Long-run dynamics of the U.S. patent classification system *

François Lafond[1,2] and Daniel Kim[3]

[1] *Institute for New Economic Thinking at the Oxford Martin School, University of Oxford*
[2] *Smith School for Enterprise and the Environment, University of Oxford*
[3] *Natural Science Research Institute, Korea Advanced Institute of Science and Technology*

February 27, 2017

Abstract

Almost by definition, radical innovations create a need to revise existing classification systems. As a result, the evolution of technological classification systems reflects technological evolution. We present three sets of findings regarding classification volatility in the U.S. Patent Classification System. First, we study the evolution of the number of distinct classes. Reconstructed time series based on the current classification scheme are very different from historical data. This suggests that using the current classification to analyze the past produces a distorted view of the evolution of the system. Second, we study the relative sizes of classes. The size distribution is exponential so classes can be of quite different sizes, but the largest classes are not necessarily the oldest. To explain this pattern with a simple stochastic growth model, we introduce the assumption that classes have a regular chance to be split. Third, we study reclassification. The share of patents that are in a different class now than they were at birth can be quite high. Reclassification mostly occurs across classes belonging to the same 1-digit NBER category, but not always. We also document that reclassified patents tend to be more cited than non-reclassified ones, even after controlling for grant year and class of origin. More generally we argue that classification changes and patent reclassification are quite common, reveal interesting information about technological evolution, and must be taken into account when using classification as a basis for forecasting.

Keywords: patents, classification, reclassification.
JEL codes: O30, O39.

1 Introduction

The U.S. patent system contains around 10 million patents classified in about 500 main classes. However, some classes are much larger than others, some classes are much older than others, and more importantly none of these classes can be thought of as a once-and-for-all well defined entity. Due to its important legal role, the U.S. Patent and Trademark Office (USPTO) has constantly devoted resources to improve the classification of inventions, so that the classification system has greatly evolved over time, reflecting contemporaneous technological evolution. Classification evolves because new classes are created but also because existing classes are abolished, merged and split. In fact, all current classes in 2015 have been established in the U.S. Patent Classification System (USPCS) after 1899, even though the first patent was granted in 1790 and the first classification system was created in 1829-1830. To give just another example, out of all patents granted in 1976, 40% are in a different main class now than they were in 1976.

*This paper reuses material from an unpublished chapter of the first author PhD thesis at UNU-MERIT, Maastricht University. This work was supported by the European Commission project FP7-ICT-2013-611272 (GROWTHCOM), the National Research Foundation of Korea (NRF) grant funded by the Korea government (No. 2011-0028908), and by the Institute for New Economic Thinking at the Oxford Martin School. We have benefited from excellent comments from many colleagues, including Jeff Alstott, Yuki Asano, Mariano Beguerisse Díaz, J. Doyne Farmer, Marco Pangallo, Emanuele Pugliese, and Giorgio Triulzi. We are also grateful to Diana Greenwald and the USPTO for helping us locating data sources. All remaining errors are ours. Contacts: francois.lafond@inet.ox.ac.uk, daniel.youngho.kim@gmail.com

To maintain the best possible level of searchability, the USPTO reclassifies patents so that at every single moment in time the patents are classified according to a coherent, up-to-date taxonomy. The downside of this is that the current classification is not meant to reflect the historical description of technological evolution as it unfolded. In other words, while the classification system provides a consistent classification of all the patents, this consistency is not time invariant. Observers at different points in time have a different idea of what is a consistent classification of the past, even when classifying the same set of past patents. In this paper, we focus on the historical evolution of the U.S. patent classification. We present three sets of findings.

First we study the evolution of the number of distinct classes, contrasting current and historical classification systems. Recent studies (Strumsky et al. 2012, Strumsky & Lobo 2015, Youn et al. 2015) have shown that it is possible to reconstruct the long-run evolution of the number of subclasses using the current classification system. This allowed them to obtain interesting results on the types of recombinations and on the relative rates of introduction of new subclasses and new combinations. An alternative way to count the number of distinct categories is to go back to the archives and to check how many classes did actually exist at different points of the past. We found important differences between the historical and reconstructed evolution of the classification system. In particular, we find that historically the growth of the number of distinct classes has been more or less linear, with about 2.5 classes added per year. By contrast, the reconstructed evolution – which considers how many current classes are needed to classify all patents granted before a given date -– suggests a different pattern with most classes created in the XIXth century and a slowdown in the rate of introduction of novel classes afterwards. Similarly, using the historical classes we find that relationship between the number of classes and the number of patents is compatible with Heaps' law, a power law scaling of the number of categories with the number of items, originally observed between the number of different words and the total number of words in a text (Heaps 1978). Using the reconstructed evolution Heaps' law does not hold over the long run.

Knowing the number of distinct classes, the next question is about their growth and relative size (in terms of the number of patents). Thus our second set of findings concerns the size distribution of classes. We find that it is exponential, confirming a result of Carnabuci (2013) on a much restricted sub-sample. We also find that there is no clear relationship between the size and the age of classes, which rules out an explanation of the exponential distribution in terms of simple stochastic growth models in which classes are created once and for all.

Third, we hypothesize that new technology fields and radical innovations tend to be associated with a higher reclassification activity. This suggests that the history of reclassification contains interesting information on the most transformative innovations. Our work here is related to Wang et al. (2016) who study how a range of metrics (claims, references, extensions, etc.) correlate with reclassification for 3 million utility patents since 1994. We used the data since 1976, for which we observe the class of origin and the citations statistics. It appears that reclassified patents are more cited than non-reclassified patents. We also construct a reclassification flow diagram, with aggregation at the level of NBER patent categories (Hall et al. 2001). This reveals that a non-negligible share of patents are reclassified across NBER categories. We find that patents in "Computers" and in "Electronics" are often reclassified in other NBER categories, which is not the case of other categories such as "Drugs".

Finally, we argue that it is not possible to explain the observed patterns without accounting for reclassification. We develop a simple model in which classes grow according to preferential attachment but have a probability of being split. The model's only inputs are the number of patents and classes in 2015 and the Heaps' law exponent. Despite this extreme parsimony, the model is able to reproduce i) the historical and reconstructed patterns of growth of the number of classes, ii) the size distribution and (partially) the lack of age-size relationship, and iii) the time evolution of the reclassification rates.

The empirical evidence that we present and the assumptions we need to make for the model make it clear that the U.S. patent classification system has evolved considerably and it is hardly possible to think of patent classes as technological domains with a stable definition. The classification system cannot be well understood as a system in which categories are created once-and-for-all and accumulate patents over time. Instead, it is better understood as a system that is constantly re-

organized. Because of this, using the current classification system to study a set of older patents is akin to looking at the past with today's glasses. In this paper, not only we show the differences between the historical and reconstructed reality, but we explain how these differences emerged.

The paper is organized as follows. Section 2 details our motivation, gives some background on categorization and reviews the literature on technological categories. Section 3 describes the USPCS and our data sources. Section 4 presents our results on the evolution of the number of classes. Section 5 discusses the size distribution of classes. Section 6 presents our results on reclassification since 1976. Section 7 presents a model that reproduces the main empirical patterns discovered in the previous sections. The last section discusses the results, motivates further research and concludes.

2 Why is studying classification systems important?

Classification systems are pervasive because they are extremely useful. At a fundamental level, categorization is at the basis of pattern recognition, learning, and sense-making. Producing a discourse regarding technologies and their evolution is no exception. As a matter of fact, theoretical and *a fortiori* empirical studies almost always rely on some sort of grouping – or aim at defining one.

Historically, the interest in technology classifications has been mostly driven by the need to match technological and industrial activities (Schmookler 1966, Scherer 1984, Verspagen 1997). Since patented technologies are classified according to their function, not their industry of use or origin, this problem is particularly difficult. Clearly, a good understanding of both industry and patent classification systems is crucial to build a good crosswalk. Here we highlight the need to acknowledge that both classification systems *change*. For this reason our results give a strong justification for automated, probabilistic, data-driven approaches to the construction of concordance tables such as the recent proposal by Lybbert & Zolas (2014) which essentially works by looking for keywords of industry definitions in patents to construct technology-industry tables.

With the rise of interest in innovation itself many studies have used existing patent classifications to study spillovers across technology domains, generally considering classification as static. For instance Kutz (2004) studied the growth and distribution of patent classes since 1976; Leydesdorff (2008), Antonelli et al. (2010), Strumsky et al. (2012) and Youn et al. (2015) studied co-classification patterns; and Caminati & Stabile (2010) and Acemoglu et al. (2016) studied the patterns of citations across USPCS or NBER technology classes. Similarly, technological classification systems are used to estimate technological distance, typically between firms or inventors in the "technology space" based on the classification of their patent portfolio (Breschi et al. 2003, Nooteboom et al. 2007, Aharonson & Schilling 2016, Alstott et al. 2016). Additional methodological contributions include Benner & Waldfogel (2008), who have pointed out that using all the codes listed on patents allows to increase the sample size and thus reduce bias in measuring proximity, and McNamee (2013) who argues for using the hierarchical structure of the classification system[1].

In spite of this wide use of the current patent classification system, there has been no quantitative studies of the historical evolution of the system apart from the counts of the number of distinct classes by Bailey (1946) and Stafford (1952), which we update here. Recently though, Strumsky et al. (2012) originated a renewed interest in patent classification by arguing that the classification of patents in multiple fields is indicative of knowledge recombination. Using the complete record of US patents classified according the current classification system, Youn et al. (2015) studied the subclasses ("technology codes"). They found that the number of subclasses used up to a given year is proportional to the cumulative number of patents until about 1870, but grew less and less fast afterwards. Remarkably, however, this slowdown in the introduction of new subclasses does not apply to new *combinations* of subclasses. Youn et al. (2015) found that the number of combinations has been consistently equal to 60% of the number of patents. This finding confirms Strumsky et al.'s (2012) argument that patent classifications contain useful information to understand technological change over the long-

[1] In a related context (how professional diversity scales with city size), Bettencourt et al. (2014) and Youn et al. (2016) exploited the different layers of industry and occupation classifications classification systems to identify resolution-independent quantities. Measuring diversity depends on which layer of the classification system one uses, but in such a way that the infinite resolution limit (deepest classification layer) exists and can be used to characterise universal quantities.

run. Furthermore, the detailed study of combinations can reveal the degree of novelty of specific patents (Strumsky & Lobo 2015, Kim et al. 2016).

Besides their use for simplifying the analysis and creating crosswalks, technology taxonomies are also interesting *per se*. A particularly interesting endeavour would be to construct systematic technology phylogenies showing how a technology descends from others (Basalla 1988, Solé et al. 2013) (for specific examples, see Tëmkin & Eldredge (2007) for cornets and Valverde & Solé (2015) for programming languages).

But categories are not simply useful to describe reality, they are often used to *construct* it (Foucault 1966). When categories are created as nouns, they can have a predicate and become a subject. As a result, classification systems are institutions that allow agents to coordinate and agree on how things should be called and on where boundaries should be drawn. Furthermore, classification systems may create a feedback on the system it describes, for instance by legitimizing the items that it classifies or more simply by biasing which items are found through search and reused in recombination to create other items. Categorization thus affects the future evolution of the items and their relation (boundaries) with other items. Along this line of argument, the process of categorization is performative. In summary, data on the evolution of technological classification systems provides a window on how society understands its technological artefacts and legitimizes them through the process of categorization. According to Latour (2005), social scientists should not over impose their own categories over the actors that they analyze. Instead a researcher should follow the actors and see how they create categories themselves.

Nelson (2006) described technological evolution as the co-evolution of a body of practice and a body of understanding. The role of the body of understanding is to "rationalize" the practice. According to him this distinction has important implication for understanding the evolutionary dynamics, since each body has its own selection criteria. Our argument here is that the evolution of the USPCS reflects how the beliefs of the community of technologists about the mesoscale structure of technological systems coevolves with technological advancements. We consider patent categorization as a process of codification of an understanding concerning the technological system. To see why studying patent categories goes beyond studying patents, it is useful to remember that examiners and applicants do not need to prove that a technology improves our *understanding* of a natural phenomenon; they simply need to show that a device or process is novel and effective at solving a problem. However, to establish a new class, it is necessary to agree that bringing together inventions under this new header actually improves understanding, and thus searchability of the patent system. In that sense we believe that the dynamics of patent classes constitute a window on the "community of technologists"[2] Since classification systems are designed to optimize search, they reflect how search takes place which in turn is indicative of what thought processes are in place. These routines are an integral part of normal problem-solving within a paradigm. As a result, classification systems must be affected by paradigm-switching radical innovations. As noted by e.g. Pavitt (1985) and Hicks (2011), new technology which fits perfectly in the existing classification scheme may be considered an incremental innovation, as opposed to a radical innovation which challenges existing schemes. A direct consequence is that the historical evolution of the classification system contains a great deal of information on technological change beyond the information contained in the patents[3]. We now describe our attempt at reconstructing the dynamics of the U.S. patent classification system.

3 The data: the USPCS

We chose the USPCS for several reasons. First of all, we chose a patent system, because of our interest in technological evolution but also because due to their important legal role patent systems benefit from resources necessary to be maintained up to

[2] Patent officers are generally highly skilled workers. Besides anecdotal evidence on particularly smart patent examiners (Albert Einstein), patent officers are generally highly qualified (often PhDs). That said, Rotkin et al. (1999) mention that classification work was not particularly attractive and that the Classification division had difficulties attracting volunteers. More recently Paradise (2012) eludes to "high turnover, less than ideal wages and heavy workloads". There is an emerging literature on patent officers' biases and incentives (Cockburn et al. 2003, Schuett 2013) but it is focused on the decision to grant the patent. Little is known about biases in classification.

[3] In labor economics, some studies have exploited classification system changes. Xiang (2005) finds that new goods, as measured by changes to the SIC system, have a higher skill intensity than existing goods. Lin (2011) and Berger & Frey (2015) used changes in the index of industries and the dictionary of occupation to evaluate new work at the city level.

date. Among the patent classification systems, the USPCS is the oldest still in use (Wolter 2012). It is also fairly well documented, and in english. The major drawback of this choice is that the USPCS is now discontinued. This means that the latest years may include a classificatory dynamics that anticipate the transition to the Cooperative Patent Classification[4], and also implies that our research will not be updated and cannot make predictions specific to this system that can be tested in the future. Nevertheless, we think that the USPCS had a major influence over technology classifications and makes a good case study.

3.1 The early history of the USPCS

The U.S. patent system was established on 31 July 1790, but the need for examination was abolished 3 years later and reestablished only in 1836. As a result, there was no need to search for prior art and therefore the need for a classification was weak.

The earliest known official subject matter classification appeared in 1823 as an appendix to the Secretary of State's report to the Congress for that year (Rotkin et al. 1999). It classified 635 patents models in 29 categories such as "Bridges and Locks", 1184 in a category named "For various purposes", and omitted those which were not "deemed of sufficient importance to merit preservations".

In 1829, a report from the Superintendent proposed that with the prospect of the new, larger apartments for the Patent office, there would be enough room for a systematic arrangement and classification of models. He appended a list of 14 categories to the report.[5]

In 1830 the House of representatives ordered the publication of a list of all patents, which appeared in December 1830/January 1831 with a table of contents organizing patents in 16 categories, which were almost identical to the 14 categories of 1829 plus "Surgical instruments" and "Horology".[6]

In July 1836, the requirement of novelty examination came into effect, making the search for prior art more pressing. Incidentally, in December the Patent office was completely destroyed by a fire. In 1837, a new classification system of 21 classes was published, including a Miscellaneous class and a few instances of cross noting[7]. The following year another schedule was published, with some significant reorganization and a total number of classes of 22. A new official classification appeared in 1868 and contained 36 main classes. Commenting on this increase in the number of classes, the Commissioner of patents wrote that (Rotkin et al. 1999)

> "The number of classes has risen from 22 to 36, a number of subjects being now recognized individually which were formally merged with others under a more generic title. Among these are builder's hardware, felting, illumination, paper, and sewing machines, to each of which subject so much attention has been directed by inventors that a division became a necessity to secure a proper apportionment of work among the corps of examiners."

Clearly, one of the rationale behind the creation and division of classes is to balance the class sizes, but this was not only to facilitate search. This class schedule was designed with administrative problems in mind, including the assignment of patent applications to the right examiners and the "equitable apportionment of work among examiners" (Rotkin et al. 1999).

Shortly after 1868 a parallel classification appeared, containing 176 classes used in the newly set up patent subscription service. This led to a new official classification containing 145 classes and published as a book in 1872. The number of classes grew to 158 in 1878 and 164 in 1880. Rotkin et al. (1999) note that the 1880 classification did not contain any form of cross-noting and cross references. In 1882 classification reached 167 classes and introduced indentation of subclasses at more than one level. The classification of 1882 also introduced the class 36, "Electricity".

In 1893 it was made clear in the annual report that a Classification division was required "so that [the history of invention] would be readily accessible to searchers upon the novelty of any alleged

[4] `http://www.cooperativepatentclassification.org/index.html`

[5] The main titles were Agriculture, Factory machine, Navigation, Land works, Common trades, Wheel carriages, Hydraulicks (the spelling of which was changed in 1830), Calorific and steam apparatus, Mills, Lever and screw power, Arms, Mathematical instruments, Chemical compositions and Fine arts.

[6] An interesting remark on this classification (Rotkin et al. 1999) is that it already contains classes based on industry categories (agriculture, navigation, ...) and classes based on a "specific mechanical force system" (such as Lever and screw power).

[7] The first example given by Rotkin et al. (1999) is a patent for a pump classified in both "Navigation" and in "Hydraulics and Hydrostatics"

invention". After that, the need for a classification division (and the associated claim for extra budget) was consistently legitimated by this need to "oppose the whole of prior art" to every new application. In 1898 the "Classification division" was created with a head, two assistants and two clerks, with the purpose of establishing clearer classification principles and reclassifiying all existing patents. This marked the beginning of professional classification at the USPTO.

Since then the classification division has been very active and the patent classification system has evolved considerably, as we document extensively in this paper. But before, we need to explain the basic organizing principles of the classification system.

3.2 Rationale and organization of the modern USPCS

The USPCS attributes to each patent at least one subject matter. A subject matter includes a main class, delineating the main technology, and a subclass, delineating processes, structural features and functional features. All classes and most subclasses have a definition. Importantly, these are the patent claims which are classified, not the whole patent itself. The patent inherits the classification of its claims; its main classification is the classification of its main ("most comprehensive") claim.

There are different types of patents, and they are translated into different types of classes. According to the USPTO[8], "in general terms, a utility patent protects the way an article is used and works, while a design patent protects the way an article looks." The "classification of design patents is based on the concept of function or intended use of the industrial design disclosed and claimed in the Design patent."[9].

During the XIXth century classification was based on which industry or profession was using the invention, for instance "Bee culture" (449) or "Butchering" (452). The example of choice (Falasco 2002, USPTO 2005, Strumsky et al. 2012) is that of cooling devices which were classified separately if they were used to cool different things, such as beer or milk. Today's system would classify both as cooling devices into the class "Heat exchange" (165), which is the utility or function of the invention. Another revealing example (Schmookler 1966, Griliches 1990) is that a subclass dealing with the dispensing of liquids contains both a patent for a water pistol and one for a holy water dispenser. This change in the fundamental principles of classification took place at the turn of the century with the establishment of the Classification division (Falasco 2002, Rotkin et al. 1999). Progressively, the division undertook to redesign the classification system so that inventions would be classified according their utility. The fundamental principle which emerged is that of "utility classification by *proximate* function" (Falasco 2002) where the emphasis on "proximate" means that it is the fundamental function of the invention, not some example application in a particular device or industry. For instance "Agitating" (366) is the relevant class for inventions which perform agitation, whether this is to wash clothes, churn butter, or mix paint (Simmons 2014). Another classification by utility is the classification by effect or product, where the result may be tangible (e.g. Semiconductors device and manufacture, 438) or intangible (e.g. Audio signal system, 381). Finally, the classification by structure ("arrangement of components") is sometimes used for simple subject matter having general function. This rationale is the most often used for chemical compounds and stock material. It is rarely used for classes and more often used at the subclass level (USPTO 2005)

Even though the classification by utility is the dominant principle, the three classification rationales (by industry, utility and structure) coexist. Each class "reflects the theories of classification that existed at the time it was reclassified" (USPTO 2005). In addition, the system keeps evolving as classes (and even more so subclasses) are created, merged and split. New categories emerge when the need is felt by an examiner and approved by the appropriate Technology Center; in this case the USPCS is revised through a "Classification order" and all patents that need to are reclassified (Strumsky et al. 2012).

One of the latest class to have been created is "Nanotechnology (977)", in October 2004. As noted by Strumsky et al. (2012), using the current classification system one finds that after reclassification the first nanotechnology patent was granted much earlier[10]. According to Paradise

[8] `http://www.uspto.gov/web/offices/pac/mpep/s1502.html`

[9] `http://www.uspto.gov/page/seven-classification-design-patents`

[10] 1986 for Strumsky et al. (2012), 1978 for Paradise (2012) and 1975 (US3896814) according to the data that we use here. Again, these differences reflect the importance

(2012), large federal research funding led to the emergence of "nanotechnology" as a unifying term, which became reflected in scientific publications and patents. Because nanotechnologies were new, received lots of applications and require interdisciplinary knowledge, it was difficult to ensure that prior art was reviewed properly. The USPTO engaged in a classification project in 2001, which started by defining nanotechnologies and establishing their scope, through an internal process as well as by engaging with other stakeholders such as users or other patent offices. In 2004 the Nanotechnology cross-reference digest was established; cross-reference means that this class cannot be used as a primary class. Paradise (2012) argues that class 977 has been defined with a too low threshold of 1 to 100 nanometers. Also, reclassification has been encouraged but is not systematic, so that many important nanopatents granted before 2004 may not be classified as such.

Another example of class creation worth mentioning is given by Érdi et al. (2013) who argue that the creation of "Fabric (woven, knitted, or nonwoven textile or cloth, etc.)" (422) created in 1997, could have been predicted based on clustering analysis of citations.

Finally, a last example is that of organic chemistry[11]. Class 260 used to contain the largest array of patent documents but it was decided that this class needed to be reclassified "because its concepts did not necessarily address new technology and several of its subclasses were too difficult to search because of their size.". To make smaller reclassification projects immediately available it was decided to split the large class into many individual classes in the range of Classes 518-585. Each of these classes is "considered an independent class under the Class 260 umbrella"; many of these classes have the same general name such as "Organic coumpounds – part of the class 532-570 series"[12]

As argued by Strumsky et al. (2012), this procedure of introducing new codes and modifying existing ones ensures that the current classification of patents is consistent and makes it possible to study the development of technologies over a long period of time. However, while looking at the past with today's glasses ensures that we look at different periods of the past in a consistent way, it is not the same as reporting what the past was in the eyes of those who lived it. In this sense, we believe that it is also interesting to try and reconstruct the classification systems that were in place in the past. We now describe our preliminary attempt to do so, by listing available sources and constructing a simple count of the number of classes used in the past.

3.3 Dataset construction

In this paper we focus on main classes, due to the difficulty of collecting historical data at the subclass level. We relied on several sources. First of all, we use the Master Classification File for patent grants (version mcfpat1512) for patent classifications and the Patent Grant Authority File (version 20160130) for patent grant years[13].

Our most important data collection effort concerns the historical number of classes. For the early years our main sources are Bailey (1946) and Rotkin et al. (1999), complemented by Reingold (1960) and the "Manual of Classification" for the 5 years within the period 1908–1923.

For the 1950–60's, we used mostly a year-specific source named "General information concerning Patents" which contained a sentence as "Patents are classified into x classes". Unfortunately starting 1969 the sentence becomes "Patents are classified into more than 310 classes".

We therefore switched to another source named "Index of patents issued from the United States Patent Office", which contains the list of classes. Starting 1963, it contains the list of classes with their name and number on a separate page[14]. For 1985, we used a report of the Office of Technology

of reclassification.

[11] see `http://www.uspto.gov/page/addendum-reclassification-classes-518-585`

[12] These classes also have a hierarchy indicated by their number, as subclasses within a class schedule usually do.

[13] We removed 93 patents dated January 1st 1800, 4 patents dated 1700, and whenever we use the main classification only we also remove 298 patents with no main (OR) classification

[14] We had to make assumptions about what counts as Design class or subclass. In the 60's, the list shows that Designs are subdivided into "Industrial arts" and "Household, personal and fine arts". We thus assume that the number of design classes is 2, up to the year 1977 where Design classes appear on this list with their name and number. We implicitly assume that prior to 1977 the design classes were actually subclasses, since in 1977, there were 39 Design classes, whereas the number of (sub)classes used for design patents in 1976 was more than 60. It should be noted though that according to the dates established, some of the current design classes were created in the late 60's. Moreover for 1976 the number of Organic compound classes in not clear; We assumed it was 6, as listed in 1977. Note also that sometimes we have two slightly different values for the same year due to contradictory sources or because the sources refer to a different month.

Assessment and Forecast (OTAF) of the Patent and Trademark Office (OTAF 1985). For the years 2001 to 2013, we collected data from the Internet Archive.[15]

As of February 2016 there are 440 utility classes (including the Miscellaneous 001 and and the "Information storage" G9B (established in 2008)), 33 design classes, and the class PLT "Plant", giving a total of 474 classes.[16].

4 Dynamics of the number of classes and Heaps' law

Our first result concerns the growth of the number of classes (Fig. 1). We have computed the growth of the number of classes according to several methods.

First, we used the raw data collected from the historical sources mentioned in Section 3.3. Quite unexpectedly, the data suggests a linear growth, with appreciable fluctuations mainly due to the introduction of an entirely new system in 1872 and to design classes in 1977. The grey line shows the linear fit with an estimated slope of 2.41 (s.e. 0.06) and R^2 of 0.96. Second, we have computed, using the Master Classification File for 2015, the number of distinct classes in which the patents granted up to year t are classified (black line). To do so, we have used all classes in which patents are classified (i.e. including cross-reference classes).[17] The pattern of growth is quite different from the historical data. If we consider only the post-1836 data, the growth of the number of classes is sublinear – less and less classes are introduced every year. Before 1836, the trend was linear or perhaps exponential, giving a somewhat asymmetric S-shape to the overall picture. Third, we computed the growth of the number of classes based on the dates at which all current classes were established (blue line)[18]. According to this measure, the first class was created in 1899, when the reorganization of classification started with the creation of the classification division[19].

Fig. 2 displays the number of classes against the number of patents in a log-log scale. In many systems, it has been found that the number of categories grows as a power law of the number of items that they classify, a result known as Heaps' law (for an example based on a classification system –the medical subject headings– instead of a language, see Petersen et al. (2016)). Here we find that using the 2015 classification, Heaps' law is clearly violated[20]. Using the historical data, Heaps' law appears as a reasonable approximation. We estimate the Heaps' exponent to be 0.376 with standard error of 0.009 and $R^2 = 0.97$. The inset on the bottom right of Fig. 2 shows that for the latest years, Heaps' law fails: for the latest 2 million patents (about 20% of the total), almost no classes were created. We do not know whether this slowdown in the introduction of classes is due to a slowdown of radical innovation, or to a more institutionally-driven reason such as a lack of investment in the USPCS due to the expected switch to the Common Patent Classification. Since the joint classification system was first announced on 25 October 2010 (Blackman 2011), we show this date (more precisely, patent number 7818817 issue on the 26^{th}) as a suggestive indicator (dashed line on the inset). Another consideration is that the system may be growing more "vertically", in terms of the number of layers of subclasses – unfortunately here we have to focus on classes, so we here we are not able to test for this.

Generally, we would observe Heaps' law whenever the number of items and the number of categories grow exponentially in time. However, this is not the case here, since Fig. 1 suggests a linear growth for the number of classes and a subexponential growth for the number of patents (inset of Fig. 2). More importantly for our point, models that would explain Heaps' law based on a growing

[15] `https://archive.org/index.php` where we can find the evolution of the url `http://www.uspto.gov/web/patents/classification/selectnumwithtitle.htm`. We added the class "001" to the count.

[16] The list of classes available with their dates established contains 476 classes, but it does not contain 001, and it contains 364, 389, and 395 which have been abolished.

[17] The (reconstructed) number of classes is slightly lower if we consider only Primary classes, because some classes are used only as a cross-reference, never as primary class. These classes are 902: Electronic funds transfer, 903: Hybrid electric vehicles, 901: Robots, 930: Peptide or protein sequence, 977: Nanotechnology, 976: Nuclear technology, 968: Horology, 987: Organic compounds containing a bi, sb, as, or p atom or containing a metal atom of the 6th to 8th group of the periodic system, 984: Musical instruments, G9B: Information storage based on relative movement between record carrier and transducer.

[18] Collected from `https://www.uspto.gov`, page USPCS dates-established

[19] "Buckles, Buttons, clasps, etc." is an example of a class that was created early under a slightly different name (1872 according to Simmons (2014), see Bailey (1946) for details) but has a posterior "date established" (1904 according to the USPTO). Another example is "Butchering".

[20] It is possible to obtain a good fit by limiting the fit to the latest periods, however this is arbitrary, and gives a very low Heaps' exponent, leaving unexplained the creation of the vast majority of classes.

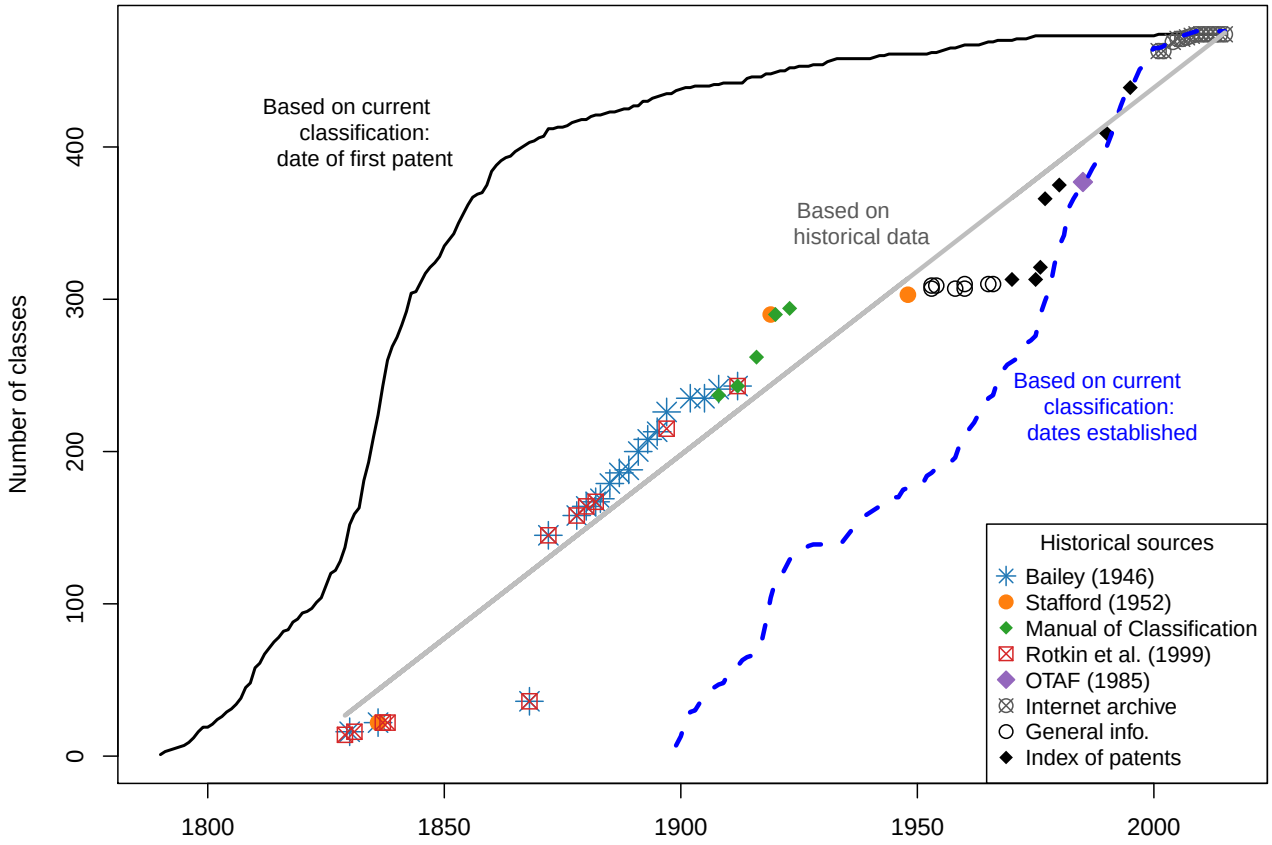

Figure 1: Evolution of the number of distinct classes.

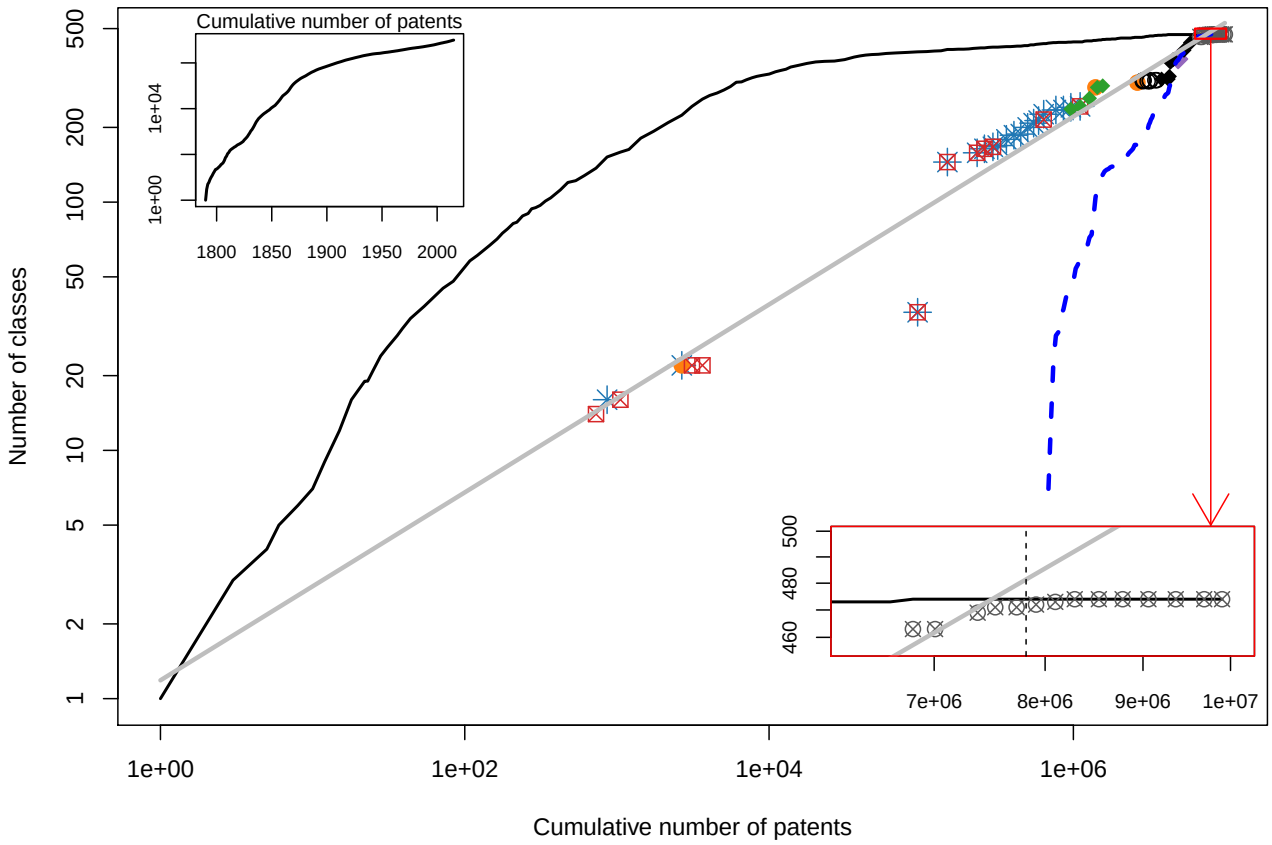

Figure 2: Heaps' law.

number of well defined, immutable categories and their own accumulation of patents would miss the point. We now expand on this by studying a striking empirical fact regarding the size distribution.

5 The size distribution and the age-size relationship

Besides the creation and reorganization of technological categories, we are interested in their growth and relative sizes. More generally, our work is motivated by the Schumpeterian idea that the economy is constantly reshaping itself by introducing novelty (Dopfer et al. 2004, Saviotti & Pyka 2004). The growth of technological domains has been deeply scrutinized in the economics of technical change and development (Schumpeter 1934, Dosi 1982, Pasinetti 1983, Pavitt 1984, Freeman & Soete 1997, Saviotti 1996, Malerba 2002). A recurring theme in this literature is the high heterogeneity among sectors. When sectors or technological domains grow at different rates, structural change occurs: the relative sizes of different domains is modified. To study this question in a parsimonious way, one may opt for a mesoscale approach, that is, study the size distribution of categories.

Our work here is most directly related to Carnabuci (2013) who first showed on data for 1963–1999 that the size distribution of classes is close to exponential. This is an interesting and at first surprising finding, because based on the assumption that all domains grow at the same average rate stochastic growth models such as Gibrat (1931) or Yule (1925) predict a Log-normal or a Pareto distribution, which are much more fat tailed. Instead, we do not see the emergence of relatively very large domains, and this may at first suggest that older sectors do not keep growing as fast as younger ones, perhaps due to technology life-cycles (Vernon 1966, Klepper 1997, Andersen 1999). However, as we will discuss, we are able to explain the exponential size distribution by keeping Gibrat's law, but assuming that categories are split randomly.

5.1 The size distribution of categories

In this section we study the size distribution of classes, where size is the number of patents in 2015 and classes are defined using the current classification system. We use only the primary classification, so we have only 464 classes. Fig. 3 suggests a linear relationship between the size of a class and the log of its rank, that is, class sizes are exponentially distributed[21]. To see this, let $p(k)$ be the probability density of the sizes k. If it is exponential, it is $p(k) = \lambda e^{-\lambda k}$. By definition, the rank $r(k)$ of a class of size k is the number of classes that have a larger size, which is $r(k) = N \int_k^\infty \lambda e^{-\lambda x} dx = N e^{-\lambda k}$, where N is the number of classes. This is equivalent to size being linear in the logarithm of the rank. We estimate the parameter λ by maximum likelihood and obtained $\hat{\lambda} = 4.69 \times 10^{-5}$ with standard error 0.22×10^{-5}. Note that $\hat{\lambda}$ is one over the mean size, 21332. We use this estimate to plot the resulting fit in Fig. 3.

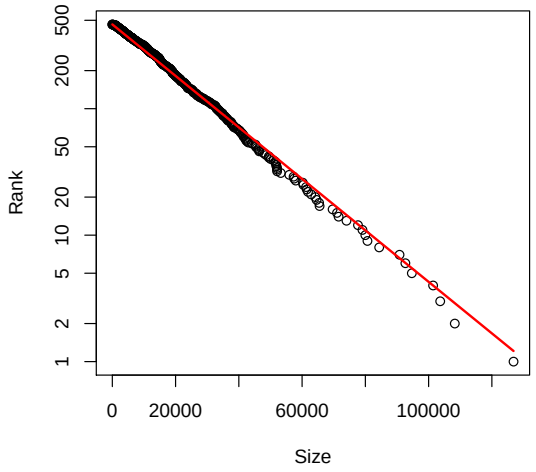

Figure 3: Rank-size relationship.

It is interesting to find an exponential distribution, since one may have expected a power law, which is quite common as a size distribution, and appears often with Heaps' law (Lü et al. 2010, Petersen et al. 2016). Since the exponential distribution is a good representation of the data, it is worth looking for a simple mechanism that generates this distribution, which we will do in Sec-

[21] For simplicity we used the (continuous) exponential distribution instead of the more appropriate (discrete) geometric distribution, but this makes no difference to our point. We have not rigorously tested whether or not the exponential hypothesis can be rejected, because the proper hypothesis is geometric and classical test statistics such as Kolmogorov-Smirnov do not easily apply to discrete distributions. Likelihood ratio tests interpreted at the 5% level showed that it is possible to obtain better fits using two-parameters distributions that extends the exponential/geometric, namely the Weibull and the Negative binomial, especially after removing the two smallest categories which are outliers (contain 4 and 6 patents) and are part of larger series (520 and 532).

tion 7. But since many models can generate an exponential distribution we first need to present additional empirical evidence that will allow us to discriminate between different candidate models.

5.2 The age-size relationship

To determine whether older classes contain more patents than younger ones, we first need to note that there are two ways of measuring age: the official date at which the class was established, and the year in which its first patent was granted. As expected, it appears that the year in which a class is established is always posterior to the date of its first patent[22].

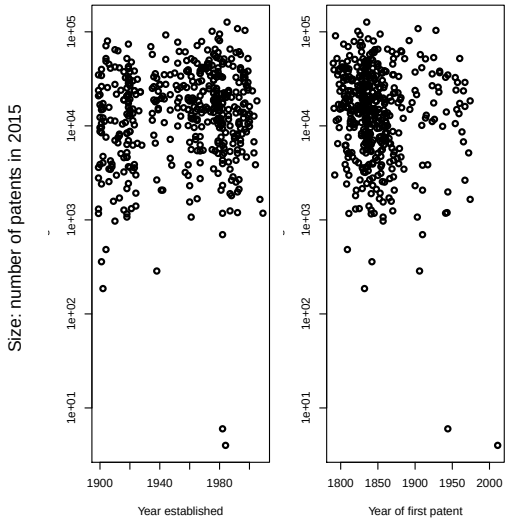

Figure 4: Age-size relationship.

Since these two ways of measuring age can be quite different, we show the age-size (or rather size-birth date) relationship for both in Fig. 4. If stochastic growth models without reclassification were valid, we would observe a negative slope, that is, newer classes should have fewer patents because they have had less time for accumulation from random growth. Instead, we find no clear relationship. In the case of the year established, linear regressions indicated a positive relationship significant at the 10% but not at the 5% confidence level, whether or not the two "outliers" were removed. Using a log-linear model, we found a highly significant coefficient of 0.004 after removing the two outliers. In the case of the year of the first patent, we found a highly significant negative coefficient of -0.005, although it becomes significant only at the 10% level after removing the two outliers); In all cases (two different age variables and two different models) the R^2 was below 1%. We conclude that these relationships are at best very weak, and in one case of the "wrong" sign (with classes established in recent years being on average larger). Whether they are significant or not, their magnitude and the goodness of fits are much lower than what one would expect from growth-only models such as Simon (1955), or its modification with uniform attachment (to match the exponential size distribution). We will come back to the discussion of models later, but first we want to show another empirical pattern and explain why we think reclassification and classification system changes are interesting indicators of technological change.

6 Reclassification activity as an indicator of technological change

It seems almost tautological to say that a radical innovation is hard to categorize when it appears. If an innovation is truly "radical", it should profoundly change how we think about a technology, a technological domain, or a set of functions performed by technologies. If this is the case a patent related to a radical innovation is originally hard to classify. It is likely that it will have to be reclassified in the future, when a more appropriate set of concepts has been developed and institutionalized (that is, when the community of technologists have codified a novel understanding about the radical innovation). It is also well accepted that radical innovations may create a new wave of additional innovations, which may or may not cluster in time (Silverberg & Verspagen 2003) but when they are general purpose we do expect a rise in innovative activity (Bresnahan & Trajtenberg 1995). A less commented consequence of the emergence and diffusion of General Purpose Technologies (GPTs) is that both due to the sheer increase in the number of patents in this technology, and to the impact of this technology on others, we should expect higher classification volatility. Classification volatility is to be expected particularly in relation to GPTs because by definition a GPT interacts with existing technologies and create or reorganize interactions among existing technologies. From the point of view of the classification, the very definition of the objects and their boundaries are transformed. In

[22] Apart from class 532. We confirmed this by manually searching the USPTO website. 532 is part of the Organic compound classes, which have been reorganized heavily, as discussed in section 3.2

short, some categories become too large and need to be split; some definitions become obsolete and need to be changed; and the "best" grouping of technologies is affected by the birth and death of conceptual relationships between the function, industry of origin or application, and structural features of technologies.

In this section we provide a preliminary study. First we establish that this indicator does exist (reclassification rates can be quite high if we can look far enough in the past). Second, we show that reclassified patents are more cited. And third, we show that reclassification can take place across fairly distant technological domains, as measured by 1-digit NBER categories.

6.1 Reclassification rates

How many patents have been reclassified? To start with, since no classification existed prior to 1829, all patents published before that have been "(re)classified" in the sense that their category has been determined several and potentially many years after being granted. The same applies to all patents granted at times where completely different classification systems prevailed, which is the case before 1899. In modern times, classification has evolved but as discussed in Section 3, the overall classification framework put in place at the turn of the century stayed more or less the same. For the period after 1976, we know the original classification of each patent because we can read it on the digitized version of the original paper. This allows us to calculate, for each year t, a rate of reclassification defined as the share of patents granted in year t which have a different classification in 2015 than in t.

Figure 5 shows the evolution of the reclassification rate. It appears that as much as 40% of the 1976's patents belong to a different class now than when they first appear. This reclassification rate declines sharply after that, reaching about 10% in the 1990's and almost zero thereafter. This is an expected result, since the longer the time since granting the patent, the higher the chances that the classification system has changed.

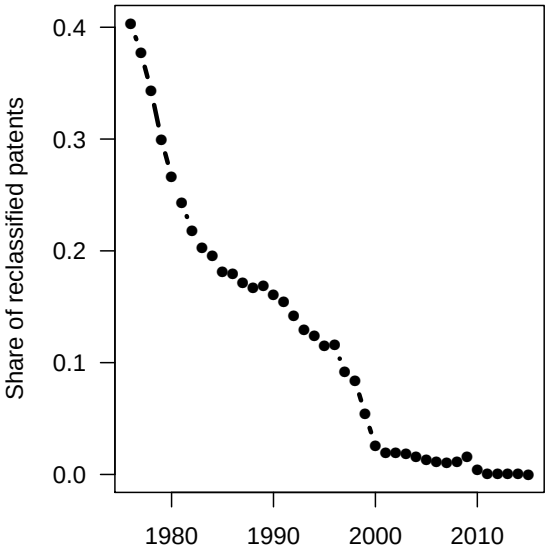

Figure 5: Share of patents granted in a given year that are in a different class in 2015, as compared to when they were granted.

6.2 Are reclassified patents more cited?

Since there is an established relationship between patent value and the number of citations received (Hall et al. 2005), it is interesting to check if reclassified patents are more cited. Of course, we are only observing correlations, and the relationship between citations and reclassification can work in multiple ways. A plausible hypothesis is that the more active is a technological domain (in terms of new patents and thus new citations being made), the more likely it is that there will be a need for reclassification, if only to keep the classes at a manageable size. Another hypothesis is that highly innovative patents are intrinsically ambiguously defined in terms of the classification system existing when they first appear. In any case, since we only have the class number at birth and the class number in 2015, we cannot make subtle distinction between different mechanisms. However, we can check whether reclassified patents are on average more cited, and we can do so after controlling for the grant year and class at birth. To check this, we use the Patent Grant Bibliographic (Front Page) Text Data (January 1976 – December 2015) provided by the USPTO[23].

Table 1 shows basic statistics. Reclassified patents constitute 7% of the sample, and have received on average more than 24 citations, which is more than twice as much as the non reclassified patents. Figure 6 shows the distribution of citations[24] for reclassified and non reclassified patents separately. It appears that at every number of

[23] `https://bulkdata.uspto.gov/` (Access date: March 2, 2016)

[24] The almost linear slope on a log-log plot suggests that this distribution may be Pareto, at least in the tail. Fortunately, the Pareto exponent is not extremely low, so the expectation is well defined (see Csárdi et al. (2007), Valverde et al. (2007) and Silverberg & Verspagen (2007)).

citations received k less than the maximum number of citations received by a reclassified patent, the share of patents with a number of citations received higher than k is higher for reclassified than for non reclassified patents.

	share	mean	median	s.d.
All	1	11.37	4	27.49
Non reclass	0.93	10.35	3	24.77
Reclass.	0.07	24.77	11	49.17

Table 1: Patent citations summary statistics.

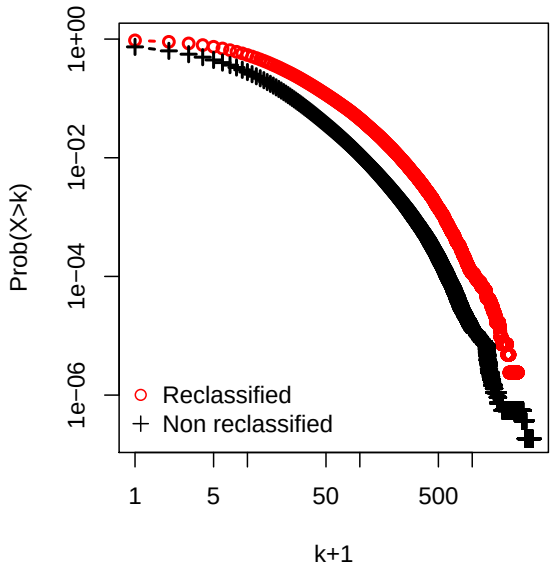

Figure 6: Complementary cumulative distribution function for citations received by reclassified and non-reclassified patents.

To investigate the relationship between reclassification and citations in more detail, we regressed the log of total citations received in 2015 on the reclassification dummy and on dummies for the class at birth, for each year separately (and keeping only the patents with at least one citation received, about 75%):

$$\log(c_i) = \alpha_t + \beta_t R_i + \sum_{j=1}^{J_t - 1} \gamma_{j,t} D_{i,j}$$

where c_i is the number of citations received by patent i between its birth (time t) and 2015, R_i is a dummy that takes the value of 1 if patent i has a main class code in 2015 different from the one it had when it appeared (i.e. in year t), J_t is the number of distinct classes in which the patents born in year t were classified at birth, and $D_{i,j}$ is a dummy that takes the value of 1 if patent i was classified in class j at birth.

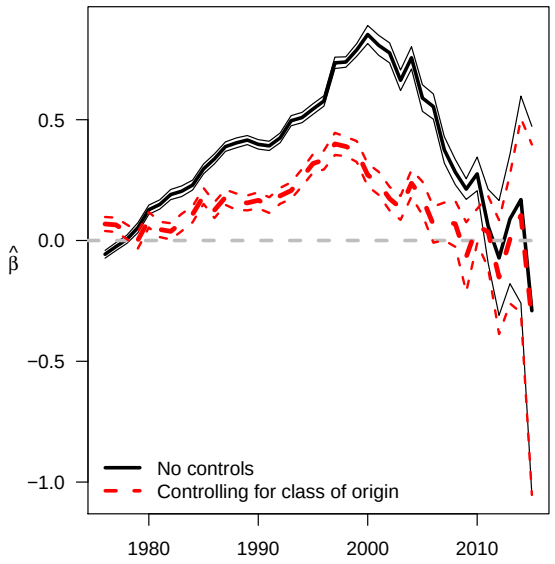

Figure 7: Coefficient of the year-specific regressions of the log of citations received on the reclassification dummy (including dummies for the class of origin or not).

Note that we estimate this equation separately for every grant year. We include the class at birth dummies because this allows us to consider patents that are "identical twins" in the sense of being born in the same class in the same year. The coefficient β then shows if reclassified patents have on average received more citations. The results are reported in Fig. 7, showing good evidence that reclassification is associated to more citations received. As expected, recent years are not significant since there has not been enough time for reclassification to take place and citations to accumulate (the bands represent standard approximate 95% confidence intervals). We also note that controlling for the class at birth generally weakens the effect (red dashed line compared to black solid line).

6.3 Reclassification flows

To visualize the reclassification flows, we consider only the patents that have been reclassified. As in Wang et al. (2016) we want to construct a bipartite graph showing the original class on one side and the current class on the other side. Since we identify classes by their code number, a potentially serious problem may arise if classes are renumbered, although we believe this tends to be rare given the limited time span 1976–2015. An example of this is "Bee culture" which was class number 6, but since 1988 is class number 449 and class number 6 does no longer exists. However, even in this case, even though these two classes have the same

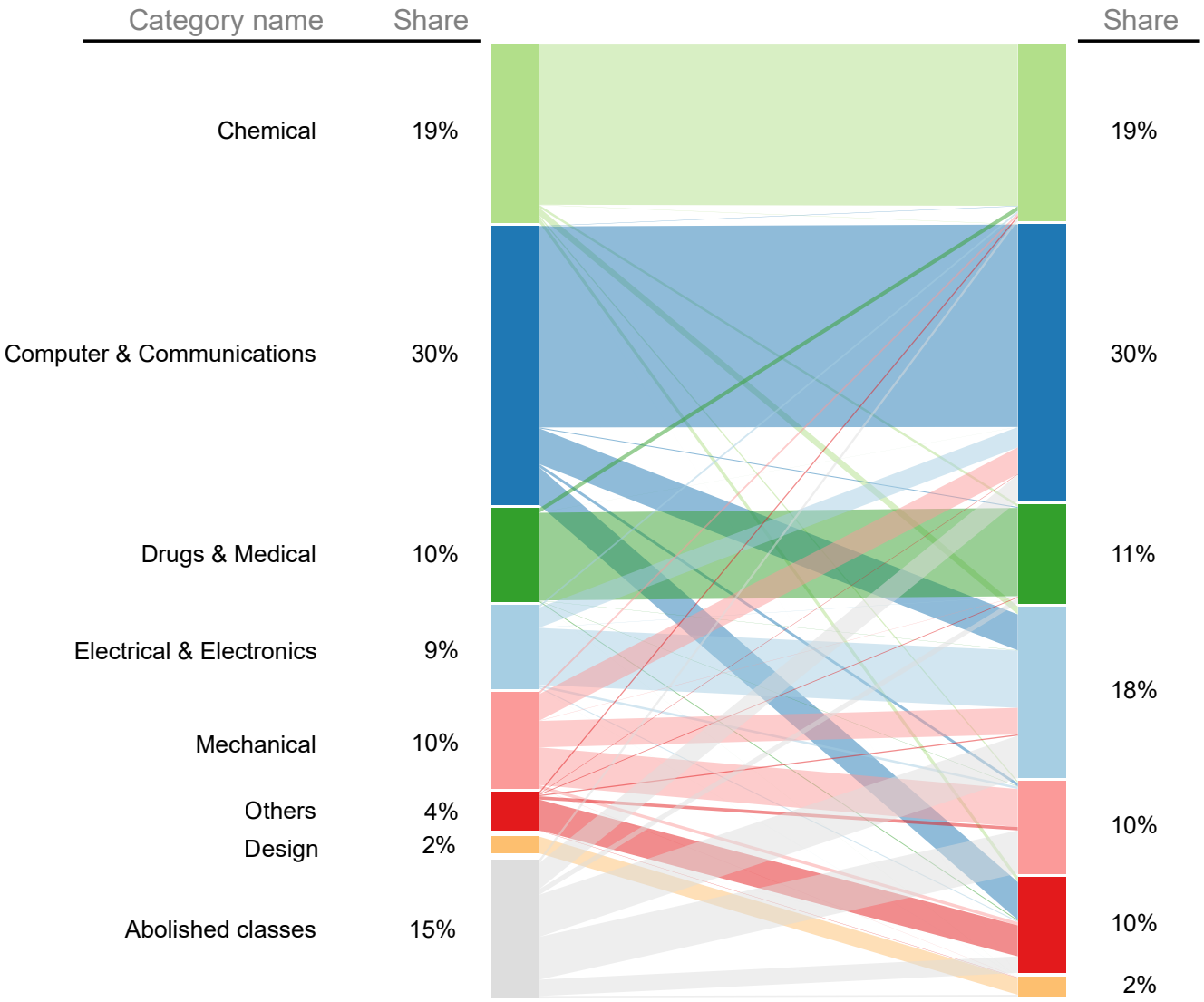

Figure 8: Reclassification flows.

name, we do not know if they are meant to encompass the same technological domain and have just been "renumbered", or if other considerations prevailed and renumbering coincides with a more substantive reorganisation. An interesting extension of our work would be to use natural language processing techniques on class definitions to define a measure reclassification distance more precisely and exclude mere renumbering.

To make the flow diagram readable and easier to interpret, we aggregate by using the NBER categories[25]. To assign each class to a NBER category, we used the 2006 version of the NBER classification, which we modified slightly by classifying the Design classes separately, and classifying USPCS 850 (Scanning probe techniques and apparatus) in NBER 4 (Electrical) and USPCS PLT (Plant) in NBER 6 (Others).

Fig. 8 shows the results[26]. The share of a category means the fraction of reclassified patents whose primary class is in a particular NBER category. The width of the lines between an original category i and a current category j is proportional to the number of reclassified patents whose original class is in category i and current class is in category j. Line colors indicate the original category.

We can see that patents originally classified in the categories Chemical tend to be reclassified in another class of the category Chemical. The same pattern is observed for the category Drugs. By contrast, the categories Computers & Communications and Electrical & Electronics display more cross-reclassifications. This may indicate that the

[25] For more details on the NBER categories, see the historical reference (Hall et al. 2001) and the recent effort by Marco et al. (2015) to attribute NBER (sub) categories to patent applications.

[26] See the online version at `http://danielykim.me/visualizations/PatentReclassificationHJTcategory/`

NBER categories related to computers and electronics are not as crisply defined as those related to Chemical and Drugs, and may be suggestive of the general purpose nature of computers. This could also suggest that that these domains were going through a lot of upheaval during this time period. While there is some ambiguity in interpreting these patterns, they are not *a priori* obvious and point to the same phenomenon as the correlation between citations and reclassifications: dynamic, impact-full, really novel, general purpose fields are associated to more taxonomic volatility.

7 A simple model

In this section, we propose a very simple model that reproduces several facts described above. As compared to other recent models for size distributions and Heaps' law in innovation systems (Tria et al. 2014, Marengo & Zeppini 2016, Lafond 2014), the key assumption that we will introduce is that classes are sometimes split and their items reclassified. We provide basic intuition instead of a rigorous discussion[27].

Let us start with the well-known model of Simon (1955). A new patent arrives every period. The patent creates a new category with probability α, otherwise it goes to an existing category which is chosen with probability proportional to its size. The former assumption is meaningful, because in reality the number of categories grows over time. The second assumption is meaningful too, because this "preferential attachment"/"cumulative advantage" is related to Gibrat's law: categories grow at a rate independent of their size, so that their probability of getting the next patent is proportional to their size.

There are three major problems with this model. First it gives the Yule-Simon distribution for the size distribution of classes. This is basically a power law so it has much fatter tails than the exponential law that we observe. In other words, it over predicts the number of very large categories by a large margin. Second, since older categories have more time to accumulate patents, it predicts a strong correlation between age and size. Third, since at each time step categories are created with probability α and patents are created with probability 1, the relationship between the number of categories αt and the number of patents t is linear instead of Heaps' constant elasticity relation.

A solution to make the size distribution exponential instead of power law is to change preferential attachment for uniform random attachment, that is to choose each category with equal probability. Besides the fact that this new assumption may seem less intuitive than Gibrat's law, this would not solve the second problem because it would still be the case that older categories accumulate more patents. The solution is to acknowledge that categories are not entities that are defined once and for all; instead, they are frequently split and their patents are reclassified.

We therefore turn to the model proposed by Ijiri & Simon (1975). It assumes that new categories are introduced over time by splitting existing ones. In its original form the model postulates a linear arrangement of stars and bars. Each star represents a patent, and bars materialize the classes. For instance, if there are 3 patents in class 1 and 1 patent in class 2, we have |***|*|. Now imagine that between any two symbols there is a space. At each period, we choose a space uniformly at random and fill it with either a bar (with probability α) or a star (with complementary probability). When a star is added, it means that an existing category acquires a new patent. When a bar is added, it means that an existing category is split into two categories. It turns out that the resulting size-distribution is exponential, as desired. But before we can evaluate the age-size relationship, we need to decide how to measure the age of a category. To do this we propose to reformulate the model as follows.

We start with one patent in one category. At each period, we first select an existing category j with probability proportional to its size k_j and add one patent in it. Next, with probability α we create two novel categories by splitting the selected category uniformly at random; that is, we draw a number s from a uniform distribution ranging from 1 to k_j. Next, each patent in j is assigned to the new category 1 with the probability being s/k_j, or to the new category 2 otherwise. This procedure leads to a straightforward interpretation: the patents are *reclassified* from j to the first or the second new category. These two categories are *established* at this step of the process, and since patents are created sequentially one by one, we also know the *date of the first patent*

[27] For instance, we do not claim that the model *in general* produces a certain type of pattern such as a lack of age-size relationship. We simply show that under a specific parametrisation taken from the empirical data (10 million patents, 474 classes, and a Heaps exponent of about 0.376), it produces patterns similar to the empirical data.

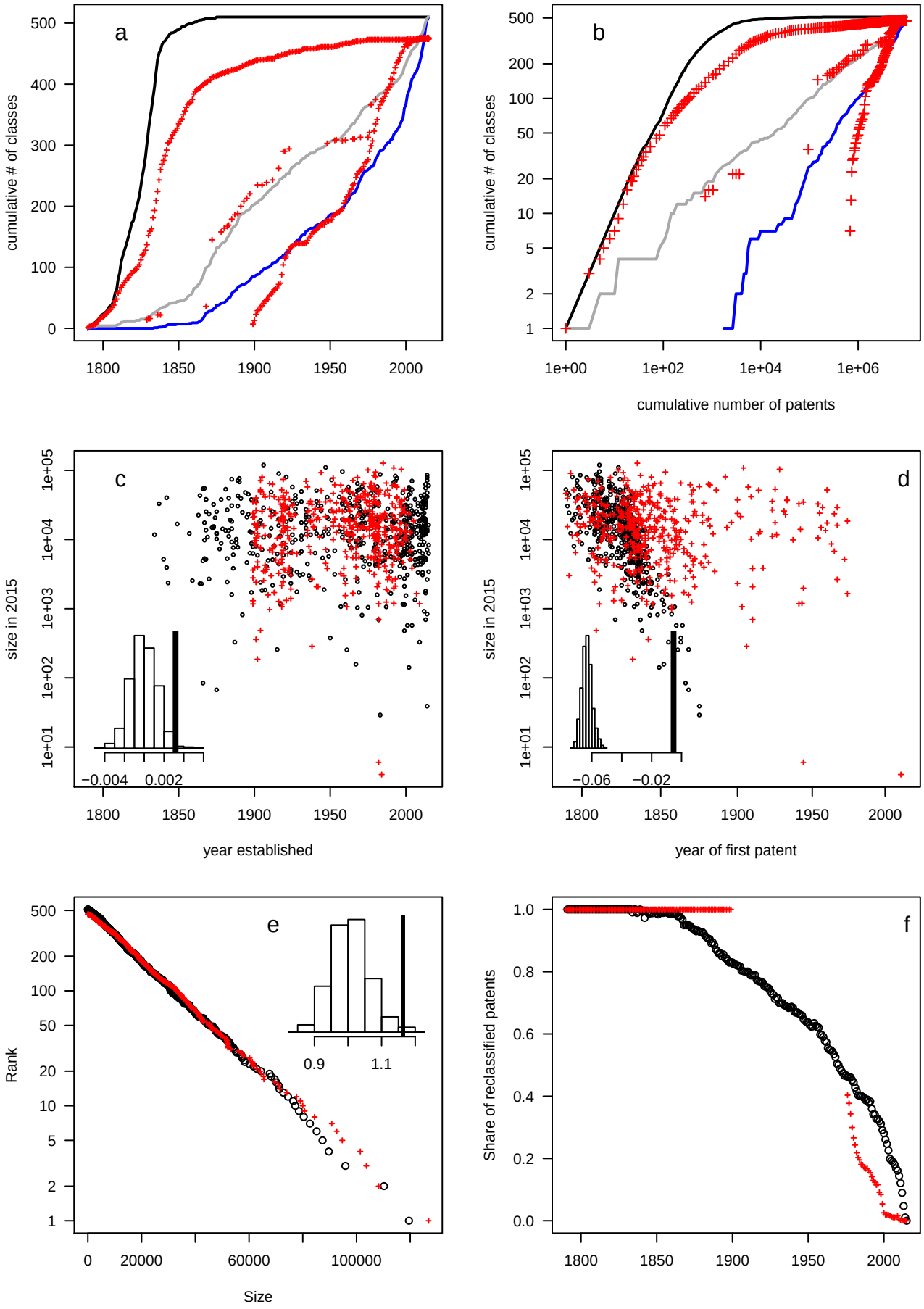

Figure 9: Simulation results against empirical data (red crosses). See Section 7 for details

of each new category. To give a date in calendar years to patents and categories, we can simply use the dates of the real patents.

Since α is constant, as in Simon's original model, we are left with the third problem (Heaps' power law is violated). We propose to make α time dependent to solve this issue[28]. Denoting the number of categories by C_t and the number of patents by t (since there is exactly one new patent per period), we want to have $C_t = C_0 t^b$ (Heap's law). This means that C_t should grow at a per period rate of $dC_t/dt = C_0 b t^{b-1}$. Since we have measured $b \approx 0.376$ and we want the number of categories to be 474 when the number of patents is 9,898,324, we can calculate $C_0 = C_t/t^b = 1.11$. This gives $\alpha_t = 1.11 \times 0.376\, t^{0.376-1}$, which we take to be 1 when $t = 1$.[29]

Note how parsimonious the model is: its only inputs are the current number of patents and categories, and the Heaps' exponent. Here we do not attempt to study it rigorously. We provide simulation results under specific parameter values. Fig. 9 shows the outcome of a single simulation run (black dots and lines), compared to empirical data (red crosses).

The first pair of panels (a and b) shows the same (empirical) data as Fig. 1 and 2 using red crosses. The results from the simulations are the curves. The simulation reproduces Heaps' law well, by direct construction (the grey middle curve on panel b). But it also reproduces fairly well the evolution of the reconstructed number of classes, both the one based on the "date of first patent" and the one based on the "dates established", and both against calendar time (years) and against the cumulative number of patents.

The second pair of panels (c and d) show the age-size relationships, with the same empirical data as in Fig. 4. Panel c shows that the model seems to produce categories whose sizes are *not* strongly correlated with the year in which they were established, as in the empirical data. However, in our model there is a fairly strong negative correlation between size and the year of the first patent and this correlation is absent (or is much weaker) in the empirical data. These results for one single run are confirmed by Monte Carlo simulations. We ran the model 1000 times and recorded the estimated coefficient of a simple linear regression between the log of size and each measure of age. The insets show the distribution of the estimated coefficients, with a vertical line showing the coefficient estimated on the empirical data.

The next panel (e) shows the size distribution in a rank-size form, as in Fig. 3. As expected, the model reproduces this feature of the empirical data fairly well. However the empirical data is not exactly exponential and may be slightly better fitted by a negative binomial model (which has one more parameter and recovers the exponential when its shape parameter equals one). The top right histogram shows the distribution of the estimated negative binomial shape parameter. The empirical value departs only slightly from the Monte Carlo distribution.

Finally, the last panel (f) shows the evolution of the share of reclassified patents, with the empirical data from Fig. 5 augmented by values of 1 between 1790 and 1899 (since no current categories existed prior to 1899, all patents have been reclassified). Here again, the model reproduces fairly well the empirical pattern. All or almost all patents from early years have been reclassified, and the share is falling over time. That said, for recent years (post 1976), the specific shape of the curve is different.

Overall, we think that given its simplicity the model reproduces a surprisingly high number of empirical facts. It allows us to understand the differences between the different patterns of growth of the reconstructed and historical number of classes. Without a built-in reclassification process it would not have been possible to match all these empirical facts – if only because without reclassification historical and reconstructed evolution coincide. This shows how important it is to consider reclassification when we look at the mesoscale evolution of the patent system.

8 Conclusion

In this paper, we have presented a quantitative history of the evolution of the main patent classes within the U.S. Patent Classification System. Our main finding is that the USPCS incurred regular and important changes. For academic researchers, these changes may be perceived as a source of problems, because this suggests that it may not

[28] An interesting alternative (instead of using the parameter α) would be to model the process by which patents grow and the process by which clerks split categories separately.

[29] There is a small inconsistency arising because the model is about primary classification only, but the historical number of classes and Heaps' law are measured using all classes, because we could not differentiate cross-reference classes in historical data. Another point of detail is that we could have used the estimated $C_0 = 0.17$ instead of the calculated one. These details do not fundamentally change our point.

always be legitimate to think that a given patent belongs to one and the same category forever. This means that results obtained using the current classification system may change in the future, when using a different classification system, and even if the very same set of patent is considered.

That said, we do not think the effect would be strong. Besides, using the current classification system is still often the best thing to do because of its consistency. Our point here is not to critique the use of the current classification, but to argue that historical changes to the classification system itself contain interesting information that has not been exploited. The changes of technological categories reflect the evolution of technology itself.

Our first result is that different methods to compute the growth of the number of classes give widely different results, establishing that the changes to the classification system are very important. Our second result suggests that we do not see very large categories in empirical data because categories are regularly split, leading to an exponential size distribution with no relationship between the age and size of a category. Our third result is that reclassification data contains useful information to understand technological evolution. Our fourth result is that a very simple model that can explain many of the observed patterns needs to include the splitting of classes and the reclassification of patents. Taken together, these results show that it is both necessary and interesting to understand the evolution of classification systems.

We believe these findings are interesting for all researchers working with economic and technological classifications, because we characterized quantitatively the volatility of the patent classification system. We do not know whether they are unstable because collective representations of technological artefacts are context-dependent, or because as more items are introduced and resources invested in classifying them appropriately, collective discovery of the "true" mesoscale partition takes place. But clearly one should bear in mind that they are dynamic when interpreting the results which rely upon them.

A case in point is the use of technological classes to produce forecasts: how can we predict the evolution of a given class or set of classes several decades ahead, when we know these classes might not even exist in the future? Even if we consider that today's categorization will not change, a subtle issue arises in the production of correct forecasting models. To see this consider developing a time series models describing the growth of some particular classes. To test the forecasting ability of the model, one should perform out-of-sample tests, as e.g. Farmer & Lafond (2016) did for technology performance time series. Part of the past data is used to predict more recent data, and the data which is not used for estimation is compared to the forecasts. Now, note that when we use the current classification, we effectively use data from the present; that is, the delineation of categories for past patents uses knowledge from the present, and it is therefore not entirely valid to evaluate forecasts (there is "data snooping" in the sense that one uses knowledge of the future to predict the future).

Classification system changes pose serious problems for forecasting but may also bring opportunities: if classification changes reflect technological change then one can in principle construct quantitative theories of that change. Since the patterns described here could be roughly understood using an extremely simple model, it may be possible to make useful forecasts with more detailed models and data. This could be useful because patent classification changes are more frequent than changes to other classification systems such as industries, products and occupation. An interesting avenue for future research would be to use the changes of the patent classification system to predict the changes of industry and occupation classification systems, thus predicting the types of jobs of the future.

Beyond innovation studies, with the rising availability of very large datasets, digitized and carefully recorded classifications and classification changes will become available. It will be possible to explore classifications as an evolving network and track the split, merge, birth and death of categories. This is an exciting new area of research, but the big data that we will accumulate will only (or mostly) cover the recent years. This makes historical studies such as the present one all the more important.

References

Acemoglu, D., Akcigit, U. & Kerr, W. R. (2016), 'Innovation network', *Proceedings of the National Academy of Sciences* p. 201613559.

Aharonson, B. S. & Schilling, M. A. (2016), 'Mapping the technological landscape: Measuring technology distance, technological foot-

prints, and technology evolution', *Research Policy* **45**(1), 81–96.

Alstott, J., Triulzi, G., Yan, B. & Luo, J. (2016), 'Mapping technology space by normalizing patent networks', *Scientometrics* pp. 1–37.

Andersen, B. (1999), 'The hunt for S-shaped growth paths in technological innovation: a patent study', *Journal of Evolutionary Economics* **9**(4), 487–526.

Antonelli, C., Krafft, J. & Quatraro, F. (2010), 'Recombinant knowledge and growth: The case of icts', *Structural Change and Economic Dynamics* **21**(1), 50–69.

Bailey, M. (1946), 'History of classification of patents', *J. Pat. Off. Soc'y* **28**, 463.

Basalla, G. (1988), *The evolution of technology*, Cambridge University Press.

Benner, M. & Waldfogel, J. (2008), 'Close to you? bias and precision in patent-based measures of technological proximity', *Research Policy* **37**(9), 1556–1567.

Berger, T. & Frey, C. B. (2015), 'Industrial renewal in the 21st century: evidence from us cities', *Regional Studies* pp. 1–10.

Bettencourt, L., Samaniego, H. & Youn, H. (2014), 'Professional diversity and the productivity of cities', *Scientific Reports* **4**, 5393.

Blackman, M. (2011), 'Classification news', *World Patent Information* **33**(3), 294.

Breschi, S., Lissoni, F. & Malerba, F. (2003), 'Knowledge-relatedness in firm technological diversification', *Research Policy* **32**(1), 69–87.

Bresnahan, T. F. & Trajtenberg, M. (1995), 'General purpose technologies 'engines of growth'?', *Journal of econometrics* **65**(1), 83–108.

Caminati, M. & Stabile, A. (2010), 'The pattern of knowledge flows between technology fields', *Metroeconomica* **61**(2), 364–397.

Carnabuci, G. (2013), 'The distribution of technological progress', *Empirical Economics* **44**(3), 1143–1154.

Cockburn, I. M., Kortum, S. & Stern, S. (2003), 'Are all patent examiners equal? examiners, patent characteristics, and litigation outcomes', *Patents in the Knowledge-Based Economy* p. 19.

Csárdi, G., Strandburg, K., Zalányi, L., Tobochnik, J. & Érdi, P. (2007), 'Modeling innovation by a kinetic description of the patent citation system', *Physica A: Statistical Mechanics and its Applications* **374**(2), 783–793.

Dopfer, K., Foster, J. & Potts, J. (2004), 'Micro-meso-macro', *Journal of Evolutionary Economics* **14**(3), 263–279.

Dosi, G. (1982), 'Technological paradigms and technological trajectories:: A suggested interpretation of the determinants and directions of technical change', *Research Policy* **11**(3), 147–162.

Érdi, P., Makovi, K., Somogyvári, Z., Strandburg, K., Tobochnik, J., Volf, P. & Zalányi, L. (2013), 'Prediction of emerging technologies based on analysis of the us patent citation network', *Scientometrics* **95**(1), 225–242.

Falasco, L. (2002), 'Bases of the United States Patent Classification', *World Patent Information* **24**(1), 31–33.

Farmer, J. D. & Lafond, F. (2016), 'How predictable is technological progress?', *Research Policy* **45**(3), 647–665.

Foucault, M. (1966), *Les mots et les choses*, Vol. 42, Gallimard Paris.

Freeman, C. & Soete, L. (1997), *The economics of industrial innovation*, third edn, MIT Press, Cambridge, MA.

Gibrat, R. (1931), *Les inégalités économiques*, Recueil Sirey.

Griliches, Z. (1990), 'Patent statistics as economic indicators: A survey', *Journal of Economic Literature* **28**(4), 1661–1707.

Hall, B. H., Jaffe, A. & Trajtenberg, M. (2005), 'Market value and patent citations', *RAND Journal of economics* pp. 16–38.

Hall, B., Jaffe, A. & Trajtenberg, M. (2001), 'The NBER patent citation data file: Lessons, insights and methodological tools'.

Heaps, H. S. (1978), *Information retrieval: Computational and theoretical aspects*, Academic Press, Inc.

Hicks, D. (2011), 'Structural change and industrial classification', *Structural Change and Economic Dynamics* **22**(2), 93–105.

Ijiri, Y. & Simon, H. A. (1975), 'Some distributions associated with bose-einstein statistics', *Proceedings of the National Academy of Sciences* **72**(5), 1654–1657.

Kim, D., Cerigo, D. B., Jeong, H. & Youn, H. (2016), 'Technological novelty profile and inventions future impact', *EPJ Data Science* **5**(1), 1.

Klepper, S. (1997), 'Industry life cycles', *Industrial and Corporate Change* **6**(1), 145–182.

Kutz, D. O. (2004), Examining the evolution and distribution of patent classifications, *in* 'Information Visualisation, 2004. IV 2004. Proceedings. Eighth International Conference on', IEEE, pp. 983–988.

Lafond, F. (2014), *The evolution of knowledge systems*, UNU-MERIT PhD Thesis, Maastricht University Press, chapter The size distribution of patent categories: USPTO 1976-2006, pp. 91—109.

Latour, B. (2005), *Reassembling the social: An introduction to actor-network-theory*, Oxford University Press, USA.

Leydesdorff, L. (2008), 'Patent classifications as indicators of intellectual organization', *Journal of the American Society for Information Science and Technology* **59**(10), 1582–1597.

Lin, J. (2011), 'Technological adaptation, cities, and new work', *Review of Economics and Statistics* **93**(2), 554–574.

Lü, L., Zhang, Z.-K. & Zhou, T. (2010), 'Zipf's law leads to heaps' law: Analyzing their relation in finite-size systems', *PloS one* **5**(12), e14139.

Lybbert, T. J. & Zolas, N. J. (2014), 'Getting patents and economic data to speak to each other: An algorithmic links with probabilities approach for joint analyses of patenting and economic activity', *Research Policy* **43**(3), 530–542.

Malerba, F. (2002), 'Sectoral systems of innovation and production', *Research Policy* **31**(2), 247–264.

Marco, A. C., Carley, M., Jackson, S. & Myers, A. F. (2015), 'The uspto historical patent data files: Two centuries of innovation', *SSRN 2616724* .

Marengo, L. & Zeppini, P. (2016), 'The arrival of the new', *Journal of Evolutionary Economics* **26**(1), 171–194.

McNamee, R. C. (2013), 'Can t see the forest for the leaves: Similarity and distance measures for hierarchical taxonomies with a patent classification example', *Research Policy* **42**(4), 855–873.

Nelson, R. (2006), Perspectives on technological evolution, *in* K. Dopfer, ed., 'The evolutionary foundations of economics', Cambridge University Press, pp. 461–471.

Nooteboom, B., Van Haverbeke, W., Duysters, G., Gilsing, V. & Van den Oord, A. (2007), 'Optimal cognitive distance and absorptive capacity', *Research policy* **36**(7), 1016–1034.

OTAF (1985), Review and assessment of the otaf concordance between the u.s. patent classification and the standard industrial classification systems: Final report., Technical report, Office of Technology Assessment, USPTO.

Paradise, J. (2012), 'Claiming nanotechnology: improving uspto efforts at classification of emerging nano-enabled pharmaceutical technologies', *Northwestern. Journal of Technology & Intellectual Property* **10**.

Pasinetti, L. L. (1983), *Structural change and economic growth: a theoretical essay on the dynamics of the wealth of nations*, CUP Archive.

Pavitt, K. (1984), 'Sectoral patterns of technical change: towards a taxonomy and a theory', *Research Policy* **13**(6), 343–373.

Pavitt, K. (1985), 'Patent statistics as indicators of innovative activities: possibilities and problems', *Scientometrics* **7**(1-2), 77–99.

Petersen, A. M., Rotolo, D. & Leydesdorff, L. (2016), 'A triple helix model of medical innovation: Supply, demand, and technological capabilities in terms of medical subject headings', *Research Policy* **45**(3), 666–681.

Reingold, N. (1960), 'Us patent office records as sources for the history of invention and technological property', *Technology and Culture* **1**(2), 156–167.

Rotkin, I. J., Dood, K. J. & Thexton, M. A. (1999), *A history of patent classification in the United States Patent and Trademark Office*, Patent Documentation Society.

Saviotti, P. (1996), *Technological Evolution, Variety, and the Economy*, Edward Elgar Pub.

Saviotti, P. P. & Pyka, A. (2004), 'Economic development by the creation of new sectors', *Journal of evolutionary economics* **14**(1), 1–35.

Scherer, F. (1984), Using linked patent and r&d data to measure interindustry technology flows, *in* 'R&D, patents, and productivity', University of Chicago Press, pp. 417–464.

Schmookler, J. (1966), 'Invention and economic growth'.

Schuett, F. (2013), 'Patent quality and incentives at the patent office', *The RAND Journal of Economics* **44**(2), 313–336.

Schumpeter, J. A. (1934), *The theory of economic development: An inquiry into profits, capital, credit, interest, and the business cycle*, Harvard University Press.

Silverberg, G. & Verspagen, B. (2003), 'Breaking the waves: a poisson regression approach to schumpeterian clustering of basic innovations', *Cambridge Journal of Economics* **27**(5), 671–693.

Silverberg, G. & Verspagen, B. (2007), 'The size distribution of innovations revisited: an application of extreme value statistics to citation and value measures of patent significance', *Journal of Econometrics* **139**(2), 318–339.

Simmons, H. J. (2014), 'Categorizing the useful arts: Part, present, and future development of patent classification in the united states', *Law Libr. J.* **106**, 563.

Simon, H. (1955), 'On a class of skew distribution functions', *Biometrika* **42**(3/4), 425–440.

Solé, R. V., Valverde, S., Casals, M. R., Kauffman, S. A., Farmer, J. & Eldredge, N. (2013), 'The evolutionary ecology of technological innovations', *Complexity* **18**(4), 15–27.

Stafford, A. B. (1952), 'Is the rate of invention declining?', *American journal of sociology* pp. 539–545.

Strumsky, D. & Lobo, J. (2015), 'Identifying the sources of technological novelty in the process of invention', *Research Policy* **44**(8), 1445–1461.

Strumsky, D., Lobo, J. & van der Leeuw, S. (2012), 'Using patent technology codes to study technological change', *Economics of Innovation and New Technology* **21**(3), 267–286.

Tëmkin, I. & Eldredge, N. (2007), 'Phylogenetics and material cultural evolution', *Current anthropology* **48**(1), 146–154.

Tria, F., Loreto, V., Servedio, V. & Strogatz, S. (2014), 'The dynamics of correlated novelties.', *Scientific reports* .

USPTO (2005), Handbook of classification, Technical report, USPTO.

Valverde, S., Solé, R., Bedau, M. & Packard, N. (2007), 'Topology and evolution of technology innovation networks', *Physical Review E* **76**(5), 056118.

Valverde, S. & Solé, R. V. (2015), 'Punctuated equilibrium in the large-scale evolution of programming languages', *Journal of The Royal Society Interface* **12**(107), 20150249.

Vernon, R. (1966), 'International investment and international trade in the product cycle', *The Quarterly Journal of Economics* **80**(2), 190–207.

Verspagen, B. (1997), 'Measuring intersectoral technology spillovers: estimates from the european and us patent office databases', *Economic Systems Research* **9**(1), 47–65.

Wang, C.-C., Sung, H.-Y. & Huang, M.-H. (2016), 'Technological evolution seen from the uspc reclassifications', *Scientometrics* pp. 1–17.

Wolter, B. (2012), 'It takes all kinds to make a world–some thoughts on the use of classification in patent searching', *World Patent Information* **34**(1), 8–18.

Xiang, C. (2005), 'New goods and the relative demand for skilled labor', *Review of Economics and Statistics* **87**(2), 285–298.

Youn, H., Bettencourt, L. M., Lobo, J., Strumsky, D., Samaniego, H. & West, G. B. (2016), 'Scaling and universality in urban economic diversification', *Journal of The Royal Society Interface* **13**(114), 20150937.

Youn, H., Strumsky, D., Bettencourt, L. M. A. & Lobo, J. (2015), 'Invention as a combinatorial process: evidence from us patents', *J. R. Soc. Interface* **12**, 20150272.

Yule, G. (1925), 'A mathematical theory of evolution, based on the conclusions of Dr. J.C. Willis, F.R.S.', *Philosophical Transactions of the Royal Society of London. Series B, Containing Papers of a Biological Character* **213**, 21–87.