsewhere, with or without limitations, please do so in	All data files are available from the Dataverse repository, at urls http://dx.doi.org/10.7910/DVN/BW3ACK and http://dx.doi.org/10.7910/DVN/ZULMOY. All source code and results are available on the git repository of the project at https://github.com/JusteRaimbault/PatentsMining.

the box below. For example: “Data are available from the XXX Institutional Data Access / Ethics Committee for researchers who meet the criteria for access to confidential data.” “Data are from the XXX study whose authors may be contacted at XXX.” * typeset	
Additional data availability information:	

Editors PLoS ONE December 26, 2016

Dear Editors,

We are pleased, with my coauthors Antonin Bergeaud (London School of Economics) and Yoann Potiron (Keio University), to submit an original research article entitled “Classifying Patents Based on their Semantic Content” for consideration for publication in PLoS ONE. This article introduce a novel insight into the study of technological innovation, by describing a new methodology to classify patents according to their endogenous semantic content and by deploying it on a large scale database of US Patents. It does not rely on previously published work, but a preliminary version of the methodology and results were presented as a main track talk at Conference of Complex Systems 2016, Amsterdam, under the title “Investigating Patterns Patterns of Technological Innovation”. We did not have any previous interaction with PLOS regarding this manuscript. This manuscript has not been published and is not under consideration for publication elsewhere. We have no conflicts of interest to disclose. We do not have any opposed reviewer. The authors believe that the following editors may be the most appropriate to handle this manuscript : César A Hidalgo (MIT), Tobias Preis (University of Warwick), Wolfgang Glanzel (Katholieke Universiteit Leuven). Thank you for your consideration.

Yours faithfully,
Juste Raimbault
Université Paris 7 - UMR CNRS 8504 Géographie-cités

Classifying Patents Based on their Semantic Content

Antonin Bergeaud[1,◕], Yoann Potiron[2,◕], Juste Raimbault[3,4,◕,*]

1 Paris School of Economics - EHESS and Bank of France, Paris, France

2 Faculty of Business and Commerce, Keio University, Tokyo, Japan

3 UMR CNRS 8504 Géographie-cités, Université Paris VII, Paris, France

4 UMR-T 9403 IFSTTAR LVMT, Ecole Nationale des Ponts et Chaussées, Champs-sur-Marne, France

◕These authors contributed equally to this work.

*** Corresponding Author**

Email : juste.raimbault@polytechnique.edu

Abstract

In this paper, we extend some usual techniques of classification resulting from a large-scale data-mining and network approach. This new technology, which in particular is designed to be suitable to big data, is used to construct an open consolidated database from raw data on 4 million patents taken from the US patent office from 1976 onward. To build the pattern network, not only do we look at each patent title, but we also examine their full abstract and extract the relevant keywords accordingly. We refer to this classification as *semantic approach* in contrast with the more common *technological approach* which consists in taking the topology when considering US Patent office technological classes. Moreover, we document that both approaches have highly different topological measures and strong statistical evidence that they feature a different model. This suggests that our method is a useful tool to extract endogenous information.

1 Introduction

Innovation and technological change have been described by many scholars as the main drivers of economic growth as in [1] and [2]. [3] advertised the use of patents as an economic indicator and as a good proxy for innovation. Subsequently, the easier availability of comprehensive databases on patent details and the increasing number of studies allowing a more efficient use of these data (e.g. [4]) have opened the way to a very wide range of analysis. Most of the statistics derived from the patent databases relied on a few key features: the identity of the inventor, the type and identity of the rights owner, the citations made by the patent to prior art and the technological classes assigned by the patent office post patent's content review. Combining this information is particularly relevant when trying to capture the diffusion of knowledge and the interaction between technological fields as studied in [5]. With methods such as citation dynamics modeling discussed in [6] or co-authorship networks analysis in [7], a large body of the literature such as [8] or [9] has studied patents citation network to understand processes driving technological innovation, diffusion and the birth of technological clusters. Finally, [10] look at the dynamics of citations from different classes to show that the laser/ink-jet printer technology resulted from the recombination of two different existing technologies.

Consequently, technological classification combined with other features of patents can be a valuable tool for researchers interested in studying technologies throughout history and to predict future innovations by looking at past knowledge and interaction across sectors and technologies. But it is also crucial for firms that face an ever changing demand structure and need to anticipate future technological trends and convergence (see, e.g., [11]) to adapt to the resulting increase in competition discussed in [12] and to maintain market share. Curiously, and in spite of the large number of studies that analyze interactions across technologies [13], little is known about the underlying "innovation network" (e.g. [14]).

In this monograph, we propose an alternative classification based on semantic network analysis from patent abstracts and explore the new information emerging from it. In contrast with the regular technological classification which results from the choice of the patent reviewer, semantic classification is carried automatically based on the content of the patent abstract. Although patent officers are experts in their fields, the relevance of the existing classification is limited by the fact that it is based on the state of technology at the time the patent was granted and cannot anticipate the birth of new fields.[1] In contrast we don't face this issue with the semantic approach. The semantic links can be clues of one technology taking inspiration from another and good predictors of future technology convergence (e.g. [15] study semantic similarities from the whole text of 326 US-patents on *phytosterols* and show that semantic analysis have a good predicting power of future technology convergence). One can for instance consider the case of the word *optic*. Until more recently, this word was often associated with technologies such as photography or eye surgery, while it is now almost exclusively used in a context of semi-transistor design and electro-optic. This semantic shift did not happen by chance but contains information on the fact that modern electronic extensively uses technologies that were initially developed in optic.

Previous research has already proposed to use semantic networks to study technological domains and detect novelty. [16] was one of the first to enhance this approach with the idea of visualizing keywords network illustrated on a small technological domain. The same approach can be used to help companies identifying the state of the art in their field and avoid patent infringement as in [17] and [18]. More closely related to our methodology, [19] develop a method based on patent semantic analysis of patent to vindicate the view that this approach outperform others in the monitoring of technology and in the identification of novelty innovation. Semantic analysis has already proven its efficiency in various fields, such as in technology studies (e.g. [20] and [21]) and in political science (e.g. [22]).

Building on such previous research, we make several contributions by fulfilling some shortcomings of existing studies, such as for example the use of frequency-selected single keywords. First of all, we develop and implement a novel fully-automatized methodology to classify patents according to their semantic abstract content, which is to the best of our knowledge the first of its type. This includes the following refinements for which details can be found in Section 3: (i) use of multi-stems as potential keywords; (ii) filtering of keywords based on a second-order (co-occurrences) relevance measure and on an external independent measure (technological dispersion); (iii) multi-objective

[1] To correct for this, the USPTO regularly make changes in its classification in order to adapt to technological change (for example, the "nanotechnology" class (977) was established in 2004 and retroactively to all relevant previously granted patents).

optimization of semantic network modularity and size. The use of all this techniques in the context of semantic classification is new and essential from a practical perspective.

Furthermore, most of the existing studies rely on a subsample of patent data, whereas we implement it on the full US Patent database from 1976 to 2013. This way, a general structure of technological innovation can be studied. We draw from this application promising qualitative stylized facts, such as a qualitative regime shift around the end of the 1990s, and a significant improvement of citation modularity for the semantic classification when comparing to the technological classification. These thematic conclusions validate our method as a useful tool to extract endogenous information, in a complementary way to the technological classification.

Finally, the statistical model introduced in Section 4.4 seems to indicate that patents tend to cite more similar patents in the semantic network when fitted to data. In particular, this propensity is shown to be significantly bigger than the corresponding propensity for technological classes, and this seems to be consistent over time. On the account of this information, we believe that patent officers could benefit very much from looking at the semantic network when considering potential citation candidates of a patent in review.

The paper is organized as follows. Section 2 presents the patent data, the existing classification and provide details about the data collection process. Section 3 explains the construction of the semantic classes. Section 4 tests their relevance by providing exploratory results. Finally, section 5 discusses potential further developments and conclude. More details, including robustness checking, figures and technical derivations can be found in Supporting Information.

2 Background

In our analysis, we will consider all utility patents granted in the United States Patent and Trademark Office (USPTO) from 1976 to 2013. A clearer definition of utility patent is given in Supporting Information. Also, additional information on how to correctly exploit patent data can be found in [4] and [23].

2.1 An existing classification: the USPC system

Each USPTO patent is associated with a non-empty set of technological classes and subclasses. There are currently around 440 classes and over 150,000 subclasses constituting the United State Patent Classification (USPC) system. While a technological class corresponds to the technological field covered by the patent, a subclass stands for a specific technology or method used in this invention. A patent can have multiple technological classes, on average in our data a patent has 1.8 different classes and 3.9 pairs of class/subclass. At this stage, two features of this system are worth mentioning: (i) classes and subclasses are not chosen by the inventors of the patent but by the examiner during the granting process based on the content of the patent; (ii) the classification has evolved in time and continues to change in order to adapt to new technologies by creating or editing classes. When a change occurs, the USPTO reviews all the previous patents so as to create a consistent classification.

2.2 A bibliographical network between patents: citations

As with scientific publications, patents must give reference to all the previous patents which correspond to related prior art. They therefore indicate the past knowledge which relates to the patented invention. Yet, contrary to scientific citations, they also have an important legal role as they are used to delimit the scope of the property rights awarded by the patent. One can consult [24] for more details about this. Failing to refer to prior art can lead to the invalidation of the patent (e.g. [25]). Another crucial difference is that the majority of the citations are actually chosen by the examiners and not by the inventors themselves. From the USPTO, we gather information of all citations made by each patent (backward citations) and all citations received by each patent as of the end of 2013 (forward citations). We can thus build a complete network of citations that we will use later on in the analysis.

Turning to the structure of the lag between the citing and the cited patent in terms of application date, we see that the mean of this lag is 8.5 years and the median is 7 years. This distribution is highly skewed, the 95^{th} percentile is 21 years. We also report 164,000 citations with a negative time lag. This is due to the fact that some citations can be added during the examination process and some patents require more time to be granted than others.

In what follows, we choose to restrict attention to pairs of citations with a lag no larger than 5 years. We impose this restriction for two reasons. First, the number of citations received peaks 4-5 years after application. Second, the structure of the citation lag is necessarily biased by the truncation of our sample: the more recent patents mechanically receive less citations than the older ones. As we are restricting to citations received no later than 5 years after the application date, this effect will only affect patents with an application date after 2007.

2.3 Data collection and basic description

Each patent contains an abstract and a core text which describe the invention.[2] Although including the full core texts would be natural and probably very useful in a systematic text-mining approach as done in [26], they are too long to be included and thus we consider only the abstracts for the analysis. Indeed, the semantic analysis counts more than 4 million patents, with corresponding abstracts with an average length of 120.8 words (and a standard deviation of 62.4), a size that is already challenging in terms of computational burden and data size. In addition, abstracts are aimed at synthesizing purpose and content of patents and must therefore be a relevant object of study (see [27]). The USPTO defines a guidance stating that an abstract should be "a summary of the disclosure as contained in the description, the claims, and any drawings; the summary shall indicate the technical field to which the invention pertains and shall be drafted in a way which allows the clear understanding of the technical problem, the gist of the solution of that problem through the invention, and the principal use or uses of the invention" (PCT Rule 8).

We construct from raw data a unified database. Data is collected from USPTO patent redbook bulk downloads, that provides as raw data (specific `dat` or `xml` formats) full patent information, starting from 1976. Detailed procedure of data collection, parsing and consolidation are available in Supporting Information. The latest dump of the database in `Mongodb` format is available at `http://dx.doi.org/10.7910/DVN/BW3ACK`. Collection and homogenization of the database into a directly usable database with

[2] To see what a patent looks like in practice, one can refer to the USPTO patent full-text database `http://patft.uspto.gov/netahtml/PTO/index.html` or to Google patent which publishes USPTO patents in *pdf* format at `https://patents.google.com`.

basic information and abstracts was an important task as USPTO raw data formats are involved and change frequently.

We count 4,666,365 utility patents with an abstract granted from 1976 to 2013.[3] The number of patents granted each year increases from around 70,000 in 1976 to about 278,000 in 2013. When distributed by the year of application, the picture is slightly different. The number of patents steadily increase from 1976 to 2000 and remains constant around 200,000 per year from 2000 to 2007. Restricting our sample to patent with application date ranging from 1976 to 2007, we are left with 3,949,615 patents. These patents cite 38,756,292 other patents with the empirical lag distribution that has been extensively analyzed in [4]. Conditioned on being cited at least once, a patent receives on average 13.5 citations within a five-year window. 270,877 patents receive no citation during the next five years following application, 10% of patents receive only one citation and 1% of them receive more than 100 citations. A within class citation is defined as a citation between two patents sharing at least one common technological class. Following this definition, 84% of the citations are within class citations. 14% of the citations are between two patents that share the exact same set of technological classes.

2.4 Towards a Complementary Classification

Potentialities of text-mining techniques as an alternative way to analyze and classify patents are documented in [26]. The author's main argument, in support of an automatic classification tool for patent, is to reduce the considerable amount of human effort needed to classify all the applications. The work conducted in the field of natural language processing and/or text analysis has been developed in order to improve search performance in patent databases, build technology map or investigate the potential infringement risks prior to developing a new technology (see [28] for a review). Text-mining of patent documents is also widely used as a tool to build networks which carry additional information to the simplistic bibliographic connections model as argued in [16]. As far as the authors know, the use of text-mining as a way to build a global classification of patents remains however largely unexplored. One notable exception

[3] A very small number of patents have a missing abstract, these are patents that have been withdrawn and we do not consider them in the analysis.

can be found in [15] where semantic-based classification is shown to outperform the standard classification in predicting the convergence of technologies even in small samples. Semantic analysis reveals itself to be more flexible and more quickly adaptable to the apparition of new clusters of technologies. Indeed, as argued in [15], before two distinct technologies start to clearly converge, one should expect similar words to be used in patents from both technologies.

Finally, a semantic classification where patents are gathered based on the fact that they share similar significant keywords has the advantage of including a network feature that cannot be found in the USPC case, namely that each patent is associated with a vector of probability to belong to each of the semantic classes (more details on this feature can be found in Section 3.4). Using co-occurrence of keywords, it is then possible to construct a network of patents and to study the influence of some key topological features. As reviewed previously, the use of co-occurrences is the usual way to construct a semantic network. Other hybrid technique such as bipartite semantic/authors networks, do not have the nice feature of relying solely on endogenous semantic information contained in data.

3 Semantic Classification Construction

In this section, we describe methods and empirical analysis leading to the construction of semantic network and the corresponding classification.

3.1 Keywords extraction

Let $\mathcal{P}$ be the set of patents, we first assign to a patent $p \in \mathcal{P}$ a set of potentially significant keywords $K(p)$ from its text $\mathcal{A}(p)$ (that corresponds to the concatenation of its own title and abstract). $K(p)$ are extracted through a similar procedure as the one detailed in [29]:

1. Text parsing and Tokenization: we transform raw texts into a set of words and sentences, reading it (parsing) and splitting it into elementary entities (words organized in sentences).

2. Part-of-speech tagging: attribution of a grammatical function to each of the tokens

defined previously.

3. Stem extraction: families of words are generally derived from a unique root called stem (for example `compute`, `computer`, `computation` all yield the same stem `comput`) that we extract from tokens. At this point the abstract text is reduced to a set of stems and their grammatical functions.

4. Multi-stems construction: these are the basic semantic units used in further analysis. They are constructed as groups of successive stems in a sentence which satisfies a simple grammatical function rule. The length of the group is between 1 and 3 and its elements are either nouns, attributive verbs or adjectives. We choose to extract the semantics from such nominal groups in view of the technical nature of texts, which is not likely to contain subtle nuances in combinations of verbs and nominal groups.

Text processing operations are implemented in `python` in order to use built-in functions `nltk` library [30] for most of above operations. This library supports most of state-of-the-art natural language processing operations.[4]

3.2 Keywords relevance estimation

Relevance definition Following the heuristic in [29], we estimate relevance score in order to filter multi-stem. The choice of the total number of keywords to be extracted, which we shall denote K_w , is important, too small a value would yield similar network structures but including less information whereas very large values tend to include too many irrelevant keywords. We choose to set this parameter to $K_w = 100,000$. We first consider the filtration of $k \cdot K_w$ (with $k = 4$) to keep a large set of potential keywords but still have a reasonable number of co-occurrences to be computed. This step has only very marginal effects on the nature of the final keywords but is necessary for computational purposes. The filtration is done on the *unithood* u_i, defined for keyword i as $u_i = f_i \cdot \log(1 + l_i)$ where f_i is the multi-stem's number of apparitions over the whole corpus and l_i its length in words. A second filtration of K_w keywords is done on the *termhood* t_i, where the formal definition can be found in (1) . It is computed as a

[4] Source code is openly available on the repository of the project: `https://github.com/JusteRaimbault/PatentsMining`

chi-squared score on the distribution of the stem's co-occurrences and then compared to a uniform distribution within the whole corpus. Intuitively, uniformly distributed terms will be identified as plain language and they are thus not relevant for the classification. More precisely, we compute the co-occurrence matrix (M_{ij}), where M_{ij} is defined as the number of patents where stems i and j appear together. The *termhood* score t_i is defined as

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

Moving window estimation The previous scores are estimated on a moving window with fixed time length following the idea that the present relevance is given by the most recent context and thus that the influence vanishes when going further into the past. Consequently, the co-occurrence matrix is chosen to be constructed at year t restricting to patent which applied during the time window $\left[t - T_0; t\right]$. Note that the causal property of the window is crucial as the future cannot play any role in the current state of keywords and patents. This way, we will obtain semantic classes which are exploitable on a T_0 time span. For example, this enables us to compute the modularity of classes in the citation network as in section 4.3. In the following, we take $T_0 = 4$ (which corresponds to a five year window) consistently with the choice of maximum time lag for citations made in Section 2.2. Accordingly, the sensitivity analysis for $T_0 = 2$ can be found in Appendix S4 Text : Network Sensitivity Analysis.

3.3 Construction of the semantic network

We keep the set of most relevant keywords $\mathcal{K}_W$ and obtain their co-occurrence matrix as defined in Section 3.2. This matrix can be directly interpreted as the weighted adjacency matrix of the semantic network. At this stage, the topology of raw networks does not allow the extraction of clear communities. This is partly due to the presence of hubs that correspond to frequent terms common to many fields (e.g. `method`, `apparat`) which are wrongly filtered as relevant. We therefore introduce an additional measure to correct the network topology: the concentration of keywords across technological classes, defined as:

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

where $k_j(s)$ is the number of occurrences of the sth keyword in each of the jth technological class taken from one of the $N^{(tec)}$ USPC classes. The higher c_{tech}, the more specific to a technological class the node is. For example, the terms `semiconductor` is widely used in electronics and does not contain any significant information in this field. We use a threshold parameter, defined as θ_c, and keep nodes with $c_{tech}(s) > \theta_c$. Likewise, edges with low weights correspond to rare co-occurrences and are considered to be noise. To account for this we define the threshold parameter for edges θ_w, and we filter edges with a weight below θ_w , following the rationale that two keywords are not linked "by chance" if they appear simultaneously a minimal number of time. To control for size effect, we normalize by taking $\theta_w = \theta_w^{(0)} \cdot N_P$ where N_P is the number of patents in the corpus ($N_P = |\mathcal{P}|$). $\theta_w^{(0)}$ is thus a varying parameter interpreted as a noise threshold *per patent*. Communities are then extracted using a standard modularity maximization procedure as described in [31] to which we add the two constraints captured by θ_w and θ_c, namely that edges must have a weight greater than θ_w and nodes a concentration greater than θ_c. At this stage, both parameters θ_c and $\theta_w^{(0)}$ are unconstrained and their choice is not straightforward. Indeed, many optimization objectives are possible, such as the modularity, network size or number of communities. We find that modularity is maximized at a roughly stable value of θ_w across different θ_c for each year, corresponding to a stable $\theta_w^{(0)}$ across years, which leads us to choose $\theta_w^{(0)} = 4.1 \cdot 10^{-5}$. Then for the choice of θ_c, different candidates points lie on a Pareto front for the bi-objective optimization on number of communities and network size. There is a priori no reason to choose any specific point among the different optimums. Consequently, we have tried the analysis with all the candidate values for θ_c and found that the results are the most reasonable when taking $\theta_c = 0.06$ (see Fig. 1).

3.4 Characteristics of Semantic Classes

For each year t, we define as $N_t^{(sem)}$ the number of semantic classes which have been computed by clustering keywords from patents appeared during the period $\left[t - T_0, t\right]$ (we recall that we have chosen $T_0 = 4$). Each semantic class $k = 1, \cdots, N_t^{(sem)}$ is

Figure 1. Sensitivity analysis of network community structure to filtering parameters. We consider a specific window 2000-2004 and the obtained plots are typical. *(Left panel)* We plot the number of communities as a function of the edge threshold parameter θ_w for different values of the node threshold parameter θ_c. The maximum is roughly stable across θ_c (dashed red line). *(Right panel)* To choose θ_c, we do a Pareto optimization on communities and network size: the compromise point (red overline) on the Pareto front (purple overline: possible choices after having fixed $\theta_w^{(0)}$; blue level gives modularity) corresponds to $\theta_c = 0.06$.

Figure 2. An example of semantic network visualization. We show the network obtained for the window 2000-2004, with parameters $\theta_c = 0.06$ and $\theta_w = \theta_w^{(0)} \cdot N_P = 4.5e^{-5} \cdot 9.1e^5$. The corresponding file in a vector format (`.svg`), that can be zoomed and explored, is available as Supplementary Material.

characterized by a set of keywords $K(k,t)$ which is a subset of $\mathcal{K}_W$ selected as described in Section 3.1 to Section 3.3. The cardinal of $K(k,t)$ distribution across each semantic class k is highly skewed with a few semantic classes containing over $1,000$ keywords, most of them with roughly the same number of keywords. In contrast, there are also many semantic classes with only two keywords. There are around 30 keywords by semantic class on average and the median is 2 for any t. Fig. 3 shows that the average number of keywords is relatively stable from 1976 to 1992 and then picks around 1996 prior to going down.

Title of semantic classes USPC technological classes are defined by a title and a highly accurate definition which help retrieve patents easily. The title can be a single word (e.g.: class 101: "Printing") or more complex (e.g.: class 218: "High-voltage switches with arc preventing or extinguishing devices"). As our goal is to release a comprehensive database in which each patent is associated with a set of semantic classes, it is necessary to give an insight on what these classes represent by associating a short description or a title as in [26]. In our case, such description is taken as a subset of keywords taken from $K(k,t)$. For the vast majority of semantic classes that have less than 5 keywords, we decide to keep all of theses keywords as a description. For the remaining classes which feature around 50 keywords on average, we rely on the topological properties of the semantic network. [32] suggest to retain only the most frequently used terms in $K(k,t)$. Another possibility is to select 5 keywords based on

Figure 3. This figure plots the average number of keywords by semantic class for each time window $[t-4;t]$ from $t = 1980$ to $t = 2007$.

their network centrality with the idea that very central keywords are the best candidates to describe the overall idea captured by a community. For example, the largest semantic class in 2003-2007 is characterized by the keywords: `Support Packet; Tree Network; Network Wide; Voic Stream; Code Symbol Reader`.

Size of technological and semantic classes We consider a specific window of observations (for example 2000-2004), and we define Z the number of patents which appeared during that time window. For each patent $i = 1, \cdots, Z$ we associate a vector of probability where each component $p_{ij}^{(sem)} \in \big[0, 1\big]$, with $j = 1, \cdots, N(sem)$ and where[5]

$$\sum_{j=1}^{N^{(sem)}} p_{ij}^{(sem)} = 1.$$

On average across all time windows, a patent is associated to 1.8 semantic classes with a positive probability. Next we define the size of a semantic class as

$$S_j^{(sem)} = \sum_{i=1}^{Z} p_{ij}^{(sem)}.$$

Correspondingly, we aim to provide a consistent definition for technological classes. For that purpose, we follow the so-called "fractional count" method, which was introduced by the USPTO and consists in dividing equally the patents between all the classes they belong to. Formally, we define the number of technological classes as $N^{(tec)}$ (which is not time dependent contrary to the semantic case) and for $j = 1, \cdots, N^{(tec)}$ the corresponding matrix of probability is defined as

$$p_{ij}^{(tec)} = \frac{B_{ij}}{\sum_{k=1}^{N^{(tec)}} B_{ik}},$$

where B_{ij} equals 1 if the ith patent belongs to the jth technological class and 0 if not. When there is no room for confusion, we will drop the exponent part and write only p_{ij} when referring to either the technological or semantic matrix. Empirically, we find that both classes exhibit a similar hierarchical structure in the sense of a power-law type of distribution of class sizes as shown in Fig. 4. This feature is important, it suggests that

[5]When there is no room for confusion, we drop the subscript t in $N_t^{(sem)}$.

Figure 4. Sizes of classes. Yearly from $t = 1980$ to $t = 2007$, we plot the size of semantic classes (left-side) and technological classes (right-side) for the corresponding time window $[t-4, t]$, from the biggest to the smallest. The formal definition of size can be found in Section 3.4. Each color corresponds to one specific year. Yearly semantic classes and technological classes present a similar hierarchical structure which confirms the comparability of the two classifications. This feature is crucial for the statistical analysis in Section 4.4. Over time, curves are translated and levels of hierarchy stays roughly constant.

a classification based on the text content of patents has some separating power in the sense that it does not divide up all the patents in one or two communities.

3.5 Potential Refinements of the Method

Our semantic classification method could be refined by combining it with other techniques such as Latent Dirichlet Allocation which is a widely used topic detection method (e.g. [33]), already used on patent data as in [34] where it provides a measure of idea novelty and the counter-intuitive stylized facts that breakthrough invention are likely to come out of local search in a field rather than distant technological recombination. Using this approach should first help further evaluate the robustness of our qualitative conclusions (external validation). Also, depending on the level of orthogonality with our classification, it can potentially bring an additional feature to characterize patents, in the spirit of multi-modeling techniques where neighbor models are combined to take advantage of each point of view on a system.

Our use of network analysis can also be extended using newly developed techniques of hyper-network analysis. Indeed, patents and keywords can for example be nodes of a bipartite network, or patents be links of an hyper-network, in the sense of multiple layers with different classification links and citation links. The combination of citation network modeling by Stochastic Block Modeling with topic modeling was studied for scientific papers by [35] , outperforming previous link prediction algorithms. [36] provide a method to compare macroscopic structures of the different layers in a multilayer network that could be applied as a refinement of the overlap, modularity and statistical modeling studied in this paper. Furthermore, is has recently been shown that measures of multilayer network projections induce a significant loss of information compared to the generalized corresponding measure [37], which confirms the relevance of such development that we left for further research.

An other potential research development would be to further exploit the temporal structure of our dataset. Indeed, large progress have recently been made in complex network analysis of time-series data (see [38] for a review). For example, [39] develops a method to construct multiscale network from time series, which could in our case be a solution to identify structures in patents trajectories at different levels, and be an alternative to the single scale modularity analysis we use.

4 Results

In this section, we present some key features of our resulting semantic classification showing both complementary and differences with the technological classification. We first present several measures derived from this semantic classification at the patent level: Diversity, Originality, Generality (Section 4.1) and Overlapping (Section 4.2). We then show that the two classifications show highly different topological measures and strong statistical evidence that they feature a different model (Sections 4.3 and 4.4).

4.1 Patent Level Measures

Given a classification system (technological or semantic classes), and the associated probabilities p_{ij} for each patent i to belong to class j (that were defined in Section 3.4), one can define a patent-level diversity measure as one minus the Herfindhal concentration index on p_{ij} by

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

We show in Fig. 5 the distribution over time of semantic and technological diversity with the corresponding mean time-series. This is carried with two different settings, namely including/not including patents with zero diversity (i.e. single class patents). We call other patents “complicated patents” in the following. First of all, the presence of mass in small probabilities for semantic but not technological diversity confirms that the semantic classification contains patent spread over a larger number of classes. More interestingly, a general decrease of diversity for complicated patents, both for semantic and technological classification systems, can be interpreted as an increase in invention

Figure 5. Patent level diversities. Distributions of diversities (Left column) and corresponding mean time-series (Right column) for $t = 1980$ to $t = 2007$ (with the corresponding time window $[t-4, t]$). The first row includes all classified patents, whereas the second row includes only patents with more than one class (i.e. patents with diversity greater than 0).

specialization. This is a well-known stylized fact as documented in [40]. Furthermore, a qualitative regime shift on semantic classification occurs around 1996. This can be seen whether or not we include patents with zero diversity. The diversity of complicated patents stabilizes after a constant decrease, and the overall diversity begins to strongly decrease. This means that on the one hand the number of single class patents begins to increase and on the other hand complicated patents do not change in diversity. It can be interpreted as a change in the regime of specialization, the new regime being caused by more single-class patents.

More commonly used in the literature are the measures of originality and generality. These measures follow the same idea than the above-defined diversity in quantifying the diversity of classes (whether technological or semantic) associated with a patent. But instead of looking at the patent's classes, they consider the classes of the patents that are cited or citing. Formally, the originality O_i and the generality G_i of a patent i are defined as

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

where $z \in \{tec, sem\}$, I_i denotes the set of patents that are cited by the ith patent within a five year window (i.e. if the ith patent appears at year t, then we consider patents on $[t - T_0, t]$) when considering the originality and $\tilde{I}_i$ the set of patents that cite patent i after less than five years (i.e. we consider patents on $[t, t + T_0]$) in the case of generality. Note that the measure of generality is forward looking in the sense that $G_i^{(z)}$ used information that will only be available 5 years after patent applications. Both measures are lower on average based on semantic classification than on technological classification. Fig. 6 plots the mean value of $O_i^{(sem)}$, $O_i^{(tec)}$, $G_i^{(sem)}$ and $G_i^{(tec)}$.

Figure 6. Patent level originality (left hand side) and **generality** (right hand side) for $t = 1980$ to $t = 2007$ (with the corresponding time window $[t-4, t]$) as defined in subsection 4.1.

4.2 Classes overlaps

A proximity measure between two classes can be defined by their overlap in terms of patents. Such measures could for example be used to construct a metrics between semantic classes. Intuitively, highly overlapping classes are very close in terms of technological content and one can use them to measure distance between two firms in terms of technology as done in [41]. Formally, recalling the definition of (p_{ij}) as the probability for the ith patent to belong to the jth class and N_P as the number of patents it writes

$$Overlap_{jk} = \frac{1}{N_P} \cdot \sum_{i=1}^{N_P} p_{ij} p_{ik}. \tag{2}$$

The overlap is normalized by patent count to account for the effect of corpus size: by convention, we assume the overlap to be maximal when there is only one class in the corpus. A corresponding relative overlap is computed as a set similarity measure in the number of patents common to two classes A and B, given by $o(A, B) = 2 \cdot \frac{|A \cap B|}{|A|+|B|}$.

Intra-classification overlaps The study of distributions of overlaps inside each classification, i.e. between technological classes and between semantic classes separately, reveals the structural difference between the two classification methods, suggesting their complementary nature. Their evolution in time can furthermore give insights into trends of specialization. We show in Fig. 7 distributions and mean time-series of overlaps for the two classifications. The technological classification globally always follow a decreasing trend, corresponding to more and more isolated classes, i.e. specialized inventions, confirming the stylized fact obtained in previous subsection. For semantic classes, the dynamic is somehow more intriguing and supports the story of a qualitative regime shift suggested before. Although globally decreasing as technological overlap, normalized (resp. relative) mean overlap exhibits a peak (clearer for normalized overlap) culminating in 1996 (resp. 1999). Looking at normalized overlaps, classification structure was somewhat stable until 1990, then strongly increased to peak in 1996 and then decrease at a similar pace up to now. Technologies began to share more and more until a breakpoint when increasing isolation became the rule again. An evolutionary perspective on technological innovation [42] could shed light on possible interpretations of this regime shift: as species evolve, the fitness landscape first would have been locally favorable to cross-insemination, until each fitness reaches a threshold above which auto-specialization becomes the optimal path. It is very comparable to the establishment of an ecological niche [43], the strong interdependency originating here during the mutual insemination resulting in a highly path-dependent final situation.

Figure 7. Intra-Classification overlaps. (Left column) Distribution of overlaps O_{ij} for all $i \neq j$ (zero values are removed because of the log-scale). Right column) Corresponding mean time-series. (First row) Normalized overlaps. (Second row) Relative overlaps.

Inter-classification overlaps Overlaps *between* classifications are defined as in (2), but with j standing for the jth technological class and k for the kth semantic class: p_{ij} are technological probabilities and p_{ik} semantic probabilities. They describe the relative correspondence between the two classifications and are a good indicator to spot relative changes, as shown in Fig. 8. Mean inter-classification overlap clearly exhibits two linear trends, the first one being constant from 1980 to 1996, followed by a constant decrease. Although difficult to interpret directly, this stylized fact clearly unveils a change in the *nature* of inventions, or at least in the relation between content of inventions and technological classification. As the tipping point is at the same time as the ones observed in the previous section and since the two statistics are different, it is unlikely that this is a mere coincidence. Thus, these observations could be markers of a hidden underlying structural changes in processes.

Figure 8. Distribution of relative overlaps between classifications. (Left) Distribution of overlaps at all time steps; (Right) Corresponding mean time-series. The decreasing trend starting around 1996 confirms a qualitative regime shift in that period.

4.3 Citation Modularity

An exogenous source of information on relevance of classifications is the citation network described in Section 2.2. The correspondence between citation links and classes should provide a measure of accuracy of classifications, in the sense of an external validation since it is well-known that citation homophily is expected to be quite high (see, e.g, [14]). This section studies empirically modularities of the citation network regarding the different classifications. To corroborate the obtained results, we propose to look at a more rigorous framework in Section 4.4. Modularity is a simple measure of how communities in a network are well clustered (see [31] for the accurate definition). Although initially designed for single-class classifications, this measure can be extended to the case where nodes can belong to several classes at the same time, in our case with different probabilities as introduced in [44]. The simple directed modularity is given in our case by

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

with A_{ij} the citation adjacency matrix (i.e. $A_{ij} = 1$ if there is a citation from the ith patent to the jth patent, and $A_{ij} = 0$ if not), $k_i^{in} = |I_i|$ (resp. $k_i^{out} = \left|\tilde{I}_i\right|$) in-degree (resp. out-degree) of patents (i.e. the number of citations made by the ith patent to others and the number of citations received by the ith patent). Q_d can be defined for each of the two classification systems: $z \in \{tec, sem\}$. If $z = tec$, c_i is defined as the main patent class, which is taken as the first class whereas if $z = sem$, c_i is the class with the largest probability.

Multi-class modularity in turns is given by

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

where

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

We take $F(p_{ic}, p_{jc}) = p_{ic} \cdot p_{jc}$ as suggested in [44]. Modularity is an aggregated measure of how the network deviates from a null model where links would be randomly made according to node degree. In other words it captures the propensity for links to be inside the classes. Overlapping modularity naturally extends simple modularity by taking into account the fact that nodes can belong simultaneously to many classes. We document in Fig. 9 both simple and multi-class modularities over time. For simple modularity,

$Q_d^{(tec)}$ is low and stable across the years whereas $Q_d^{(sem)}$ is slightly greater and increasing. These values are however low and suggest that single classes are not sufficient to capture citation homophily. Multi-class modularities tell a different story. First of all, both classification modularities have a clear increasing trend, meaning that they become more and more adequate with citation network. The specializations revealed by both patent level diversities and classes overlap is a candidate explanation for this growing modularities. Secondly, semantic modularity dominates technological modularity by an order of magnitude (e.g. 0.0094 for technological against 0.0853 for semantic in 2007) at each time. This discrepancy has a strong qualitative significance. Our semantic classification fits better the citation network when using multiple classes. As technologies can be seen as a combination of different components as shown by [5], this heterogeneous nature is most likely better taken into account by our multi-class semantic classification.

Figure 9. Temporal evolution of semantic and technological modularities of the citation network. (Left) Simple directed modularity, computed with patent main classes (main technological class and semantic class with larger probability). (Right) Multi-class modularity, computed following [44]

4.4 Statistical Model

In this section, we develop a statistical model aimed at quantifying performance of both technological and semantic classification systems. In particular, we aim at corroborating findings obtained in Section 4.3. The mere difference between this approach and the citation modularity approach lies in the choice of the underlying model, and the according quantities of interest. In addition for the semantic approach, we want to see if when restricting to patents with higher probabilities to belong to a class, we obtain better results. To do that, we choose to look at within class citations proportion (for both technological and semantic approaches). We provide two obvious reasons why we choose this. First, the citations are commonly used as a proxy for performance as mentioned in Section 4.3. Second, this choice is "statistically fair" in the sense that both approaches have focused on various goals and not on maximizing directly the within class proportion. Nonetheless, the within class proportion is too sensitive to the distribution of the shape of classes. For example, a dataset where patents for each class account for 10% of the total number of patents will mechanically have a better within class proportion than if each class accounts for only 1%. Consequently, an adequate statistical model, which treats datasets fairly regardless of their distribution in classes, is needed. This effort ressembles to the previous study of citation modularity, but is complementary since the model presented here can be understood as an elementary model of citation network growth. Furthermore, the parameters fitted here can have a direct interpretation as a citation probability.

We need to introduce and recall some notations. We consider a specific window of observations $\big[t - T_0, t\big]$, and we define Z the number of patents which appeared during that time window. We let $t_1, \cdots, t_Z$ their corresponding appearance date by chronological order, which for simplicity are assumed to be such that $t_1 < \cdots < t_Z$. For each patent $i = 1, \cdots, Z$ we consider C_i the number of distinctive couples {cited patent, cited patent's class} made by the ith patent (for instance if the ith patent has only made one citation and that the cited patent is associated with three classes, then $C_i = 3$). Let $z \in \{tec, sem\}$, we define $N_i^{(z)}$ the number of patents associated to at least one of the ith classes at time t_{i-1}. For $l = 1, \cdots, C_i$ we consider the variables $B_{l,i}$, which equal 1 if the cited patent's class is also common to the ith patent. We assume that $B_{l,i}$ are independent of each other and conditioned on the past follow Bernoulli variables

$$B\Big(\min\Big\{1, \frac{N_i^{(z)}}{i-1} + \theta^{(z)}\Big\}\Big),$$

where the parameter $0 \leq \theta^{(z)} \leq 1$ indicates the propensity for any patent to cite patents of its own technological or semantic class. When $\theta^{(z)} = 0$, the probability of citing patents from its own class is simply $N_i^{(z)}(i-1)^{-1}$, which corresponds to the observed proportion of patents which belong to at least one of the ith patent's classes. Thus this corresponds to the estimated probability of citing one patent if we assume that the probability of citing any patent $k = 1, \cdots, i-1$ is uniformly distributed, which could be a reasonable assumption if classes were assigned randomly and independently from patent abstract contents. Conversely if $\theta^{(z)} = 1$, we are in the case of a model where there are 100% of within class citations. A reasonable choice of $\theta^{(z)}$ lies between those two extreme values. Finally, we assume that the number of distinctive couples C_i are a sequence of independent and identically distributed random variables following the discrete distribution C, and also independent from the other quantities.

We estimate $\theta^{(z)}$ via maximum likelihood, and obtain the corresponding maximum likelihood estimator (MLE) $\hat{\theta}^{(z)}$. The likelihood function, along with the standard deviation expression and details about the test, can be found in Supporting Information. The fitted values, standard errors and p-values corresponding to the statistical test $\theta^{(sem)} = \theta^{(tec)}$ (with corresponding alternative hypothesis $\theta^{(sem)} > \theta^{(tec)}$) on non-overlapping blocks from the period 1980-2007[6] are reported on Table 1. Semantic values are reported for four different chosen thresholds $p^{-} = .04, .06, .08, .1$. It means that we restricted to the couples (ith patent, jth class) such that $p_{ij} \geq p^{-}$.

The choice of considering non-overlapping blocks (instead of overlapping blocks) is merely statistical. Ultimately, our interest is in the significance of the test over the whole period 1980-2007. Thus, we want to compute a global p-value. This can be done considering the local p-values (by local, we mean for instance computed on the period 2001-2005) assuming independence between them. This assumption is reasonable only if the blocks are non-overlapping. All of this can be found in Supporting Information. Finally, note that from a statistical perspective, including overlapping blocks wouldn't yield more information.

The values reported in Table 1 are overwhelmingly against the null hypothesis. The global estimates of $\theta^{(sem)}$ are significantly bigger than the estimate of $\theta^{(tec)}$ for all the considered thresholds. Although the corresponding p-values (which are also very close to 0) are not reported, it is also quite clear that the bigger the threshold, the higher the corresponding $\theta^{(sem)}$ is estimated. This is consistently seen for any period, and significant for the global period. This seems to indicate that when restricting to the couples (patent, class) with high semantic probability, the propension to cite patents from its own class $\theta^{(sem)}$ is increasing. We believe that this might provide extra information to patent officers when making their choice of citations. Indeed, they could look first to patents which belong to the same semantic class, especially when patents have high probability semantic values.

Note that the introduced model can be seen as a simple model of citations network growth conditional to a classification, which can be expressed as a stochastic block model (e.g. [45], [46]). The parameters are estimated computing the corresponding MLE. In

[6] Note that the estimation included patents up until 2010 in the period 2006-2007 and not the patents from 1980 in the period 1980-1985 for homogeneity in size with other periods. This doesn't affect the significativity of the results.

view of [47], this can be thought as equivalent to maximizing modularity measures.

Table 1. Estimated values of $\theta^{(tec)}$ and $\theta^{(sem)}$ and corresponding standard errors obtained from a Maximum Likelihood estimator as presented in section 4.4.

Approach	Estimated Value	st. er.	p-value
1980-1985 period			
technological	.664	.008	
semantic $p^- = .04$	.741	.047	.053
semantic $p^- = .06$	.799	.081	.049
semantic $p^- = .08$	.828	.126	.097
semantic $p^- = .10$	.834	.166	.153
1986-1990 period			
technological	.634	.007	
semantic $p^- = .04$	.703	.022	.001
semantic $p^- = .06$	.768	.040	.0004
semantic $p^- = .08$	.804	.069	.007
semantic $p^- = .10$	.832	.114	.041
1991-1995 period			
technological	.619	.006	
semantic $p^- = .04$	.655	.009	.0004
semantic $p^- = .06$	.713	.017	9e-08
semantic $p^- = .08$	.731	.025	7e-06
semantic $p^- = .10$	.750	.037	9e-06
1996-2000 period			
technological	.551	.003	
semantic $p^- = .04$	.585	.002	≈ 0
semantic $p^- = .06$	.638	.004	≈ 0
semantic $p^- = .08$	.660	.006	≈ 0
semantic $p^- = .10$	.686	.008	≈ 0
2001-2005 period			
technological	.567	.003	
semantic $p^- = .04$	.621	.004	≈ 0
semantic $p^- = .06$	.676	.007	≈ 0
semantic $p^- = .08$	.701	.010	≈ 0
semantic $p^- = .10$	.710	.013	≈ 0
2006-2007 period			
technological	.600	.007	
semantic $p^- = .04$	.683	.016	1e-06
semantic $p^- = .06$	.732	.025	2e-07
semantic $p^- = .08$	.760	.036	6e-06
semantic $p^- = .10$	.782	.048	9e-05
1980-2007 global period			
technological	.606	.002	
semantic $p^- = .04$	.665	.009	8e-11
semantic $p^- = .06$	.721	.017	9e-12
semantic $p^- = .08$	.747	.025	9e-09
semantic $p^- = .10$	.782	.035	3e-07

5 Conclusion

The main contribution of this study was twofold. First we have defined how we built a network of patents based on a classification that uses semantic information from abstracts. We have shown that this classification share some similarities with the traditional technological classification, but also have distinct features. Second, we provide researchers with materials resulting from our analysis, which includes: (i) a database linking each patent with its set of semantic classes and the associated probabilities; (ii) a list of these semantic classes with a description based on the most relevant keywords; (iii) a list of patent with their topological properties in the semantic network (centrality, frequency, degree, etc.). The availability of this data suggests new avenues for further research. Linking our dataset with existing open ones can lead to various powerful developments. For example, using it together with the disambiguated inventor database provided by [48] could be a way to study semantic profiles of inventors, or of cities as inventor addresses are provided. The investigation of spatial diffusion of innovation between cities, which is a key component of Pumain's Evolutive Urban Theory [49] , would be made possible.

A first potential application is to use the patents' topological measures inherited from their relevant keywords. The fact that these measures are backward-looking and immediately available after the publication of the patent information is an important asset. It would for example be very interesting to test their predicting power to assess the quality of an innovation, using the number of forward citations received by a patent, and subsequently the future effect on the firm's market value.

Regarding firm innovative strategy, a second extension could be to study trajectories of firms in the two networks: technological and semantic. Merging these information with data on the market value of firms can give a lot of insight about the more efficient innovative strategies, about the importance of technology convergence or about acquisition of small innovative firms. It will also allow to observe innovation pattern over a firm life cycle and how this differ across technology field.

A third extension would be to use dig further into the history of innovation. USPTO patent data have been digitized from the first patent in July 1790. However, not all of them contain a text that is directly exploitable. We consider that the quality of patent's images is good enough to rely on Optical Character Recognition techniques to retrieve plain text from at least 1920. With such data, we would be able to extend our analysis further back in time and to study how technological progress occurs and combines in time. [50] conduct a similar work by looking at recombination and apparition of technological subclasses. Using the fact that communities are constructed yearly, one can construct a measure of proximity between two successive classes. This could give clear view on how technologies converged over the year and when others became obsolete and replaced by new methods.

Supporting Information

S1 Text : Definition of utility patent

Describes with more details the definition of patents and context.

S2 Text : Data collection procedure

Detailed description of data collection

S3 File : Semantic Network Visualization

Vector file of the semantic network (Fig.2)

S4 Text : Network Sensitivity Analysis

Extended figures for Network Sensitivity Analysis

S5 Text : Statistical definitions and derivations

Extended definitions and derivations for the statistical model

References

1. Aghion P, Howitt P. A Model of Growth through Creative Destruction. Econometrica. 1992 March;60(2):323–51. Available from: https://ideas.repec.org/a/ecm/emetrp/v60y1992i2p323-51.html.

2. Romer PM. Endogenous Technological Change. Journal of Political Economy. 1990 October;98(5):S71–102. Available from: https://ideas.repec.org/a/ucp/jpolec/v98y1990i5ps71-102.html.
3. Griliches Z. Patent Statistics as Economic Indicators: A Survey. National Bureau of Economic Research, Inc; 1990. 3301. Available from: https://ideas.repec.org/p/nbr/nberwo/3301.html.
4. Hall BH, Jaffe AB, Trajtenberg M. The NBER Patent Citations Data File: Lessons, Insights and Methodological Tools. C.E.P.R. Discussion Papers; 2001. 3094. Available from: https://ideas.repec.org/p/cpr/ceprdp/3094.html.
5. Youn H, Strumsky D, Bettencourt LMA, Lobo J. Invention as a combinatorial process: evidence from US patents. Journal of The Royal Society Interface. 2015;12(106).
6. Newman MEJ. Prediction of highly cited papers. ArXiv e-prints. 2013 Oct;.
7. Sarigöl E, Pfitzner R, Scholtes I, Garas A, Schweitzer F. Predicting Scientific Success Based on Coauthorship Networks. ArXiv e-prints. 2014 Feb;.
8. Sorenson O, Rivkin JW, Fleming L. Complexity, networks and knowledge flow. Research policy. 2006;35(7):994–1017.
9. Kay L, Newman N, Youtie J, Porter AL, Rafols I. Patent overlay mapping: Visualizing technological distance. Journal of the Association for Information Science and Technology. 2014;65(12):2432–2443.
10. Bruck P, Réthy I, Szente J, Tobochnik J, Érdi P. Recognition of emerging technology trends: class-selective study of citations in the US Patent Citation Network. Scientometrics. 2016;107(3):1465–1475.
11. Curran CS, Leker J. Patent indicators for monitoring convergence–examples from NFF and ICT. Technological Forecasting and Social Change. 2011;78(2):256–273.
12. Katz ML. Remarks on the economic implications of convergence. Industrial and Corporate Change. 1996;5(4):1079–1095.

13. Furman JL, Stern S. Climbing atop the Shoulders of Giants: The Impact of Institutions on Cumulative Research. American Economic Review. 2011 August;101(5):1933–63. Available from: http://www.aeaweb.org/articles?id=10.1257/aer.101.5.1933.

14. Acemoglu AU Daron, Kerr W. Innovation Network. Proceedings of the National Academy of Sciences (forthcoming). 2016;.

15. Preschitschek N, Niemann H, Leker J, Moehrle MG. Anticipating industry convergence: Semantic analyses vs IPC co-classification analyses of patents. Foresight. 2013 11;15(6):446–464.

16. Yoon B, Park Y. A text-mining-based patent network: Analytical tool for high-technology trend. The Journal of High Technology Management Research. 2004;15(1):37–50.

17. Park I, Yoon B. A semantic analysis approach for identifying patent infringement based on a product–patent map. Technology Analysis & Strategic Management. 2014;26(8):855–874.

18. Yoon J, Kim K. Detecting signals of new technological opportunities using semantic patent analysis and outlier detection. Scientometrics. 2011;90(2):445–461.

19. Gerken JM, Moehrle MG. A new instrument for technology monitoring: novelty in patents measured by semantic patent analysis. Scientometrics. 2012;91(3):645–670.

20. Choi J, Hwang YS. Patent keyword network analysis for improving technology development efficiency. Technological Forecasting and Social Change. 2014;83:170–182.

21. Fattori M, Pedrazzi G, Turra R. Text mining applied to patent mapping: a practical business case. World Patent Information. 2003;25(4):335–342.

22. Gurciullo S, Smallegan M, Pereda M, Battiston F, Patania A, Poledna S, et al. Complex Politics: A Quantitative Semantic and Topological Analysis of UK House of Commons Debates. ArXiv e-prints. 2015 Oct;.

23. Lerner J, Seru A. The use and misuse of patent data: Issues for corporate finance and beyond. Booth/Harvard Business School Working Paper. 2015;.

24. OECD. OECD Patent Statistics Manual. 2009;Available from: /content/book/9789264056442-en.

25. DechezleprÃªtre A, Martin R, Mohnen M. Knowledge Spillovers from Clean and Dirty Technologies. Centre for Economic Performance, LSE; 2014. dp1300. Available from: https://ideas.repec.org/p/cep/cepdps/dp1300.html.

26. Tseng YH, Lin CJ, Lin YI. Text mining techniques for patent analysis. Information Processing & Management. 2007;43(5):1216–1247.

27. Adams S. The text, the full text and nothing but the text: Part 1 - Standards for creating textual information in patent documents and general search implications. World Patent Information. 2010 March;32(1):22–29. Available from: https://ideas.repec.org/a/eee/worpat/v32y2010i1p22-29.html.

28. Abbas A, Zhang L, Khan SU. A literature review on the state-of-the-art in patent analysis. World Patent Information. 2014;37:3–13.

29. Chavalarias D, Cointet JP. Phylomemetic patterns in science evolution—the rise and fall of scientific fields. Plos One. 2013;8(2):e54847.

30. NLTK. Natural Language Toolkit, Stanford Univeristy; 2015.

31. Clauset A, Newman ME, Moore C. Finding community structure in very large networks. Physical review E. 2004;70(6):066111.

32. Yang Y, Ault T, Pierce T, Lattimer CW. Improving text categorization methods for event tracking. In: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval. ACM; 2000. p. 65–72.

33. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. Journal of machine Learning research. 2003;3(Jan):993–1022.

34. Kaplan S, Vakili K. The double-edged sword of recombination in breakthrough innovation. Strategic Management Journal. 2015;36(10):1435–1457.

35. Zhu Y, Yan X, Getoor L, Moore C. Scalable text and link analysis with mixed-topic link models. In: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM; 2013. p. 473–481.

36. Iacovacci J, Wu Z, Bianconi G. Mesoscopic Structures Reveal the Network Between the Layers of Multiplex Datasets. arXiv preprint arXiv:150503824. 2015;.

37. De Domenico M, Solé-Ribalta A, Omodei E, Gómez S, Arenas A. Ranking in interconnected multilayer networks reveals versatile nodes. Nature communications. 2015;6.

38. Gao ZK, Small M, Kurths J. Complex network analysis of time series. EPL (Europhysics Letters). 2017;116(5):50001.

39. Gao ZK, Yang YX, Fang PC, Zou Y, Xia CY, Du M. Multiscale complex network for analyzing experimental multivariate time series. EPL (Europhysics Letters). 2015;109(3):30005.

40. Archibugi D, Pianta M. Specialization and size of technological activities in industrial countries: The analysis of patent data. Research Policy. 1992;21(1):79 – 93. Available from: http://www.sciencedirect.com/science/article/pii/004873339290028З.

41. Bloom N, Schankerman M, Reenen JV. Identifying Technology Spillovers and Product Market Rivalry. Econometrica. 2013 07;81(4):1347–1393. Available from: https://ideas.repec.org/a/ecm/emetrp/v81y2013i4p1347-1393.html.

42. Ziman J. Technological innovation as an evolutionary process. Cambridge University Press; 2003.

43. Holland JH. Signals and boundaries: Building blocks for complex adaptive systems. Mit Press; 2012.

44. Nicosia V, Mangioni G, Carchiolo V, Malgeri M. Extending the definition of modularity to directed graphs with overlapping communities. Journal of Statistical Mechanics: Theory and Experiment. 2009;2009(03):P03024.

45. Decelle A, Krzakala F, Moore C, Zdeborová L. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. Physical Review E. 2011;84(6):066106.

46. Valles-Catala T, Massucci FA, Guimera R, Sales-Pardo M. Multilayer stochastic block models reveal the multilayer structure of complex networks. Physical Review X. 2016;6(1):011036.

47. Newman MEJ. Community detection in networks: Modularity optimization and maximum likelihood are equivalent. ArXiv e-prints. 2016 Jun;.

48. Li GC, Lai R, D'Amour A, Doolin DM, Sun Y, Torvik VI, et al. Disambiguation and co-authorship networks of the US patent inventor database (1975–2010). Research Policy. 2014;43(6):941–955.

49. Pumain D. Une théorie géographique des villes. Bulletin de la Société géographie de Liège. 2010;(55):5–15.

50. Akcigit U, Kerr WR, Nicholas T. The Mechanics of Endogenous Innovation and Growth: Evidence from Historical US Patents. Citeseer; 2013.

Figure 1

Click here to download Figure Fig1.tif

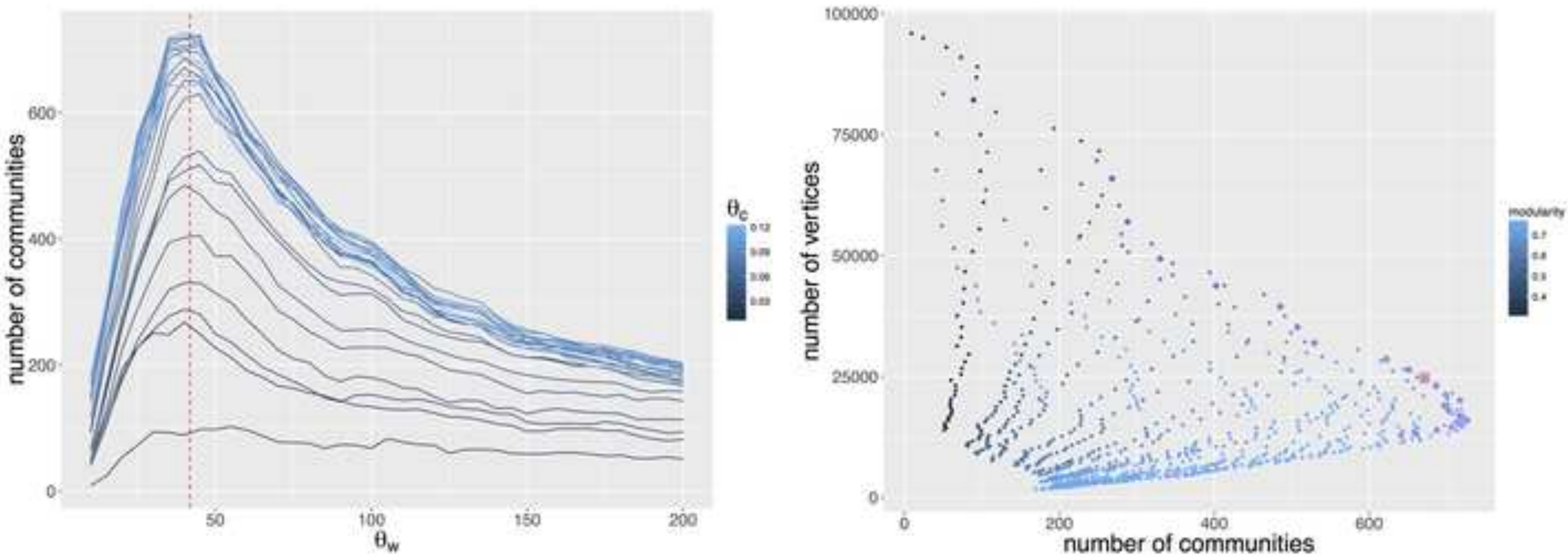


Figure 2

Click here to download Figure Fig2.tif

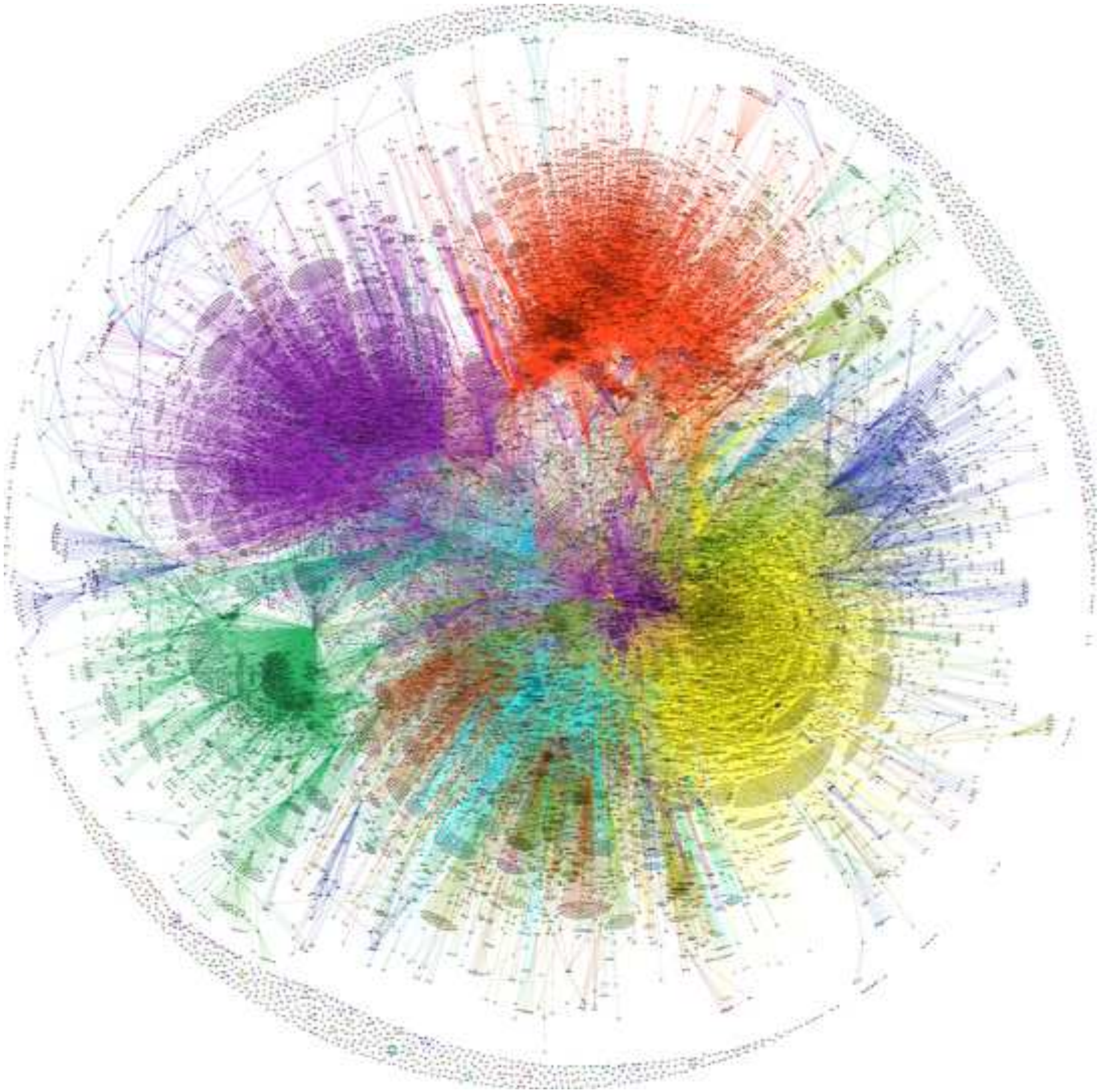


Figure 3

Click here to download Figure Fig3.tif

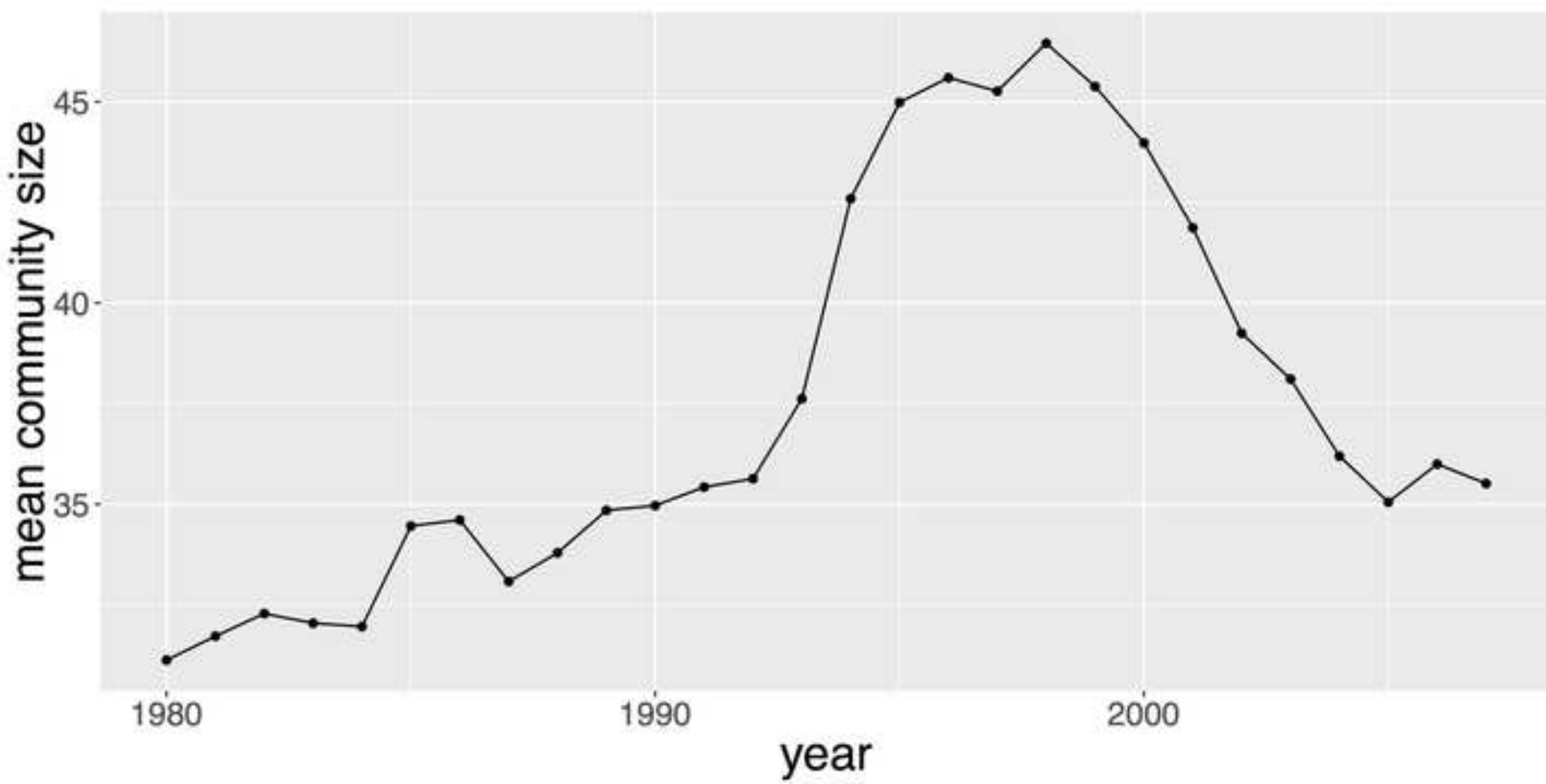
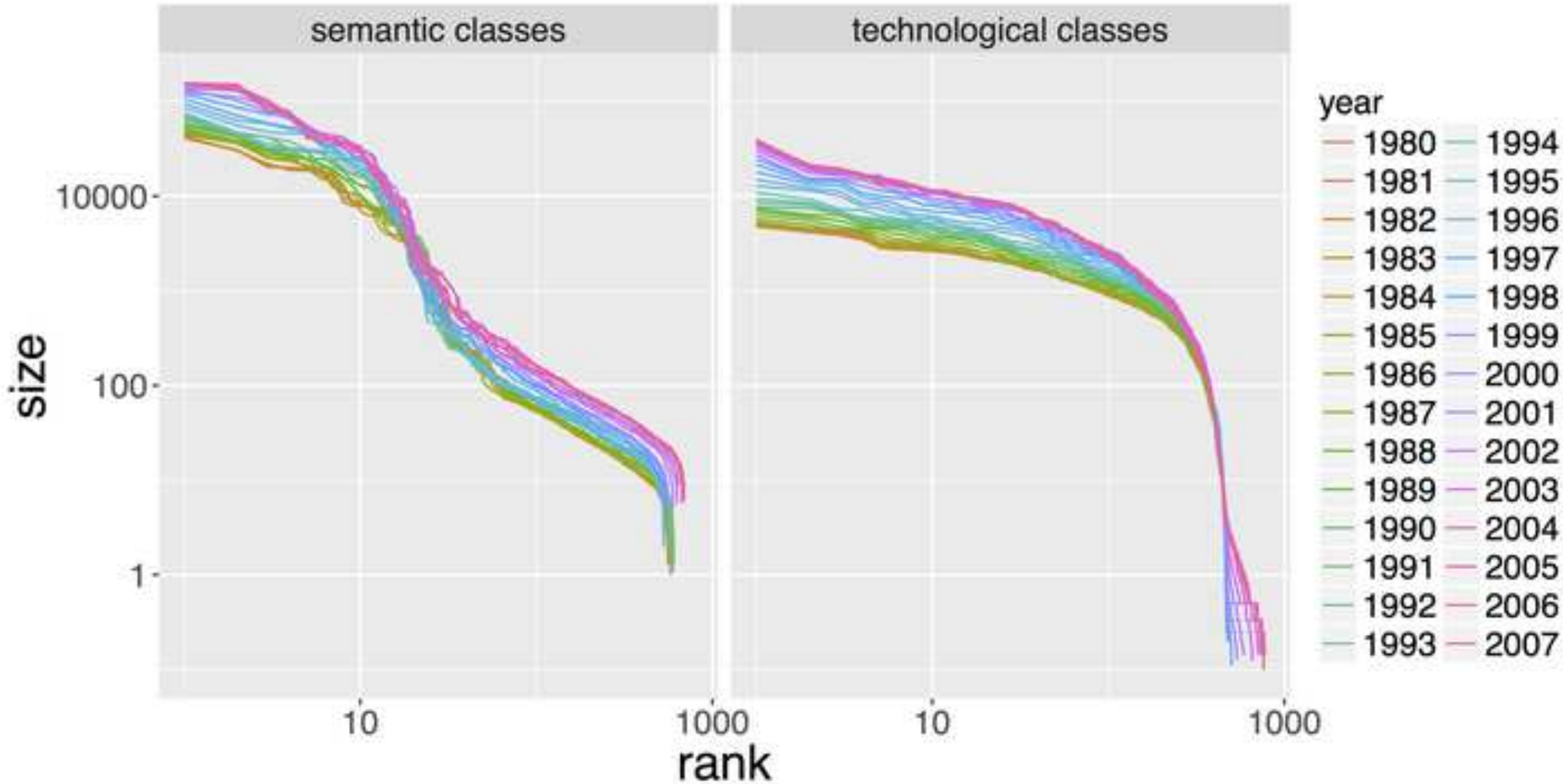


Figure 4



semantic classes
technological classes
size
10000
100
1
10
1000
10
1000
rank
year
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007

Figure 5

Click here to download Figure Fig5.tif

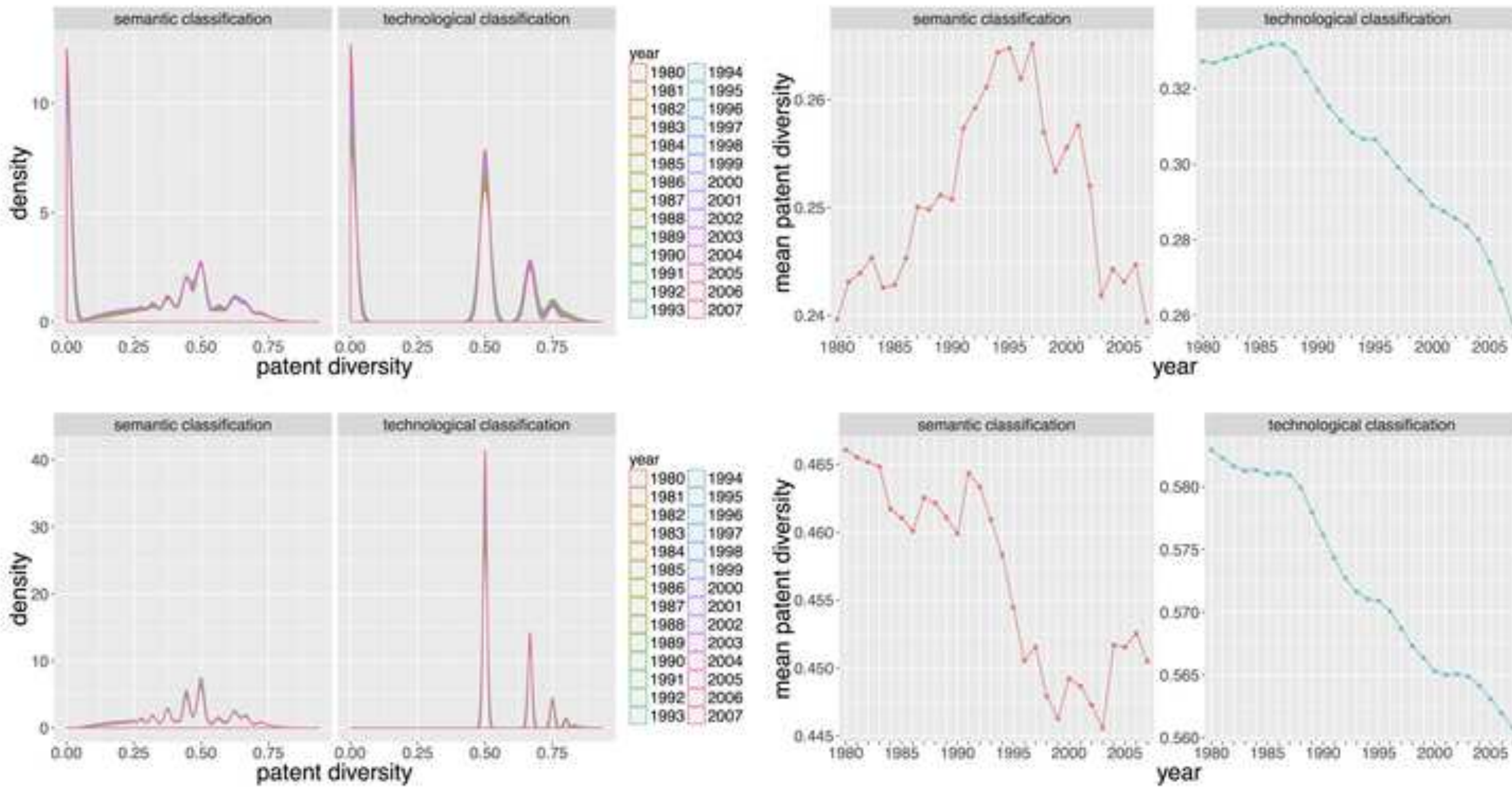


Figure 6

Click here to download Figure Fig6.tif

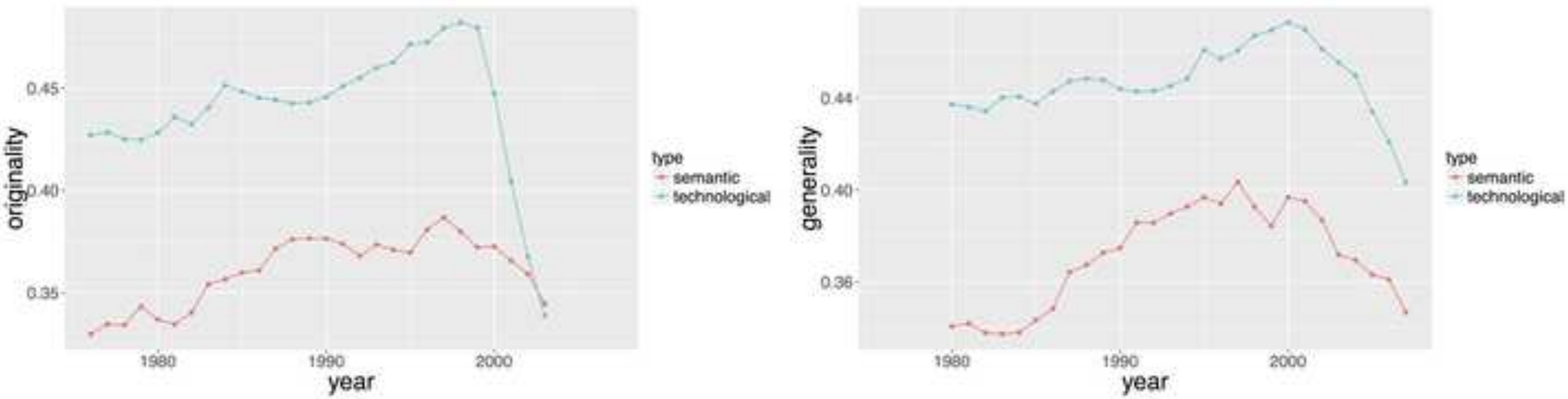


Figure 7

Click here to download Figure Fig7.tif

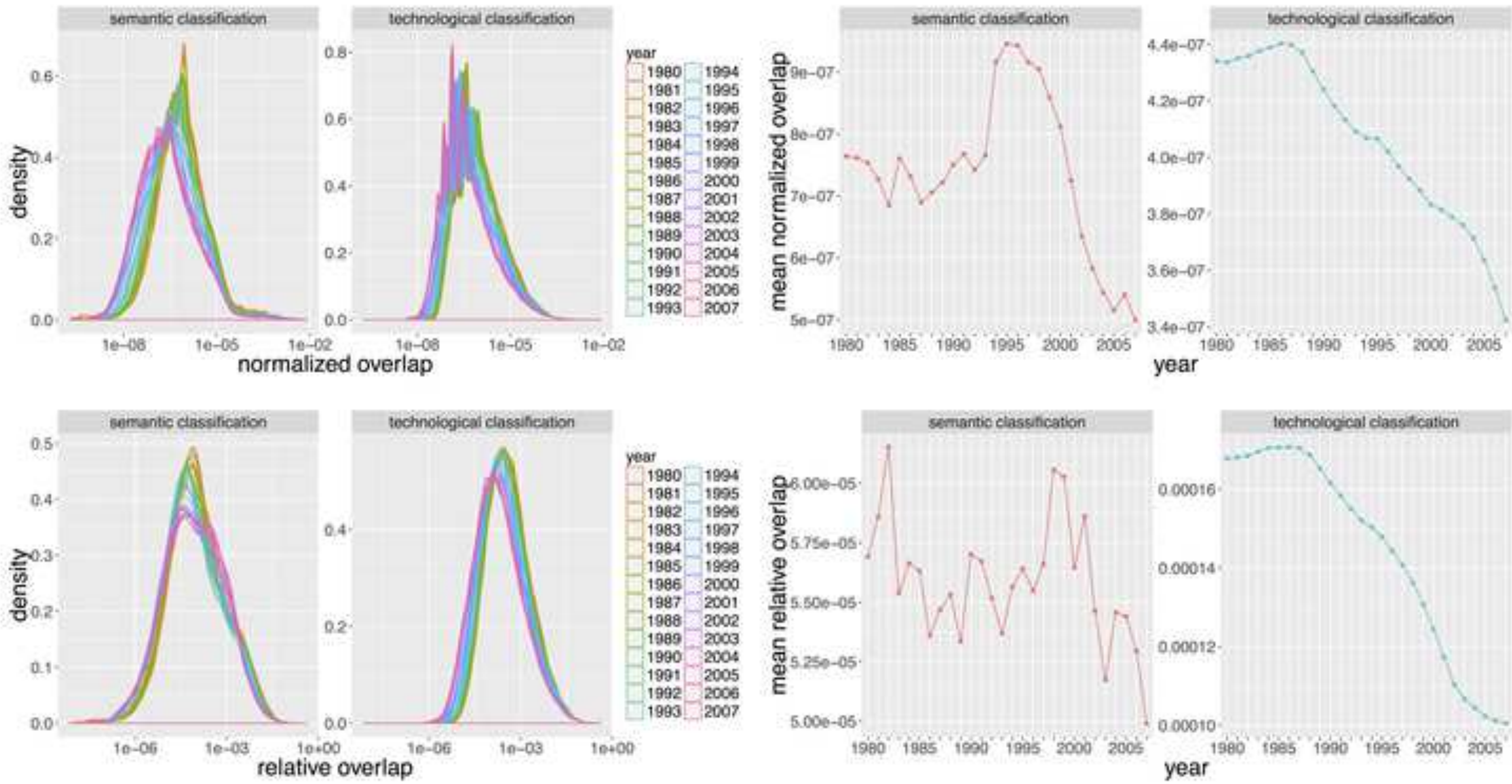


Figure 8

Click here to download Figure Fig8.tif

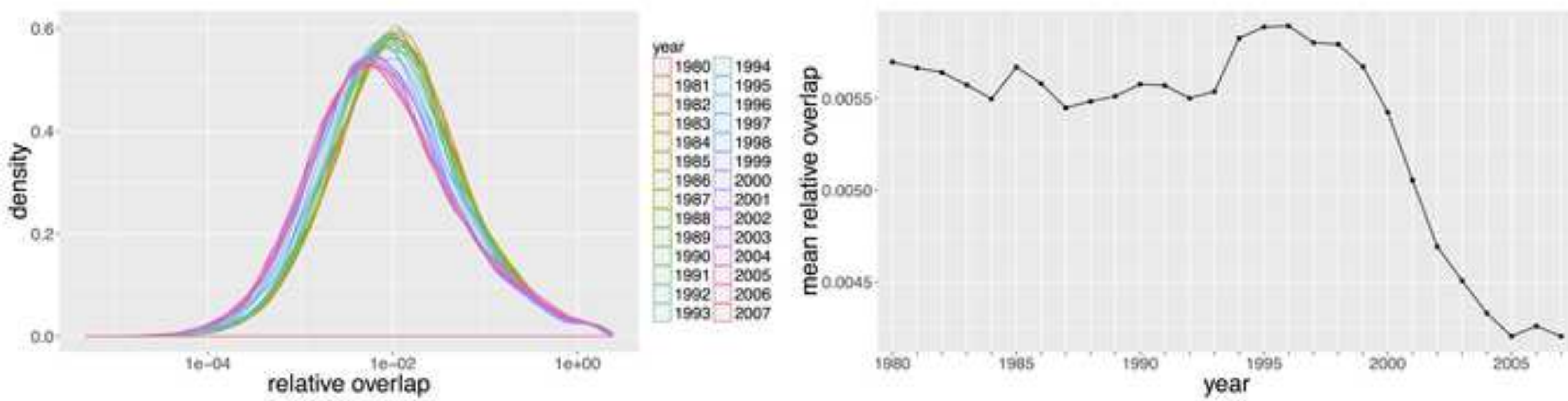
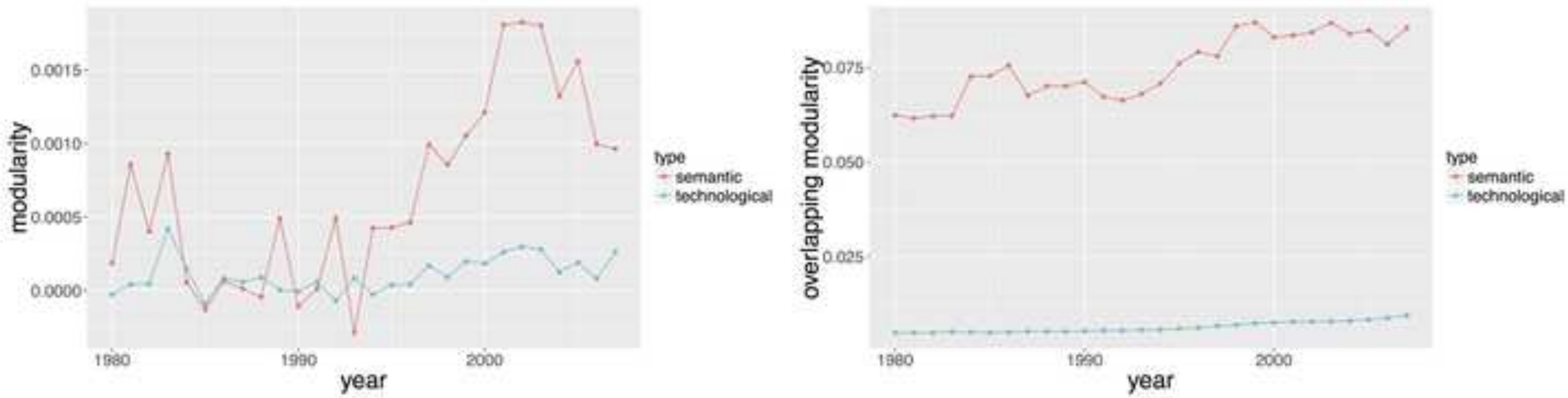


Figure 9

Click here to download Figure Fig9.tif



S1 Text

S2 Text

Click here to access/download
Supporting Information
S2_Text.pdf

S3 File

Click here to access/download
Supporting Information
S3_File.svg

S4 Text

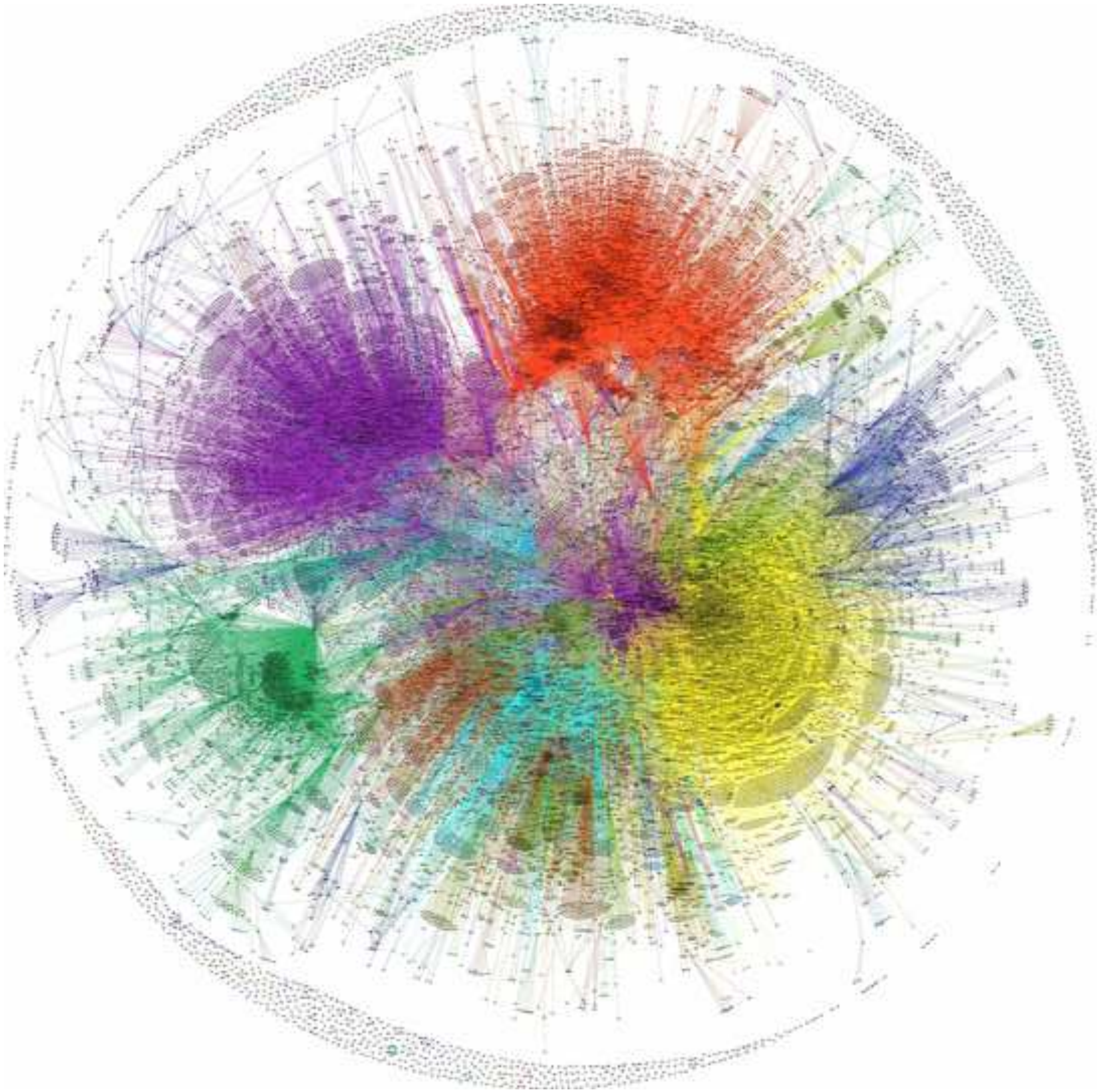
Click here to access/download
Supporting Information
S4_Text.pdf

S5 Text

Click here to access/download
Supporting Information
S5_Text.pdf

Striking Image

Click here to download Striking Image striking.jpg



Classifying Patents Based on their Semantic Content

Antonin Bergeaud[1,◕], Yoann Potiron[2,◕], Juste Raimbault[3,4,◕,*]

1 Paris School of Economics - EHESS and Bank of France, Paris, France

2 Faculty of Business and Commerce, Keio University, Tokyo, Japan

3 UMR CNRS 8504 Géographie-cités, Université Paris VII, Paris, France

4 UMR-T 9403 IFSTTAR LVMT, Ecole Nationale des Ponts et Chaussées, Champs-sur-Marne, France

◕These authors contributed equally to this work.

*** Corresponding Author**

Email : juste.raimbault@polytechnique.edu

Abstract

In this paper, we extend some usual techniques of classification resulting from a large-scale data-mining and network approach. This new technology, which in particular is designed to be suitable to big data, is used to construct an open consolidated database from raw data on 4 million patents taken from the US patent office from 1976 onward. To build the pattern network, not only do we look at each patent title, but we also examine their full abstract and extract the relevant keywords accordingly. We refer to this classification as *semantic approach* in contrast with the more common *technological approach* which consists in taking the topology when considering US Patent office technological classes. Moreover, we document that both approaches have highly different topological measures and strong statistical evidence that they feature a different model. This suggests that our method is a useful tool to extract endogenous information.

1 Introduction

Innovation and technological change have been described by many scholars as the main drivers of economic growth as in [1] and [2]. [3] advertised the use of patents as an economic indicator and as a good proxy for innovation. Subsequently, the easier availability of comprehensive databases on patent details and the increasing number of studies allowing a more efficient use of these data (e.g. [4]) have opened the way to a very wide range of analysis. Most of the statistics derived from the patent databases relied on a few key features: the identity of the inventor, the type and identity of the rights owner, the citations made by the patent to prior art and the technological classes assigned by the patent office post patent's content review. Combining this information is particularly relevant when trying to capture the diffusion of knowledge and the interaction between technological fields as studied in [5]. With methods such as citation dynamics modeling discussed in [6] or co-authorship networks analysis in [7], a large body of the literature such as [8] or [9] has studied patents citation network to understand processes driving technological innovation, diffusion and the birth of technological clusters. Finally, [10] look at the dynamics of citations from different classes to show that the laser/ink-jet printer technology resulted from the recombination of two different existing technologies.

Consequently, technological classification combined with other features of patents can be a valuable tool for researchers interested in studying technologies throughout history and to predict future innovations by looking at past knowledge and interaction across sectors and technologies. But it is also crucial for firms that face an ever changing demand structure and need to anticipate future technological trends and convergence (see, e.g., [11]) to adapt to the resulting increase in competition discussed in [12] and to maintain market share. Curiously, and in spite of the large number of studies that analyze interactions across technologies [13], little is known about the underlying "innovation network" (e.g. [14]).

In this monograph, we propose an alternative classification based on semantic network analysis from patent abstracts and explore the new information emerging from it. In contrast with the regular technological classification which results from the choice of the patent reviewer, semantic classification is carried automatically based on the content of the patent abstract. Although patent officers are experts in their fields, the relevance of the existing classification is limited by the fact that it is based on the state of technology at the time the patent was granted and cannot anticipate the birth of new fields.[1] In contrast we don't face this issue with the semantic approach. The semantic links can be clues of one technology taking inspiration from another and good predictors of future technology convergence (e.g. [15] study semantic similarities from the whole text of 326 US-patents on *phytosterols* and show that semantic analysis have a good predicting power of future technology convergence). One can for instance consider the case of the word *optic*. Until more recently, this word was often associated with technologies such as photography or eye surgery, while it is now almost exclusively used in a context of semi-transistor design and electro-optic. This semantic shift did not happen by chance but contains information on the fact that modern electronic extensively uses technologies that were initially developed in optic.

Previous research has already proposed to use semantic networks to study technological domains and detect novelty. [16] was one of the first to enhance this approach with the idea of visualizing keywords network illustrated on a small technological domain. The same approach can be used to help companies identifying the state of the art in their field and avoid patent infringement as in [17] and [18]. More closely related to our methodology, [19] develop a method based on patent semantic analysis of patent to vindicate the view that this approach outperform others in the monitoring of technology and in the identification of novelty innovation. Semantic analysis has already proven its efficiency in various fields, such as in technology studies (e.g. [20] and [21]) and in political science (e.g. [22]).

Building on such previous research, we make several contributions by fulfilling some shortcomings of existing studies, such as for example the use of frequency-selected single keywords. First of all, we develop and implement a novel fully-automatized methodology to classify patents according to their semantic abstract content, which is to the best of our knowledge the first of its type. This includes the following refinements for which details can be found in Section 3: (i) use of multi-stems as potential keywords; (ii) filtering of keywords based on a second-order (co-occurrences) relevance measure and on an external independent measure (technological dispersion); (iii) multi-objective

[1] To correct for this, the USPTO regularly make changes in its classification in order to adapt to technological change (for example, the "nanotechnology" class (977) was established in 2004 and retroactively to all relevant previously granted patents).

optimization of semantic network modularity and size. The use of all this techniques in the context of semantic classification is new and essential from a practical perspective.

Furthermore, most of the existing studies rely on a subsample of patent data, whereas we implement it on the full US Patent database from 1976 to 2013. This way, a general structure of technological innovation can be studied. We draw from this application promising qualitative stylized facts, such as a qualitative regime shift around the end of the 1990s, and a significant improvement of citation modularity for the semantic classification when comparing to the technological classification. These thematic conclusions validate our method as a useful tool to extract endogenous information, in a complementary way to the technological classification.

Finally, the statistical model introduced in Section 4.4 seems to indicate that patents tend to cite more similar patents in the semantic network when fitted to data. In particular, this propensity is shown to be significantly bigger than the corresponding propensity for technological classes, and this seems to be consistent over time. On the account of this information, we believe that patent officers could benefit very much from looking at the semantic network when considering potential citation candidates of a patent in review.

The paper is organized as follows. Section 2 presents the patent data, the existing classification and provide details about the data collection process. Section 3 explains the construction of the semantic classes. Section 4 tests their relevance by providing exploratory results. Finally, section 5 discusses potential further developments and conclude. More details, including robustness checking, figures and technical derivations can be found in Supporting Information.

2 Background

In our analysis, we will consider all utility patents granted in the United States Patent and Trademark Office (USPTO) from 1976 to 2013. A clearer definition of utility patent is given in Supporting Information. Also, additional information on how to correctly exploit patent data can be found in [4] and [23].

2.1 An existing classification: the USPC system

Each USPTO patent is associated with a non-empty set of technological classes and subclasses. There are currently around 440 classes and over 150,000 subclasses constituting the United State Patent Classification (USPC) system. While a technological class corresponds to the technological field covered by the patent, a subclass stands for a specific technology or method used in this invention. A patent can have multiple technological classes, on average in our data a patent has 1.8 different classes and 3.9 pairs of class/subclass. At this stage, two features of this system are worth mentioning: (i) classes and subclasses are not chosen by the inventors of the patent but by the examiner during the granting process based on the content of the patent; (ii) the classification has evolved in time and continues to change in order to adapt to new technologies by creating or editing classes. When a change occurs, the USPTO reviews all the previous patents so as to create a consistent classification.

2.2 A bibliographical network between patents: citations

As with scientific publications, patents must give reference to all the previous patents which correspond to related prior art. They therefore indicate the past knowledge which relates to the patented invention. Yet, contrary to scientific citations, they also have an important legal role as they are used to delimit the scope of the property rights awarded by the patent. One can consult [24] for more details about this. Failing to refer to prior art can lead to the invalidation of the patent (e.g. [25]). Another crucial difference is that the majority of the citations are actually chosen by the examiners and not by the inventors themselves. From the USPTO, we gather information of all citations made by each patent (backward citations) and all citations received by each patent as of the end of 2013 (forward citations). We can thus build a complete network of citations that we will use later on in the analysis.

Turning to the structure of the lag between the citing and the cited patent in terms of application date, we see that the mean of this lag is 8.5 years and the median is 7 years. This distribution is highly skewed, the 95^{th} percentile is 21 years. We also report 164,000 citations with a negative time lag. This is due to the fact that some citations can be added during the examination process and some patents require more time to be granted than others.

In what follows, we choose to restrict attention to pairs of citations with a lag no larger than 5 years. We impose this restriction for two reasons. First, the number of citations received peaks 4-5 years after application. Second, the structure of the citation lag is necessarily biased by the truncation of our sample: the more recent patents mechanically receive less citations than the older ones. As we are restricting to citations received no later than 5 years after the application date, this effect will only affect patents with an application date after 2007.

2.3 Data collection and basic description

Each patent contains an abstract and a core text which describe the invention.[2] Although including the full core texts would be natural and probably very useful in a systematic text-mining approach as done in [26], they are too long to be included and thus we consider only the abstracts for the analysis. Indeed, the semantic analysis counts more than 4 million patents, with corresponding abstracts with an average length of 120.8 words (and a standard deviation of 62.4), a size that is already challenging in terms of computational burden and data size. In addition, abstracts are aimed at synthesizing purpose and content of patents and must therefore be a relevant object of study (see [27]). The USPTO defines a guidance stating that an abstract should be "a summary of the disclosure as contained in the description, the claims, and any drawings; the summary shall indicate the technical field to which the invention pertains and shall be drafted in a way which allows the clear understanding of the technical problem, the gist of the solution of that problem through the invention, and the principal use or uses of the invention" (PCT Rule 8).

We construct from raw data a unified database. Data is collected from USPTO patent redbook bulk downloads, that provides as raw data (specific `dat` or `xml` formats) full patent information, starting from 1976. Detailed procedure of data collection, parsing and consolidation are available in Supporting Information. The latest dump of the database in `Mongodb` format is available at `http://dx.doi.org/10.7910/DVN/BW3ACK`. Collection and homogenization of the database into a directly usable database with

[2] To see what a patent looks like in practice, one can refer to the USPTO patent full-text database `http://patft.uspto.gov/netahtml/PTO/index.html` or to Google patent which publishes USPTO patents in *pdf* format at `https://patents.google.com`.

basic information and abstracts was an important task as USPTO raw data formats are involved and change frequently.

We count 4,666,365 utility patents with an abstract granted from 1976 to 2013.[3] The number of patents granted each year increases from around 70,000 in 1976 to about 278,000 in 2013. When distributed by the year of application, the picture is slightly different. The number of patents steadily increase from 1976 to 2000 and remains constant around 200,000 per year from 2000 to 2007. Restricting our sample to patent with application date ranging from 1976 to 2007, we are left with 3,949,615 patents. These patents cite 38,756,292 other patents with the empirical lag distribution that has been extensively analyzed in [4]. Conditioned on being cited at least once, a patent receives on average 13.5 citations within a five-year window. 270,877 patents receive no citation during the next five years following application, 10% of patents receive only one citation and 1% of them receive more than 100 citations. A within class citation is defined as a citation between two patents sharing at least one common technological class. Following this definition, 84% of the citations are within class citations. 14% of the citations are between two patents that share the exact same set of technological classes.

2.4 Towards a Complementary Classification

Potentialities of text-mining techniques as an alternative way to analyze and classify patents are documented in [26]. The author's main argument, in support of an automatic classification tool for patent, is to reduce the considerable amount of human effort needed to classify all the applications. The work conducted in the field of natural language processing and/or text analysis has been developed in order to improve search performance in patent databases, build technology map or investigate the potential infringement risks prior to developing a new technology (see [28] for a review). Text-mining of patent documents is also widely used as a tool to build networks which carry additional information to the simplistic bibliographic connections model as argued in [16]. As far as the authors know, the use of text-mining as a way to build a global classification of patents remains however largely unexplored. One notable exception

[3] A very small number of patents have a missing abstract, these are patents that have been withdrawn and we do not consider them in the analysis.

can be found in [15] where semantic-based classification is shown to outperform the standard classification in predicting the convergence of technologies even in small samples. Semantic analysis reveals itself to be more flexible and more quickly adaptable to the apparition of new clusters of technologies. Indeed, as argued in [15], before two distinct technologies start to clearly converge, one should expect similar words to be used in patents from both technologies.

Finally, a semantic classification where patents are gathered based on the fact that they share similar significant keywords has the advantage of including a network feature that cannot be found in the USPC case, namely that each patent is associated with a vector of probability to belong to each of the semantic classes (more details on this feature can be found in Section 3.4). Using co-occurrence of keywords, it is then possible to construct a network of patents and to study the influence of some key topological features. As reviewed previously, the use of co-occurrences is the usual way to construct a semantic network. Other hybrid technique such as bipartite semantic/authors networks, do not have the nice feature of relying solely on endogenous semantic information contained in data.

3 Semantic Classification Construction

In this section, we describe methods and empirical analysis leading to the construction of semantic network and the corresponding classification.

3.1 Keywords extraction

Let $\mathcal{P}$ be the set of patents, we first assign to a patent $p \in \mathcal{P}$ a set of potentially significant keywords $K(p)$ from its text $\mathcal{A}(p)$ (that corresponds to the concatenation of its own title and abstract). $K(p)$ are extracted through a similar procedure as the one detailed in [29]:

1. Text parsing and Tokenization: we transform raw texts into a set of words and sentences, reading it (parsing) and splitting it into elementary entities (words organized in sentences).

2. Part-of-speech tagging: attribution of a grammatical function to each of the tokens

defined previously.

3. Stem extraction: families of words are generally derived from a unique root called stem (for example `compute`, `computer`, `computation` all yield the same stem `comput`) that we extract from tokens. At this point the abstract text is reduced to a set of stems and their grammatical functions.

4. Multi-stems construction: these are the basic semantic units used in further analysis. They are constructed as groups of successive stems in a sentence which satisfies a simple grammatical function rule. The length of the group is between 1 and 3 and its elements are either nouns, attributive verbs or adjectives. We choose to extract the semantics from such nominal groups in view of the technical nature of texts, which is not likely to contain subtle nuances in combinations of verbs and nominal groups.

Text processing operations are implemented in `python` in order to use built-in functions `nltk` library [30] for most of above operations. This library supports most of state-of-the-art natural language processing operations.[4]

3.2 Keywords relevance estimation

Relevance definition Following the heuristic in [29], we estimate relevance score in order to filter multi-stem. The choice of the total number of keywords to be extracted, which we shall denote K_w , is important, too small a value would yield similar network structures but including less information whereas very large values tend to include too many irrelevant keywords. We choose to set this parameter to $K_w = 100,000$. We first consider the filtration of $k \cdot K_w$ (with $k = 4$) to keep a large set of potential keywords but still have a reasonable number of co-occurrences to be computed. This step has only very marginal effects on the nature of the final keywords but is necessary for computational purposes. The filtration is done on the *unithood* u_i, defined for keyword i as $u_i = f_i \cdot \log(1 + l_i)$ where f_i is the multi-stem's number of apparitions over the whole corpus and l_i its length in words. A second filtration of K_w keywords is done on the *termhood* t_i, where the formal definition can be found in (1) . It is computed as a

[4]Source code is openly available on the repository of the project: `https://github.com/JusteRaimbault/PatentsMining`

chi-squared score on the distribution of the stem's co-occurrences and then compared to a uniform distribution within the whole corpus. Intuitively, uniformly distributed terms will be identified as plain language and they are thus not relevant for the classification. More precisely, we compute the co-occurrence matrix (M_{ij}), where M_{ij} is defined as the number of patents where stems i and j appear together. The *termhood* score t_i is defined as

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

Moving window estimation The previous scores are estimated on a moving window with fixed time length following the idea that the present relevance is given by the most recent context and thus that the influence vanishes when going further into the past. Consequently, the co-occurrence matrix is chosen to be constructed at year t restricting to patent which applied during the time window $\left[t - T_0; t\right]$. Note that the causal property of the window is crucial as the future cannot play any role in the current state of keywords and patents. This way, we will obtain semantic classes which are exploitable on a T_0 time span. For example, this enables us to compute the modularity of classes in the citation network as in section 4.3. In the following, we take $T_0 = 4$ (which corresponds to a five year window) consistently with the choice of maximum time lag for citations made in Section 2.2. Accordingly, the sensitivity analysis for $T_0 = 2$ can be found in Appendix S4 Text : Network Sensitivity Analysis.

3.3 Construction of the semantic network

We keep the set of most relevant keywords $\mathcal{K}_W$ and obtain their co-occurrence matrix as defined in Section 3.2. This matrix can be directly interpreted as the weighted adjacency matrix of the semantic network. At this stage, the topology of raw networks does not allow the extraction of clear communities. This is partly due to the presence of hubs that correspond to frequent terms common to many fields (e.g. `method`, `apparat`) which are wrongly filtered as relevant. We therefore introduce an additional measure to correct the network topology: the concentration of keywords across technological classes, defined as:

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

where $k_j(s)$ is the number of occurrences of the sth keyword in each of the jth technological class taken from one of the $N^{(tec)}$ USPC classes. The higher c_{tech}, the more specific to a technological class the node is. For example, the terms `semiconductor` is widely used in electronics and does not contain any significant information in this field. We use a threshold parameter, defined as θ_c, and keep nodes with $c_{tech}(s) > \theta_c$. Likewise, edges with low weights correspond to rare co-occurrences and are considered to be noise. To account for this we define the threshold parameter for edges θ_w, and we filter edges with a weight below θ_w , following the rationale that two keywords are not linked "by chance" if they appear simultaneously a minimal number of time. To control for size effect, we normalize by taking $\theta_w = \theta_w^{(0)} \cdot N_P$ where N_P is the number of patents in the corpus ($N_P = |\mathcal{P}|$). $\theta_w^{(0)}$ is thus a varying parameter interpreted as a noise threshold *per patent*. Communities are then extracted using a standard modularity maximization procedure as described in [31] to which we add the two constraints captured by θ_w and θ_c, namely that edges must have a weight greater than θ_w and nodes a concentration greater than θ_c. At this stage, both parameters θ_c and $\theta_w^{(0)}$ are unconstrained and their choice is not straightforward. Indeed, many optimization objectives are possible, such as the modularity, network size or number of communities. We find that modularity is maximized at a roughly stable value of θ_w across different θ_c for each year, corresponding to a stable $\theta_w^{(0)}$ across years, which leads us to choose $\theta_w^{(0)} = 4.1 \cdot 10^{-5}$. Then for the choice of θ_c, different candidates points lie on a Pareto front for the bi-objective optimization on number of communities and network size. There is a priori no reason to choose any specific point among the different optimums. Consequently, we have tried the analysis with all the candidate values for θ_c and found that the results are the most reasonable when taking $\theta_c = 0.06$ (see Fig. 1).

3.4 Characteristics of Semantic Classes

For each year t, we define as $N_t^{(sem)}$ the number of semantic classes which have been computed by clustering keywords from patents appeared during the period $\left[t - T_0, t\right]$ (we recall that we have chosen $T_0 = 4$). Each semantic class $k = 1, \cdots, N_t^{(sem)}$ is

Figure 1. Sensitivity analysis of network community structure to filtering parameters. We consider a specific window 2000-2004 and the obtained plots are typical. *(Left panel)* We plot the number of communities as a function of the edge threshold parameter θ_w for different values of the node threshold parameter θ_c. The maximum is roughly stable across θ_c (dashed red line). *(Right panel)* To choose θ_c, we do a Pareto optimization on communities and network size: the compromise point (red overline) on the Pareto front (purple overline: possible choices after having fixed $\theta_w^{(0)}$; blue level gives modularity) corresponds to $\theta_c = 0.06$.

Figure 2. An example of semantic network visualization. We show the network obtained for the window 2000-2004, with parameters $\theta_c = 0.06$ and $\theta_w = \theta_w^{(0)} \cdot N_P = 4.5e^{-5} \cdot 9.1e^5$. The corresponding file in a vector format (`.svg`), that can be zoomed and explored, is available as Supplementary Material.

characterized by a set of keywords $K(k,t)$ which is a subset of $\mathcal{K}_W$ selected as described in Section 3.1 to Section 3.3. The cardinal of $K(k,t)$ distribution across each semantic class k is highly skewed with a few semantic classes containing over $1,000$ keywords, most of them with roughly the same number of keywords. In contrast, there are also many semantic classes with only two keywords. There are around 30 keywords by semantic class on average and the median is 2 for any t. Fig. 3 shows that the average number of keywords is relatively stable from 1976 to 1992 and then picks around 1996 prior to going down.

Title of semantic classes USPC technological classes are defined by a title and a highly accurate definition which help retrieve patents easily. The title can be a single word (e.g.: class 101: "Printing") or more complex (e.g.: class 218: "High-voltage switches with arc preventing or extinguishing devices"). As our goal is to release a comprehensive database in which each patent is associated with a set of semantic classes, it is necessary to give an insight on what these classes represent by associating a short description or a title as in [26]. In our case, such description is taken as a subset of keywords taken from $K(k,t)$. For the vast majority of semantic classes that have less than 5 keywords, we decide to keep all of theses keywords as a description. For the remaining classes which feature around 50 keywords on average, we rely on the topological properties of the semantic network. [32] suggest to retain only the most frequently used terms in $K(k,t)$. Another possibility is to select 5 keywords based on

Figure 3. This figure plots the average number of keywords by semantic class for each time window $[t-4;t]$ from $t = 1980$ to $t = 2007$.

their network centrality with the idea that very central keywords are the best candidates to describe the overall idea captured by a community. For example, the largest semantic class in 2003-2007 is characterized by the keywords: `Support Packet; Tree Network; Network Wide; Voic Stream; Code Symbol Reader`.

Size of technological and semantic classes We consider a specific window of observations (for example 2000-2004), and we define Z the number of patents which appeared during that time window. For each patent $i = 1, \cdots, Z$ we associate a vector of probability where each component $p_{ij}^{(sem)} \in [0, 1]$, with $j = 1, \cdots, N(sem)$ and where[5]

$$\sum_{j=1}^{N^{(sem)}} p_{ij}^{(sem)} = 1.$$

On average across all time windows, a patent is associated to 1.8 semantic classes with a positive probability. Next we define the size of a semantic class as

$$S_j^{(sem)} = \sum_{i=1}^{Z} p_{ij}^{(sem)}.$$

Correspondingly, we aim to provide a consistent definition for technological classes. For that purpose, we follow the so-called "fractional count" method, which was introduced by the USPTO and consists in dividing equally the patents between all the classes they belong to. Formally, we define the number of technological classes as $N^{(tec)}$ (which is not time dependent contrary to the semantic case) and for $j = 1, \cdots, N^{(tec)}$ the corresponding matrix of probability is defined as

$$p_{ij}^{(tec)} = \frac{B_{ij}}{\sum_{k=1}^{N^{(tec)}} B_{ik}},$$

where B_{ij} equals 1 if the ith patent belongs to the jth technological class and 0 if not. When there is no room for confusion, we will drop the exponent part and write only p_{ij} when referring to either the technological or semantic matrix. Empirically, we find that both classes exhibit a similar hierarchical structure in the sense of a power-law type of distribution of class sizes as shown in Fig. 4. This feature is important, it suggests that

[5]When there is no room for confusion, we drop the subscript t in $N_t^{(sem)}$.

Figure 4. Sizes of classes. Yearly from $t = 1980$ to $t = 2007$, we plot the size of semantic classes (left-side) and technological classes (right-side) for the corresponding time window $[t-4, t]$, from the biggest to the smallest. The formal definition of size can be found in Section 3.4. Each color corresponds to one specific year. Yearly semantic classes and technological classes present a similar hierarchical structure which confirms the comparability of the two classifications. This feature is crucial for the statistical analysis in Section 4.4. Over time, curves are translated and levels of hierarchy stays roughly constant.

a classification based on the text content of patents has some separating power in the sense that it does not divide up all the patents in one or two communities.

3.5 Potential Refinements of the Method

Our semantic classification method could be refined by combining it with other techniques such as Latent Dirichlet Allocation which is a widely used topic detection method (e.g. [33]), already used on patent data as in [34] where it provides a measure of idea novelty and the counter-intuitive stylized facts that breakthrough invention are likely to come out of local search in a field rather than distant technological recombination. Using this approach should first help further evaluate the robustness of our qualitative conclusions (external validation). Also, depending on the level of orthogonality with our classification, it can potentially bring an additional feature to characterize patents, in the spirit of multi-modeling techniques where neighbor models are combined to take advantage of each point of view on a system.

Our use of network analysis can also be extended using newly developed techniques of hyper-network analysis. Indeed, patents and keywords can for example be nodes of a bipartite network, or patents be links of an hyper-network, in the sense of multiple layers with different classification links and citation links. The combination of citation network modeling by Stochastic Block Modeling with topic modeling was studied for scientific papers by [35] , outperforming previous link prediction algorithms. [36] provide a method to compare macroscopic structures of the different layers in a multilayer network that could be applied as a refinement of the overlap, modularity and statistical modeling studied in this paper. Furthermore, is has recently been shown that measures of multilayer network projections induce a significant loss of information compared to the generalized corresponding measure [37], which confirms the relevance of such development that we left for further research.

An other potential research development would be to further exploit the temporal structure of our dataset. Indeed, large progress have recently been made in complex network analysis of time-series data (see [38] for a review). For example, [39] develops a method to construct multiscale network from time series, which could in our case be a solution to identify structures in patents trajectories at different levels, and be an alternative to the single scale modularity analysis we use.

4 Results

In this section, we present some key features of our resulting semantic classification showing both complementary and differences with the technological classification. We first present several measures derived from this semantic classification at the patent level: Diversity, Originality, Generality (Section 4.1) and Overlapping (Section 4.2). We then show that the two classifications show highly different topological measures and strong statistical evidence that they feature a different model (Sections 4.3 and 4.4).

4.1 Patent Level Measures

Given a classification system (technological or semantic classes), and the associated probabilities p_{ij} for each patent i to belong to class j (that were defined in Section 3.4), one can define a patent-level diversity measure as one minus the Herfindhal concentration index on p_{ij} by

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

We show in Fig. 5 the distribution over time of semantic and technological diversity with the corresponding mean time-series. This is carried with two different settings, namely including/not including patents with zero diversity (i.e. single class patents). We call other patents "complicated patents" in the following. First of all, the presence of mass in small probabilities for semantic but not technological diversity confirms that the semantic classification contains patent spread over a larger number of classes. More interestingly, a general decrease of diversity for complicated patents, both for semantic and technological classification systems, can be interpreted as an increase in invention

Figure 5. Patent level diversities. Distributions of diversities (Left column) and corresponding mean time-series (Right column) for $t = 1980$ to $t = 2007$ (with the corresponding time window $[t-4, t]$). The first row includes all classified patents, whereas the second row includes only patents with more than one class (i.e. patents with diversity greater than 0).

specialization. This is a well-known stylized fact as documented in [40]. Furthermore, a qualitative regime shift on semantic classification occurs around 1996. This can be seen whether or not we include patents with zero diversity. The diversity of complicated patents stabilizes after a constant decrease, and the overall diversity begins to strongly decrease. This means that on the one hand the number of single class patents begins to increase and on the other hand complicated patents do not change in diversity. It can be interpreted as a change in the regime of specialization, the new regime being caused by more single-class patents.

More commonly used in the literature are the measures of originality and generality. These measures follow the same idea than the above-defined diversity in quantifying the diversity of classes (whether technological or semantic) associated with a patent. But instead of looking at the patent's classes, they consider the classes of the patents that are cited or citing. Formally, the originality O_i and the generality G_i of a patent i are defined as

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

where $z \in \{tec, sem\}$, I_i denotes the set of patents that are cited by the ith patent within a five year window (i.e. if the ith patent appears at year t, then we consider patents on $[t - T_0, t]$) when considering the originality and $\tilde{I}_i$ the set of patents that cite patent i after less than five years (i.e. we consider patents on $[t, t + T_0]$) in the case of generality. Note that the measure of generality is forward looking in the sense that $G_i^{(z)}$ used information that will only be available 5 years after patent applications. Both measures are lower on average based on semantic classification than on technological classification. Fig. 6 plots the mean value of $O_i^{(sem)}$, $O_i^{(tec)}$, $G_i^{(sem)}$ and $G_i^{(tec)}$.

Figure 6. Patent level originality (left hand side) and **generality** (right hand side) for $t = 1980$ to $t = 2007$ (with the corresponding time window $[t-4, t]$) as defined in subsection 4.1.

4.2 Classes overlaps

A proximity measure between two classes can be defined by their overlap in terms of patents. Such measures could for example be used to construct a metrics between semantic classes. Intuitively, highly overlapping classes are very close in terms of technological content and one can use them to measure distance between two firms in terms of technology as done in [41]. Formally, recalling the definition of (p_{ij}) as the probability for the ith patent to belong to the jth class and N_P as the number of patents it writes

$$Overlap_{jk} = \frac{1}{N_P} \cdot \sum_{i=1}^{N_P} p_{ij} p_{ik}. \quad (2)$$

The overlap is normalized by patent count to account for the effect of corpus size: by convention, we assume the overlap to be maximal when there is only one class in the corpus. A corresponding relative overlap is computed as a set similarity measure in the number of patents common to two classes A and B, given by $o(A, B) = 2 \cdot \frac{|A \cap B|}{|A|+|B|}$.

Intra-classification overlaps The study of distributions of overlaps inside each classification, i.e. between technological classes and between semantic classes separately, reveals the structural difference between the two classification methods, suggesting their complementary nature. Their evolution in time can furthermore give insights into trends of specialization. We show in Fig. 7 distributions and mean time-series of overlaps for the two classifications. The technological classification globally always follow a decreasing trend, corresponding to more and more isolated classes, i.e. specialized inventions, confirming the stylized fact obtained in previous subsection. For semantic classes, the dynamic is somehow more intriguing and supports the story of a qualitative regime shift suggested before. Although globally decreasing as technological overlap, normalized (resp. relative) mean overlap exhibits a peak (clearer for normalized overlap) culminating in 1996 (resp. 1999). Looking at normalized overlaps, classification structure was somewhat stable until 1990, then strongly increased to peak in 1996 and then decrease at a similar pace up to now. Technologies began to share more and more until a breakpoint when increasing isolation became the rule again. An evolutionary perspective on technological innovation [42] could shed light on possible interpretations of this regime shift: as species evolve, the fitness landscape first would have been locally favorable to cross-insemination, until each fitness reaches a threshold above which auto-specialization becomes the optimal path. It is very comparable to the establishment of an ecological niche [43], the strong interdependency originating here during the mutual insemination resulting in a highly path-dependent final situation.

Figure 7. Intra-Classification overlaps. (Left column) Distribution of overlaps O_{ij} for all $i \neq j$ (zero values are removed because of the log-scale). Right column) Corresponding mean time-series. (First row) Normalized overlaps. (Second row) Relative overlaps.

Inter-classification overlaps Overlaps *between* classifications are defined as in (2), but with j standing for the jth technological class and k for the kth semantic class: p_{ij} are technological probabilities and p_{ik} semantic probabilities. They describe the relative correspondence between the two classifications and are a good indicator to spot relative changes, as shown in Fig. 8. Mean inter-classification overlap clearly exhibits two linear trends, the first one being constant from 1980 to 1996, followed by a constant decrease. Although difficult to interpret directly, this stylized fact clearly unveils a change in the *nature* of inventions, or at least in the relation between content of inventions and technological classification. As the tipping point is at the same time as the ones observed in the previous section and since the two statistics are different, it is unlikely that this is a mere coincidence. Thus, these observations could be markers of a hidden underlying structural changes in processes.

Figure 8. Distribution of relative overlaps between classifications. (Left) Distribution of overlaps at all time steps; (Right) Corresponding mean time-series. The decreasing trend starting around 1996 confirms a qualitative regime shift in that period.

4.3 Citation Modularity

An exogenous source of information on relevance of classifications is the citation network described in Section 2.2. The correspondence between citation links and classes should provide a measure of accuracy of classifications, in the sense of an external validation since it is well-known that citation homophily is expected to be quite high (see, e.g, [14]). This section studies empirically modularities of the citation network regarding the different classifications. To corroborate the obtained results, we propose to look at a more rigorous framework in Section 4.4. Modularity is a simple measure of how communities in a network are well clustered (see [31] for the accurate definition). Although initially designed for single-class classifications, this measure can be extended to the case where nodes can belong to several classes at the same time, in our case with different probabilities as introduced in [44]. The simple directed modularity is given in our case by

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

with A_{ij} the citation adjacency matrix (i.e. $A_{ij} = 1$ if there is a citation from the ith patent to the jth patent, and $A_{ij} = 0$ if not), $k_i^{in} = |I_i|$ (resp. $k_i^{out} = \left|\tilde{I}_i\right|$) in-degree (resp. out-degree) of patents (i.e. the number of citations made by the ith patent to others and the number of citations received by the ith patent). Q_d can be defined for each of the two classification systems: $z \in \{tec, sem\}$. If $z = tec$, c_i is defined as the main patent class, which is taken as the first class whereas if $z = sem$, c_i is the class with the largest probability.

Multi-class modularity in turns is given by

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

where

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

We take $F(p_{ic}, p_{jc}) = p_{ic} \cdot p_{jc}$ as suggested in [44]. Modularity is an aggregated measure of how the network deviates from a null model where links would be randomly made according to node degree. In other words it captures the propensity for links to be inside the classes. Overlapping modularity naturally extends simple modularity by taking into account the fact that nodes can belong simultaneously to many classes. We document in Fig. 9 both simple and multi-class modularities over time. For simple modularity,

$Q_d^{(tec)}$ is low and stable across the years whereas $Q_d^{(sem)}$ is slightly greater and increasing. These values are however low and suggest that single classes are not sufficient to capture citation homophily. Multi-class modularities tell a different story. First of all, both classification modularities have a clear increasing trend, meaning that they become more and more adequate with citation network. The specializations revealed by both patent level diversities and classes overlap is a candidate explanation for this growing modularities. Secondly, semantic modularity dominates technological modularity by an order of magnitude (e.g. 0.0094 for technological against 0.0853 for semantic in 2007) at each time. This discrepancy has a strong qualitative significance. Our semantic classification fits better the citation network when using multiple classes. As technologies can be seen as a combination of different components as shown by [5], this heterogeneous nature is most likely better taken into account by our multi-class semantic classification.

Figure 9. Temporal evolution of semantic and technological modularities of the citation network. (Left) Simple directed modularity, computed with patent main classes (main technological class and semantic class with larger probability). (Right) Multi-class modularity, computed following [44]

4.4 Statistical Model

In this section, we develop a statistical model aimed at quantifying performance of both technological and semantic classification systems. In particular, we aim at corroborating findings obtained in Section 4.3. The mere difference between this approach and the citation modularity approach lies in the choice of the underlying model, and the according quantities of interest. In addition for the semantic approach, we want to see if when restricting to patents with higher probabilities to belong to a class, we obtain better results. To do that, we choose to look at within class citations proportion (for both technological and semantic approaches). We provide two obvious reasons why we choose this. First, the citations are commonly used as a proxy for performance as mentioned in Section 4.3. Second, this choice is "statistically fair" in the sense that both approaches have focused on various goals and not on maximizing directly the within class proportion. Nonetheless, the within class proportion is too sensitive to the distribution of the shape of classes. For example, a dataset where patents for each class account for 10% of the total number of patents will mechanically have a better within class proportion than if each class accounts for only 1%. Consequently, an adequate statistical model, which treats datasets fairly regardless of their distribution in classes, is needed. This effort ressembles to the previous study of citation modularity, but is complementary since the model presented here can be understood as an elementary model of citation network growth. Furthermore, the parameters fitted here can have a direct interpretation as a citation probability.

We need to introduce and recall some notations. We consider a specific window of observations $\big[t - T_0, t\big]$, and we define Z the number of patents which appeared during that time window. We let $t_1, \cdots, t_Z$ their corresponding appearance date by chronological order, which for simplicity are assumed to be such that $t_1 < \cdots < t_Z$. For each patent $i = 1, \cdots, Z$ we consider C_i the number of distinctive couples {cited patent, cited patent's class} made by the ith patent (for instance if the ith patent has only made one citation and that the cited patent is associated with three classes, then $C_i = 3$). Let $z \in \{tec, sem\}$, we define $N_i^{(z)}$ the number of patents associated to at least one of the ith classes at time t_{i-1}. For $l = 1, \cdots, C_i$ we consider the variables $B_{l,i}$, which equal 1 if the cited patent's class is also common to the ith patent. We assume that $B_{l,i}$ are independent of each other and conditioned on the past follow Bernoulli variables

$$B\Big(\min\Big\{1, \frac{N_i^{(z)}}{i-1} + \theta^{(z)}\Big\}\Big),$$

where the parameter $0 \leq \theta^{(z)} \leq 1$ indicates the propensity for any patent to cite patents of its own technological or semantic class. When $\theta^{(z)} = 0$, the probability of citing patents from its own class is simply $N_i^{(z)}(i-1)^{-1}$, which corresponds to the observed proportion of patents which belong to at least one of the ith patent's classes. Thus this corresponds to the estimated probability of citing one patent if we assume that the probability of citing any patent $k = 1, \cdots, i-1$ is uniformly distributed, which could be a reasonable assumption if classes were assigned randomly and independently from patent abstract contents. Conversely if $\theta^{(z)} = 1$, we are in the case of a model where there are 100% of within class citations. A reasonable choice of $\theta^{(z)}$ lies between those two extreme values. Finally, we assume that the number of distinctive couples C_i are a sequence of independent and identically distributed random variables following the discrete distribution C, and also independent from the other quantities.

We estimate $\theta^{(z)}$ via maximum likelihood, and obtain the corresponding maximum likelihood estimator (MLE) $\hat{\theta}^{(z)}$. The likelihood function, along with the standard deviation expression and details about the test, can be found in Supporting Information. The fitted values, standard errors and p-values corresponding to the statistical test $\theta^{(sem)} = \theta^{(tec)}$ (with corresponding alternative hypothesis $\theta^{(sem)} > \theta^{(tec)}$) on non-overlapping blocks from the period 1980-2007[6] are reported on Table 1. Semantic values are reported for four different chosen thresholds $p^- = .04, .06, .08, .1$. It means that we restricted to the couples (ith patent, jth class) such that $p_{ij} \geq p^-$.

The choice of considering non-overlapping blocks (instead of overlapping blocks) is merely statistical. Ultimately, our interest is in the significance of the test over the whole period 1980-2007. Thus, we want to compute a global p-value. This can be done considering the local p-values (by local, we mean for instance computed on the period 2001-2005) assuming independence between them. This assumption is reasonable only if the blocks are non-overlapping. All of this can be found in Supporting Information. Finally, note that from a statistical perspective, including overlapping blocks wouldn't yield more information.

The values reported in Table 1 are overwhelmingly against the null hypothesis. The global estimates of $\theta^{(sem)}$ are significantly bigger than the estimate of $\theta^{(tec)}$ for all the considered thresholds. Although the corresponding p-values (which are also very close to 0) are not reported, it is also quite clear that the bigger the threshold, the higher the corresponding $\theta^{(sem)}$ is estimated. This is consistently seen for any period, and significant for the global period. This seems to indicate that when restricting to the couples (patent, class) with high semantic probability, the propension to cite patents from its own class $\theta^{(sem)}$ is increasing. We believe that this might provide extra information to patent officers when making their choice of citations. Indeed, they could look first to patents which belong to the same semantic class, especially when patents have high probability semantic values.

Note that the introduced model can be seen as a simple model of citations network growth conditional to a classification, which can be expressed as a stochastic block model (e.g. [45], [46]). The parameters are estimated computing the corresponding MLE. In

[6] Note that the estimation included patents up until 2010 in the period 2006-2007 and not the patents from 1980 in the period 1980-1985 for homogeneity in size with other periods. This doesn't affect the significativity of the results.

view of [47], this can be thought as equivalent to maximizing modularity measures.

Table 1. Estimated values of $\theta^{(tec)}$ and $\theta^{(sem)}$ and corresponding standard errors obtained from a Maximum Likelihood estimator as presented in section 4.4.

Approach	Estimated Value	st. er.	p-value
1980-1985 period			
technological	.664	.008	
semantic $p^- = .04$	.741	.047	.053
semantic $p^- = .06$	.799	.081	.049
semantic $p^- = .08$	.828	.126	.097
semantic $p^- = .10$	.834	.166	.153
1986-1990 period			
technological	.634	.007	
semantic $p^- = .04$	.703	.022	.001
semantic $p^- = .06$	.768	.040	.0004
semantic $p^- = .08$	.804	.069	.007
semantic $p^- = .10$	.832	.114	.041
1991-1995 period			
technological	.619	.006	
semantic $p^- = .04$	.655	.009	.0004
semantic $p^- = .06$	.713	.017	9e-08
semantic $p^- = .08$	.731	.025	7e-06
semantic $p^- = .10$	.750	.037	9e-06
1996-2000 period			
technological	.551	.003	
semantic $p^- = .04$	.585	.002	≈ 0
semantic $p^- = .06$	.638	.004	≈ 0
semantic $p^- = .08$	.660	.006	≈ 0
semantic $p^- = .10$	.686	.008	≈ 0
2001-2005 period			
technological	.567	.003	
semantic $p^- = .04$	.621	.004	≈ 0
semantic $p^- = .06$	.676	.007	≈ 0
semantic $p^- = .08$	.701	.010	≈ 0
semantic $p^- = .10$	.710	.013	≈ 0
2006-2007 period			
technological	.600	.007	
semantic $p^- = .04$	.683	.016	1e-06
semantic $p^- = .06$	.732	.025	2e-07
semantic $p^- = .08$	.760	.036	6e-06
semantic $p^- = .10$	.782	.048	9e-05
1980-2007 global period			
technological	.606	.002	
semantic $p^- = .04$	.665	.009	8e-11
semantic $p^- = .06$	.721	.017	9e-12
semantic $p^- = .08$	.747	.025	9e-09
semantic $p^- = .10$	.782	.035	3e-07

5 Conclusion

The main contribution of this study was twofold. First we have defined how we built a network of patents based on a classification that uses semantic information from abstracts. We have shown that this classification share some similarities with the traditional technological classification, but also have distinct features. Second, we provide researchers with materials resulting from our analysis, which includes: (i) a database linking each patent with its set of semantic classes and the associated probabilities; (ii) a list of these semantic classes with a description based on the most relevant keywords; (iii) a list of patent with their topological properties in the semantic network (centrality, frequency, degree, etc.). The availability of this data suggests new avenues for further research. Linking our dataset with existing open ones can lead to various powerful developments. For example, using it together with the disambiguated inventor database provided by [48] could be a way to study semantic profiles of inventors, or of cities as inventor addresses are provided. The investigation of spatial diffusion of innovation between cities, which is a key component of Pumain's Evolutive Urban Theory [49] , would be made possible.

A first potential application is to use the patents' topological measures inherited from their relevant keywords. The fact that these measures are backward-looking and immediately available after the publication of the patent information is an important asset. It would for example be very interesting to test their predicting power to assess the quality of an innovation, using the number of forward citations received by a patent, and subsequently the future effect on the firm's market value.

Regarding firm innovative strategy, a second extension could be to study trajectories of firms in the two networks: technological and semantic. Merging these information with data on the market value of firms can give a lot of insight about the more efficient innovative strategies, about the importance of technology convergence or about acquisition of small innovative firms. It will also allow to observe innovation pattern over a firm life cycle and how this differ across technology field.

A third extension would be to use dig further into the history of innovation. USPTO patent data have been digitized from the first patent in July 1790. However, not all of them contain a text that is directly exploitable. We consider that the quality of patent's images is good enough to rely on Optical Character Recognition techniques to retrieve plain text from at least 1920. With such data, we would be able to extend our analysis further back in time and to study how technological progress occurs and combines in time. [50] conduct a similar work by looking at recombination and apparition of technological subclasses. Using the fact that communities are constructed yearly, one can construct a measure of proximity between two successive classes. This could give clear view on how technologies converged over the year and when others became obsolete and replaced by new methods.

Supporting Information

S1 Text : Definition of utility patent

Describes with more details the definition of patents and context.

S2 Text : Data collection procedure

Detailed description of data collection

S3 File : Semantic Network Visualization

Vector file of the semantic network (Fig.2)

S4 Text : Network Sensitivity Analysis

Extended figures for Network Sensitivity Analysis

S5 Text : Statistical definitions and derivations

Extended definitions and derivations for the statistical model

References

1. Aghion P, Howitt P. A Model of Growth through Creative Destruction. Econometrica. 1992 March;60(2):323–51. Available from: https://ideas.repec.org/a/ecm/emetrp/v60y1992i2p323-51.html.

2. Romer PM. Endogenous Technological Change. Journal of Political Economy. 1990 October;98(5):S71–102. Available from: https://ideas.repec.org/a/ucp/jpolec/v98y1990i5ps71-102.html.

3. Griliches Z. Patent Statistics as Economic Indicators: A Survey. National Bureau of Economic Research, Inc; 1990. 3301. Available from: https://ideas.repec.org/p/nbr/nberwo/3301.html.

4. Hall BH, Jaffe AB, Trajtenberg M. The NBER Patent Citations Data File: Lessons, Insights and Methodological Tools. C.E.P.R. Discussion Papers; 2001. 3094. Available from: https://ideas.repec.org/p/cpr/ceprdp/3094.html.

5. Youn H, Strumsky D, Bettencourt LMA, Lobo J. Invention as a combinatorial process: evidence from US patents. Journal of The Royal Society Interface. 2015;12(106).

6. Newman MEJ. Prediction of highly cited papers. ArXiv e-prints. 2013 Oct;.

7. Sarigöl E, Pfitzner R, Scholtes I, Garas A, Schweitzer F. Predicting Scientific Success Based on Coauthorship Networks. ArXiv e-prints. 2014 Feb;.

8. Sorenson O, Rivkin JW, Fleming L. Complexity, networks and knowledge flow. Research policy. 2006;35(7):994–1017.

9. Kay L, Newman N, Youtie J, Porter AL, Rafols I. Patent overlay mapping: Visualizing technological distance. Journal of the Association for Information Science and Technology. 2014;65(12):2432–2443.

10. Bruck P, Réthy I, Szente J, Tobochnik J, Érdi P. Recognition of emerging technology trends: class-selective study of citations in the US Patent Citation Network. Scientometrics. 2016;107(3):1465–1475.

11. Curran CS, Leker J. Patent indicators for monitoring convergence–examples from NFF and ICT. Technological Forecasting and Social Change. 2011;78(2):256–273.

12. Katz ML. Remarks on the economic implications of convergence. Industrial and Corporate Change. 1996;5(4):1079–1095.

13. Furman JL, Stern S. Climbing atop the Shoulders of Giants: The Impact of Institutions on Cumulative Research. American Economic Review. 2011 August;101(5):1933–63. Available from: http://www.aeaweb.org/articles?id=10.1257/aer.101.5.1933.

14. Acemoglu AU Daron, Kerr W. Innovation Network. Proceedings of the National Academy of Sciences (forthcoming). 2016;.

15. Preschitschek N, Niemann H, Leker J, Moehrle MG. Anticipating industry convergence: Semantic analyses vs IPC co-classification analyses of patents. Foresight. 2013 11;15(6):446–464.

16. Yoon B, Park Y. A text-mining-based patent network: Analytical tool for high-technology trend. The Journal of High Technology Management Research. 2004;15(1):37–50.

17. Park I, Yoon B. A semantic analysis approach for identifying patent infringement based on a product–patent map. Technology Analysis & Strategic Management. 2014;26(8):855–874.

18. Yoon J, Kim K. Detecting signals of new technological opportunities using semantic patent analysis and outlier detection. Scientometrics. 2011;90(2):445–461.

19. Gerken JM, Moehrle MG. A new instrument for technology monitoring: novelty in patents measured by semantic patent analysis. Scientometrics. 2012;91(3):645–670.

20. Choi J, Hwang YS. Patent keyword network analysis for improving technology development efficiency. Technological Forecasting and Social Change. 2014;83:170–182.

21. Fattori M, Pedrazzi G, Turra R. Text mining applied to patent mapping: a practical business case. World Patent Information. 2003;25(4):335–342.

22. Gurciullo S, Smallegan M, Pereda M, Battiston F, Patania A, Poledna S, et al. Complex Politics: A Quantitative Semantic and Topological Analysis of UK House of Commons Debates. ArXiv e-prints. 2015 Oct;.

23. Lerner J, Seru A. The use and misuse of patent data: Issues for corporate finance and beyond. Booth/Harvard Business School Working Paper. 2015;.

24. OECD. OECD Patent Statistics Manual. 2009;Available from: /content/book/9789264056442-en.

25. DechezleprÃªtre A, Martin R, Mohnen M. Knowledge Spillovers from Clean and Dirty Technologies. Centre for Economic Performance, LSE; 2014. dp1300. Available from: https://ideas.repec.org/p/cep/cepdps/dp1300.html.

26. Tseng YH, Lin CJ, Lin YI. Text mining techniques for patent analysis. Information Processing & Management. 2007;43(5):1216–1247.

27. Adams S. The text, the full text and nothing but the text: Part 1 - Standards for creating textual information in patent documents and general search implications. World Patent Information. 2010 March;32(1):22–29. Available from: https://ideas.repec.org/a/eee/worpat/v32y2010i1p22-29.html.

28. Abbas A, Zhang L, Khan SU. A literature review on the state-of-the-art in patent analysis. World Patent Information. 2014;37:3–13.

29. Chavalarias D, Cointet JP. Phylomemetic patterns in science evolution—the rise and fall of scientific fields. Plos One. 2013;8(2):e54847.

30. NLTK. Natural Language Toolkit, Stanford Univeristy; 2015.

31. Clauset A, Newman ME, Moore C. Finding community structure in very large networks. Physical review E. 2004;70(6):066111.

32. Yang Y, Ault T, Pierce T, Lattimer CW. Improving text categorization methods for event tracking. In: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval. ACM; 2000. p. 65–72.

33. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. Journal of machine Learning research. 2003;3(Jan):993–1022.

34. Kaplan S, Vakili K. The double-edged sword of recombination in breakthrough innovation. Strategic Management Journal. 2015;36(10):1435–1457.

35. Zhu Y, Yan X, Getoor L, Moore C. Scalable text and link analysis with mixed-topic link models. In: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM; 2013. p. 473–481.

36. Iacovacci J, Wu Z, Bianconi G. Mesoscopic Structures Reveal the Network Between the Layers of Multiplex Datasets. arXiv preprint arXiv:150503824. 2015;.

37. De Domenico M, Solé-Ribalta A, Omodei E, Gómez S, Arenas A. Ranking in interconnected multilayer networks reveals versatile nodes. Nature communications. 2015;6.

38. Gao ZK, Small M, Kurths J. Complex network analysis of time series. EPL (Europhysics Letters). 2017;116(5):50001.

39. Gao ZK, Yang YX, Fang PC, Zou Y, Xia CY, Du M. Multiscale complex network for analyzing experimental multivariate time series. EPL (Europhysics Letters). 2015;109(3):30005.

40. Archibugi D, Pianta M. Specialization and size of technological activities in industrial countries: The analysis of patent data. Research Policy. 1992;21(1):79 – 93. Available from: http://www.sciencedirect.com/science/article/pii/004873339290028З.

41. Bloom N, Schankerman M, Reenen JV. Identifying Technology Spillovers and Product Market Rivalry. Econometrica. 2013 07;81(4):1347–1393. Available from: https://ideas.repec.org/a/ecm/emetrp/v81y2013i4p1347-1393.html.

42. Ziman J. Technological innovation as an evolutionary process. Cambridge University Press; 2003.

43. Holland JH. Signals and boundaries: Building blocks for complex adaptive systems. Mit Press; 2012.

44. Nicosia V, Mangioni G, Carchiolo V, Malgeri M. Extending the definition of modularity to directed graphs with overlapping communities. Journal of Statistical Mechanics: Theory and Experiment. 2009;2009(03):P03024.

45. Decelle A, Krzakala F, Moore C, Zdeborová L. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. Physical Review E. 2011;84(6):066106.

46. Valles-Catala T, Massucci FA, Guimera R, Sales-Pardo M. Multilayer stochastic block models reveal the multilayer structure of complex networks. Physical Review X. 2016;6(1):011036.

47. Newman MEJ. Community detection in networks: Modularity optimization and maximum likelihood are equivalent. ArXiv e-prints. 2016 Jun;.

48. Li GC, Lai R, D'Amour A, Doolin DM, Sun Y, Torvik VI, et al. Disambiguation and co-authorship networks of the US patent inventor database (1975–2010). Research Policy. 2014;43(6):941–955.

49. Pumain D. Une théorie géographique des villes. Bulletin de la Société géographie de Liège. 2010;(55):5–15.

50. Akcigit U, Kerr WR, Nicholas T. The Mechanics of Endogenous Innovation and Growth: Evidence from Historical US Patents. Citeseer; 2013.

Editor PLOS ONE April 3, 2017

Dear Editor,

Thank you for considering our manuscript "Classifying Patents Based on their Semantic Content" for possible publication in PLOS ONE. We are also very grateful to your suggestions and comments. This will undoubtedly be of great value to the paper.

We have read carefully your suggestions and comments, and have updated the paper accordingly. We provide you now the point-by-point response to the Editor and referees' reports.

We deal first with the Editor's comments.

1. We made adjustments to fully meet PLOS ONE requirements. All figures were converted and assessed using the PACE tool.
2. "Authors used co-occurrence of keywords to construct a patent network. Is this a new way? Or at least a discussion of the advantages should be provided."
   → The use of co-occurrences to construct a semantic network has already been used, and is the best way to extract the endogenous semantic structure. We added a discussion on this point.
3. "Authors introduced a measure to correct the network topology. But how to properly determine the threshold? Why choose the value 0.06 ?"
   → Thank you for pointing out this issue. We have been more specific about it now, adding three sentences. Indeed, the explanation at the end of Section 3.3 was sloppy.
4. "The research background include complex network analysis, community detection and data analysis, some recent progress in these areas should be reviewed"
   → We included some of the references suggested in Section 3.5, as time series complex network analysis is indeed an interesting potential development.
5. "All the parameters should be clearly explained."
   → Thank you for also pointing this out. We have been clearer on the definition of K_w, k and t_i (Section 3.2), θ_c, θ_w and $\theta_w^{(0)}$ (Section 3.3).

We also provide you responses to Reviewer #3. Reviewers #1 and #2 did not ask anything specific but their comments were taken into account in the adjustments we made.

1. "Maybe a minor typo in Page 24 - "availability of these data" –> "this data" or "these datasets"".
   → We corrected accordingly.
2. "I also did not see a caption for the figures and it is quite hard to read the text in the figures due to their current small size. New readers are attracted to tables and figures, and thus it is useful to have descriptive captions - the current caption for Table 1 does not describe the variables being used (I know the text does it) - a brief description of theta being the likelihood estimates (or similar) would be useful."
   → We have made the according changes. All captions inside figures were magnified as large as possible to ensure readability.
3. "I like the authors' approach overall, but would also recommend adding some discussion on how a semantic approach enables information integration and reuse - possibly with how their dataset /

ontology can be linked to others already existing in Linked Open Data. If such linking already exists, it should be shown - otherwise, this is a strong direction for future work."
$\rightarrow$ This is indeed a very good suggestion and we add ideas of interesting potential developments by joining our database with existing open databases. We added accordingly a part to the discussion.

Yours faithfully,
Juste Raimbault
Université Paris 7 - UMR CNRS 8504 Géographie-cités

Click here to access/download
Data Review URL
http://dx.doi.org/10.7910/DVN/BW3ACK

Click here to access/download
Data Review URL
http://dx.doi.org/10.7910/DVN/ZULMOY

Data Review URL
https://github.com/JusteRaimbault/PatentsMining