Tracking the Evolution of Product Inventions with Trademark Data

Tsung-Han Chou, Po-Hsuan Hsu, Jui-Chung Yang*

January 25, 2024

Abstract

In the "identification" section of trademark documents, assignees need to clearly describe goods and services covered by their trademarks in ways that general readers can easily understand. We use Natural Language Processing (NLP) to identify the first appearance and subsequent follow-ups of product inventions disclosed in trademark identifications, and then propose several measures to evaluate the novelty of these product inventions. Our novelty measures offer timely information on invention value as they are not only available on filing dates, but also positively associated with trademarks' survival duration. More importantly, we construct a unique identifier that disambiguates all trademark assignees of trademarks, which allows us to evaluate companies' product inventions based on their trademark portfolios. We plan to make our datasets publicly available to help the research community explore the untapped potential of trademark documents.

1 Introduction

Companies rely on trademarks to distinguish their products and services from those of their competitors and also to protect their intellectual property. The modern US federal trademark system originates from the 1946 Lanham Act,[1] and the United States Patent and Trademark Office (USPTO) defines a trademark as "any word, name, symbol, device, or any combination, used or intended to be used to identify and distinguish the goods or services of one seller or provider from those of others, and to indicate the source of the goods or services."

To register a trademark with the USPTO, applicants must provide identification to document their

*Chou is from the College of Management of Technology, École Polytechnique Fédérale de Lausanne (tsung-han.chou@epfl.ch). Hsu is from the College of Technology Management, National Tsing Hua University (pohsuanhsu@mx.nthu.edu.tw), and Yang is from the Department of Economics, National Taiwan University (jcyang1225@ntu.edu.tw). We express our appreciation to Sam Arts, Bernhard Ganglmair, Michael Meurer, Ronja Christine Roettger, Yanzhi Wang, and all participants of the TPRI Brownbag Seminar and Taiwan Symposium on Innovation Economics and Entrepreneurship. The insightful discussions and exchanges greatly contributed to the development of this paper. We also gratefully acknowledge financial support from the National Science and Technology Council, Taiwan (NSTC 112-2410-H-002-236 for J.-C. Yang).

[1]See Internet Appendix A for basics about trademarks as well as laws and procedures for trademark applications, registrations, and renewal.

NIKE

Word Mark	NIKE
Goods and Services	IC 009. US 021 023 026 036 038. G & S: Downloadable virtual goods, namely, computer programs featuring footwear, clothing, headwear, eyewear, bags, sports bags, backpacks, sports equipment, art, toys and accessories for use online and in online virtual worlds IC 035. US 100 101 102. G & S: Retail store services featuring virtual goods, namely, footwear, clothing, headwear, eyewear sports bags, backpacks, sports equipment, art, toys and accessories for use online; on-line retail store services featuring virtual merchandise, namely, footwear, clothing, headwear, eyewear, bags, sports bags, backpacks, sports equipment, art, toys and accessories IC 041. US 100 101 107. G & S: Entertainment services, namely, providing on-line, non-downloadable virtual footwear, clothing, headwear, eyewear, bags, sports bags, backpacks, sports equipment, art, toys and accessories for use in virtual environments
Standard Characters Claimed	
Mark Drawing Code	(4) STANDARD CHARACTER MARK
Serial Number	**97095855**
Filing Date	October 27, 2021
Current Basis	1B
Original Filing Basis	1B
Owner	(APPLICANT) Nike, Inc. CORPORATION OREGON One Bowerman Drive Beaveron OREGON 97005
Attorney of Record	Jaime M. Lemons
Prior Registrations	3406594;4704670;5275562;AND OTHERS
Type of Mark	TRADEMARK. SERVICE MARK
Register	PRINCIPAL
Live/Dead Indicator	LIVE

Notes: Nike's new filed application. We focus on the goods and services portion, as outlined in the red rectangle.

Figure 1: Nike's newly filed trademark

products or services. These identifications are important for innovation research because their descriptions not only reflect firms' plans for product development but also may capture potential inventions.[2] For example, this identification for the trademark application "Nike" (No. 97095855, see Figure 1) on October 27, 2021, describes "downloadable virtual goods, such as footwear, clothing, and headwear," which highlights Nike's initiatives in the Non-Fungible Token (NFT) market. In another example, the identification for the trademark application "AWS" (No. 87604205) on September 12, 2017, which includes "software for cloud computing," symbolizes Amazon's entry into the cloud computing business. Meanwhile, the identification of the trademark application for "Meta" (No. 97097363) on October 28, 2021, which includes "virtual communities, social network, digital currency," reflects Mark Zuckerberg's ambition to create a metaverse.

Figure 1 presents the basic information about Nike's trademark for NFT (No. 97095855), including its appearance, its serial number, registration number (if registered), key dates (e.g., filing, registration, abandonment, cancellation), the registration status of trademarks (i.e., registered, abandoned, renewed, or canceled), and owner information. The section "Goods and Services" on top of the figure lists the identification in international classes 009, 035, and 041, which refer to electrical and scientific apparatus, advertising and business, and education and entertainment, respectively. For example, its identification for class 009 states "downloadable virtual goods, namely, computer programs featuring footwear, clothing, headwear, eyewear, bags, sports bags, backpacks, sports equipment, art, toys and accessories for use online and in online virtual worlds".

According to the Trademark Manual of Examining Procedure (TMEP), an identification should not include extra or unnecessary information, and must describe goods or services in ways that general readers can easily understand the goods or services themselves. Applicants can directly use the official ID Manual[3] to prepare identifications for their applications. However, when applicants find that their products or services are unique and do not match existing identifications in the USPTO ID Manual, they may "create" new keywords in identifications, which must meet the TMEP's standards

[2] https://tmep.uspto.gov/RDMS/TMEP/current#/current/TMEP-1400d1e1982.html

[3] The USPTO provides acceptable identifications of goods and services and information related thereto. https://www.uspto.gov/trademarks/guides-and-manuals/searching-trademark-id-manual.

of specificity, definiteness, clarity, accuracy, and conciseness. After applicants complete this process, a trademark attorney in the USPTO then reviews these new keywords/identifications and determines whether to approve them (or not).

In this paper, we use Natural Language Processing (NLP) to track the first appearance and follow-up of new keywords (which reflect product and service inventions) disclosed in trademark identifications. Specifically, we collect uni-grams (one word), bi-grams (two consecutive words), and tri-grams (three consecutive words) by tokenization, as well as remove stop words and stemming sentences from trademark identifications; we call our grams that result from this process "invention keywords." We then track the frequency of an invention keyword in trademark documents since its first appearance, and label it as a "successful" invention if it has been widely used by subsequent trademarks. We classify all invention keywords in the top 1%, top 1-5%, top 5-10%, and other groups to indicate their appearance frequencies in the decades in which they were created. Our analysis shows a skewed pattern: more successful inventions appear in a disproportionately larger number of trademarks compared to less successful ones.

We also examine the future development of invention keywords by examining their frequencies from one to three decades after their creation. We find that, on average, the frequencies of the top 1% group are 4 to 5 times higher than those of the top 1-5% group and 6 to 11 times higher than those of the top 5-10% group in the next decade. Our findings support the phenomenon of "winners-being-winners," which suggests that successful inventions tend to dominate the market. More importantly, we observe a clear trend: top groups grow even faster than bottom groups in subsequent decades, suggesting strong momentum in the market's acceptance of successful inventions.[4]

We then develop several novelty measures to evaluate the novelty and impact of product inventions. Different from conventional measures that reflect trademark value/importance from an *ex post* perspective (e.g., survival duration), our novelty measures are informative about trademark value from an *ex ante* perspective. Our novelty measures are validated through their positive correlations with trademark survival duration, which takes decades to determine. This finding supports the predictive ability of our novelty measures for trademark value.

Finally, and importantly, we "harmonize" the assignees of trademarks. Because the USPTO does not require applicants to use standardized names, one entity often has multiple different applicant names. We design a disambiguation process that utilizes both companies' self-reported information as blocking rules and similarities in assignee names to determine if different names belong to the same company. We also link our matched results to DISCERN to correct potential matching errors.[5] The unique identifier we construct not only allows researchers to obtain more precise and comprehensive knowledge of a company's trademark portfolio, but also enables researchers to link public firms to their

[4]Nevertheless, some successful inventions still fail in the decade after their creation.

[5]We use trademark assignment data, trademark owner data, and DISCERN (Arora et al., 2021b) data to identify and disambiguate trademark assignees. Internet Appendix B reports our source data, and Section 4 illustrates the process of disambiguation.

trademark portfolios using GVKEY through DISCERN.

This research note serves as a step forward toward fully exploiting the untapped potential of trademark data for studies in related disciplines (e.g., economics, innovation, management). The dataset we construct can be used to (i) investigate explanatory variables that influence the performance of product inventions, (ii) identify new driving forces of firms' product inventions, (iii) determine how patents and technological innovation shape product inventions and resultant market success, (iv) track the evolution and development of an invention, especially when "add-ons" of special invention keywords (e.g., "XX + cryptocurr") are used, and (v) examine the similarity in invention keywords of new trademarks from peer firms to monitor changes in industry structure and competition dynamics.

Previous studies have adopted NLP to identify the creation and impact of new technologies. For example, Arts et al. (2021) uses NLP to detect novel keywords in US patent documents, revealing new technologies and evaluating the novelty of patents at the time of filing. Arts et al. (2023) uses NLP to identify new scientific ideas at the time of publication in scientific journals, and evaluate their impacts on subsequent research. Kelly et al. (2021) also uses NLP to develop new indicators for identifying critical patents. However, it is well-known that patents only cover a limited scope of technologies due to various factors (e.g., patentability) and cannot therefore capture inventions in all sectors (e.g., service industries).[6] Our paper bridges this gap in the literature by demonstrating the application of NLP to trademark documents and highlighting valuable (and even unique) features of trademark information with respect to innovation research.

This research note also demonstrates the use of blocking methods in disambiguation, which may capture more potential matches and improve efficiency. Previous research, for example, uses iterative blocking schemes on patent assignees (Li et al., 2014) or probabilistic blocking methods (Steorts, 2015; Steorts, 2016; Marchant et al., 2021). However, our approach differs from these studies by incorporating additional information, such as "name change" and "correction" events in trademark assignee data, and then connecting to the DISCERN data (Arora et al., 2021a; Arora et al., 2021b).[7] This method allows us to track name and ownership changes over time and may be applied to future matching practices.

It is also noteworthy that our foundation – trademark documents – are advantageous in comparison with other data sources for product inventions in the following ways. Trademark data provide a broader scope and are more objective and comprehensive than firms' new product announcement in news or financial statements (Mukherjee et al., 2017; Hoberg and Phillips, 2010). Moreover, trademark data cover a much wider range of categories than retail sales data like Nielsen's Retail Scanner (Argente

[6] According to the 2008 U.S. National Science Foundation's Business R&D and Innovation Survey (NSF BRDIS), 60% of R&D firms rate trademarks as very important, whereas 41%, 33%, 50%, and 67% of such firms rate utility patents, design patents, copyrights, and trade secrets as very important, respectively. Thus, Hall et al. (2014) concludes that registered trademarks are probably the most widely used form of IP protection, as they apply to essentially any product or service.

[7] Additionally, Dinlersoz et al. (2021) matches trademark assignees to the U.S. Census Bureau's Business Register by using name and address.

et al., 2020; Aparicio et al., 2021).

The data and measures we construct and propose enable researchers to capture the evolution of various inventions in products and services. Firms may intentionally not file patents for their inventions because of the business secret, patentability, market strategy, or legal considerations. However, firms will still launch new product lines and file trademarks to protect these inventions. Therefore, when studying innovations, trademarks can be more advantageous than patents and other intellectual property forms. These arguments are supported by the literature: trademarks are the most widely used form of IP protection (Hall et al., 2014),[8] and have been used by firms to protect their new products or services from imitation (Millot, 2009), to market their new products or services (Gao and Hitt, 2012; Flikkema et al., 2019), or to maintain their market power and customer loyalty (Block et al., 2015). Thus, several prior studies have used the number of (new) trademarks as a proxy for firm-level product innovations,[9] However, different from prior studies that use the simple count of trademarks, our measures not only track the the evolution of each specific invention but also reflect the novelty and quality of each trademark.

The remainder of this paper is structured as follows. In Section 2, we describe how we identify the creation of invention keywords. In Section 3, we discuss the measures we construct to capture the novelty of invention keywords and trademarks. In Section 4, we introduce the process of disambiguation of trademark assignees. Section 5 discusses the applications and possible extensions of our dataset. We leave all technical details for our the Internet Appendix, which includes the background information of trademarks in Section A, the details of our data collection in Section B, the details of the USPTO trademark case file dataset in Section C, the details of the USPTO trademark assignment dataset in Section D, the details of the DISCERN dataset in Section E, and the process we use to identify and develop measures for invention keywords, along with examples and summary statistics, in Section F.

2 Invention Keywords

2.1 Identifying an invention keyword and its creation time

We use trademark identifications to identify the creation and development of new inventions by implementing the method proposed in Arts et al. (2021) for patents. First, for each trademark's identification, we tokenize the text to uni-grams, bi-grams, and tri-grams. Next, we remove tokens only composed of numbers, one-character words, and stop words.[10] We then apply stemming to each token,

[8] According to the U.S. National Science Foundation's new Business R&D and Innovation Survey (BRDIS) survey, among all firms in the survey, 15% firms answer trademarks being important in the protection of their IP (utility patents 5%, design patents 6%, copyrights 12%, and trade secret 14%).

[9] See Mendonça et al. (2004); Hipp and Grupp (2005); Greenhalgh and Rogers (2006); Sandner and Block (2011); Flikkema et al. (2014); Schautschick and Greenhalgh (2016); Faurel et al. (2017); Crass et al. (2019); Crown et al. (2020); Nasirov (2020); Hsu et al. (2022a). Hsu et al. (2022b), on the other hand, use new trademarks to measure firms' market entry.

[10] Our stop words list consists of Python Natural Language Toolkit (NLTK) stop words, stop words provided by Arts et al. (2021), and introductory words used in trademarks. Examples of stop words from NLTK include the following: am, are, is, the, of, any. The USPTO trademark examining procedure suggests using these definite terms to further define introductory wording. https://tmep.uspto.gov/RDMS/TMEP/current#/current/TMEP-1400d1e2196.html. Introductory

which is a procedure to reduce words to their respective word stems. For example, both "cryptocurrency" and "cryptocurrencies" are stemmed into "cryptocurr."

Furthermore, we eliminate n-grams that appear in fewer than 10 trademarks in all years. We exclude infrequent n-grams because they might not refer to specific inventions.[11] The n-grams that remain, created after 1980, are "invention keywords." After we follow the process described here, our identified invention keywords consist of 15,702 uni-grams, 406,437 bi-grams, and 645,579 tri-grams. In Internet Appendix F.1, we demonstrate the entire process of the NLP algorithm.

The creation year of an invention keyword is defined by the filing year of the first trademark that uses it. Additionally, the corresponding decade is referred to as the "creation decade." For example, the first trademark to use ('cryptocurr,' 'payment') is No. 79163995, filed in 2014. Therefore, the creation year of ('cryptocurr,' 'payment') is 2014, and the corresponding creation decade is the 2010s.

2.2 Prevalence and impact of invention keywords

We then evaluate an invention keyword's prevalence and impact by using the following three measures to analyze its frequencies: (1) document frequency ($document\_freq\_ngrams$) refers to the number of trademarks that contain a particular invention keyword, (2) class frequency ($class\_freq\_ngrams$) refers to the number of unique international classes of trademarks that contain a specific invention keyword, and (3) overall frequency (all_freq_ngrams) refers to the number of times that an invention keyword is mentioned in all trademarks documents.[12] Specifically, we denote the impact of uni-grams as doc_freq_1grams, $class\_freq\_1grams$, and all_freq_1grams; denote the impact of bi-grams as doc_freq_2grams, $class\_freq\_2grams$, and all_freq_2grams; and denote the impact of tri-grams as doc_freq_3grams, $class\_freq\_3grams$, and all_freq_3grams. We present the summary statistics in Appendix F.3.

For instance, Twitter's trademark (No. 77721751, see Figure 2) contains three identifications: IC 038, 041, and 045. The identification in IC 038 mentions "social network" only once, while the identification in IC 045 mentions the phrase three times. Hence, "social network" weights once in doc_freq_2grams, twice in $class\_freq\_2grams$, and four times in all_freq_2grams (see more details in Appendix F.2). By following this rule for all trademarks, we determine that the doc_freq_2grams, $class\_freq\_2grams$, and all_freq_2grams of "social network" are 26,993, 32,527, and 49,386, respectively. These frequency measures reflect the frequency of bi-grams term usage across all trademarks and serve as proxies for the impact of invention keywords. The same idea is applied to uni-grams and tri-grams.

Table 1 lists the most frequent invention keywords by document frequency in the 1980s, 1990s, 2000s,

words in the USPTO include the following: namely, consisting, particularly.

[11] For example, we eliminate the bi-grams ('reduc,' 'return') due to its low frequency. This bi-grams can be derived from two different objects:"Metal Pipe Fittings-Namely, Elbows, Tees, Crosses, Couplings, *Reducers, Return* Bends, Caps, Locknuts, Flanges, Bushings, Plugs, Laterals, Y-Branches, Adapters" and "mailing list processing services, namely, checking addresses and postal regulations to *reduce return* mail, provide the best price and eliminate duplication of names."

[12] Note that all_freq may exceed $class\_freq\_ngrams$ and $class\_freq\_ngrams$ may exceed doc_freq_ngrams, if a trademark contains multiple international classes and/or if a n-gram term appears multiple times within one class.

Word Mark	**TWITTER**
Goods and Services	IC 038. US 100 101 104. G & S: Telecommunications services, namely, providing online and telecommunication facilities for real-time interaction between and among users of computers, mobile and handheld computers, and wired and wireless communication devices; enabling individuals to send and receive messages via email, instant messaging or a website on the internet in the field of general interest; providing on-line chat rooms and electronic bulletin boards for transmission of messages among users in the field of general interest; providing an online community forum for users to share information, photos, audio and video content about themselves, their likes and dislikes and daily activities, to get feedback from their peers, to form virtual communities, and to engage in social networking. FIRST USE: 20061201. FIRST USE IN COMMERCE: 20061201 IC 041. US 100 101 107. G & S: Providing on-line journals, namely, blogs featuring user-defined content. FIRST USE: 20061201. FIRST USE IN COMMERCE: 20061201 IC 045. US 100 101. G & S: Online social networking services; providing a website on the internet for the purpose of social networking; providing on-line computer databases and on-line searchable databases in the field of social networking. FIRST USE: 20061201. FIRST USE IN COMMERCE: 20061201
Mark Drawing Code	(5) WORDS, LETTERS, AND/OR NUMBERS IN STYLIZED FORM
Serial Number	77721751
Filing Date	April 24, 2009

Figure 2: Twitter's trademark. The figure displays the TESS document for No. 77721751. The document records the word mark, descriptions of goods and services, serial number, and filing date.

and 2010s; each invention keyword itself is reduced to its word stem in Table 1. For instance, the 1980s saw the creation of "comput network" and "global comput network" with frequencies of 160,520 and 109,304, respectively.

2.3 Growth of important invention keywords

To show the evolution of innovation over the past few decades, we list noteworthy inventions and their document frequencies in trademark documents from the 1980s to the 2010s in Tables 2.

We show examples such as "laptop," "e-book," "biotechnolog," "comput network," "mobil phone," "databas manag," "comput game program," "comput hardwar comput," and "health care cost," which were all created in the 1980s. Their frequencies steadily increased over time. The same trend can be observed for invention keywords from the 1990s and 2000s, such as "smartphon," "3d," "e-commerc," "internet-bas," "podcast," "cloud-bas," "cybersecur," "teleseminar," "websit featur," "on-lin retail," "social network," "cell phone," "cloud comput," "social media," "blog featur," "phone tablet," "on-lin retail store," "download comput softwar," "on-lin chat room," "portabl media player," "onlin social network," and "book mark transmiss." However, initial success does not guarantee long-term prevalence. For example, the frequencies of "cd-rom," "on-lin chat room," and "electron retail servic" show strong growth in the first two decades, but then later declined.

2.4 The heterogeneity in the prevalence of invention keywords

To understand the diffusion and prevalence of product and service inventions, we track the cross-sectional distribution of the frequencies of invention keywords in trademark documents for subsequent decades. We categorize invention keywords into eight frequency groups based on their document frequency in the decade of their creation: the top 1%, the top 1-5%, the top 5-10%, the top 10-20%, the top 20-30%, the top 30-40%, the top 40-50%, and the bottom 50%. We first focus on the

Table 1: Top invention keywords for each decade (with document frequency in parentheses)

1980-1989:
uni-grams: saa (38452), laptop (26967), cd-rom (20087), searchabl (16833), upload (16501), desktop (16191). **bi-grams:** comput network (160520), global comput (121118), web site (77748), mobil phone (63198), non-download softwar (45770), download comput (43458). **tri-grams:** global comput network (109304), servic saa servic (30470), comput game softwar (29208), web site featur (28139), servic featur softwar (28057), saa servic featur (26332).
1990-1999:
uni-grams: smartphon (37066), usb (28649), hoodi (20435), web-bas (18975), mp3 (18357), 3d (16259). **bi-grams:** websit featur (78272), on-lin retail (53823), cell phone (35133), social network (26993), media player (23380), search engin (20847). **tri-grams:** on-lin retail store (48396), onlin retail store (31800), download comput softwar (26840), websit featur inform (25582), download electron public (21001), jacket footwear hat (19037).
2000-2009:
uni-grams: podcast (17306), webinar (10121), cloud-bas (5728), sharabl (3601), usb-pow (2704), pilat (2091). **bi-grams:** cloud comput (14923), download mobil (14850), blog featur (14758), yoga pant (11798), blank usb (10022), wireless charger (9043). **tri-grams:** portabl media player (20821), websit featur non-download (13605), social network servic (11739), overal sleepwear pajama (11396), sleepwear pajama romper (11378), jumper overal sleepwear (11348).
2010-2021:
uni-grams: smartwatch (9486), smartglass (4267), selfi (3737), e-liquid (3269), jeg (1642), iot (1625). **bi-grams:** wearabl activ (6018), children treat (4564), smartphon protect (4527), toy drone (4357), smartphon mount (4141), babi adult (3695). **tri-grams:** wearabl activ tracker (6004), wireless charger wireless (4054), adult children women (3653), viral communic channel (3639), electron cigarett liquid (3603), babi adult children (3570).

Notes: The table shows the decade that invented words were created. The frequency of each word appearing in trademark documents is indicated in parentheses. Invention keywords have been stemmed (i.e., reduced to their respective root form). For further information about the stemming process, please refer to Appendix F.1.

Table 2: Examples for invention keywords

Panel A: Examples for uni-grams invention keywords				
Keywords	*doc*_1980*s*	*doc*_1990*s*	*doc*_2000*s*	*doc*_2010*s*
laptop	16	231	2321	243994
cd-rom	11	5052	9253	5771
e-book	1	1	310	7581
biotechnolog	129	642	2511	5156
smartphon	0	1	178	36887
3d	0	128	758	15373
e-commerc	0	368	2714	6620
internet-bas	0	18	1102	6444
podcast	0	0	1624	15682
cloud-bas	0	0	12	5716
cybersecur	0	0	5	1501
teleseminar	0	0	147	934
Panel B: Examples for bi-grams invention keywords				
Keywords	*doc*_1980*s*	*doc*_1990*s*	*doc*_2000*s*	*doc*_2010*s*
comput network	255	51790	115939	126653
mobil phone	12	241	9747	81884
databas manag	154	4814	14107	23444
laser printer	164	1001	1315	886
websit featur	0	434	13737	101281
on-lin retail	0	4454	22027	55478
social network	0	11	7340	42098
cell phone	0	84	4419	37966
social media	0	0	520	25814
cloud comput	0	0	165	21082
blog featur	0	0	3120	16847
phone tablet	0	0	36	9532
Panel C: Examples for tri-grams invention keywords				
Keywords	*doc*_1980*s*	*doc*_1990*s*	*doc*_2000*s*	*doc*_2010*s*
comput game program	524	4131	9354	17938
comput hardwar comput	34	2883	8064	11797
health care cost	78	389	798	1469
prerecord comput program	1038	298	168	88
on-lin retail store	0	1913	17865	51808
download comput softwar	0	254	4251	28560
on-lin chat room	0	711	4015	3442
electron retail servic	0	676	917	107
portabl media player	0	0	1038	26109
onlin social network	0	0	1561	9996
book mark transmiss	0	0	2865	4104
rice hot dog	0	0	18	14

Notes: This table shows noteworthy uni-grams, bi-grams, and tri-grams, along with their respective frequencies of occurrence in documents every ten years.

prevalence of these frequency groups in their creation decades. We report our summary statistics in Internet Appendix F.3. As illustrated in Figures F1, F2, and F3 in the Internet Appendix, we find a skewed distribution in document frequencies of these invention keywords in their creation decades, which indicates that only a small number of inventions become very successful.

Second, we examine the future prevalence of invention keywords (and frequency groups) by examining their frequencies in the next decade (we include all details in Internet Appendix F.4). Figures F4, F5, and F6 present the prevalence of different frequency groups in the next decade. On average, the document frequencies of the top 1% of invention keywords are 4 to 5 times greater than those in the top 1-5% group, and 6 to 11 times greater than those in the top 5-10% group. This supports the "winners-being-winners" phenomenon, as the most successful inventions tend to dominate markets in the future, as evidenced by their continued prevalence in the next decade.

Third, we find a clear trend that the top groups grow *faster* than the bottom groups, as presented in Internet Appendix F.5. Using bi-grams created in the 1980s as an example, Figures F7, F8, and F9 in the Internet Appendix show the long-term development of invention keywords. For the top 1% group, the average frequencies in the 1980s, 1990s, 2000s, and 2010s were 171, 1,844, 3,839, and 5,901, respectively, while the average frequencies of the bottom 50% group in the 1980s, 1990s, 2000s, and 2010s were 2, 5, 19, and 64, respectively. This pattern of successful invention keywords becoming even more prevalent and growing even faster suggests momentum in the market's acceptance of successful product inventions.

3 Trademark Novelty

We next turn to the construction of measures for the novelty of trademarks based on the combination, importance, and age of new invention keywords. Different from conventional measures (e.g., survival duration) that reflect the value of trademarks from an *ex post* perspective, our new measures provide valuable information about trademark value from an *ex ante* perspective.

3.1 New invention keywords and the novelty of trademarks

We construct our first and second measures, *novelty counts* and *novelty ratio*, by counting the number of "new" invention keywords in a trademark's identification. A new invention keyword is an invention keyword that has been used for the first time within *five years* before the filing year of the trademark that we wish to capture its novelty.

Novelty counts is obtained by counting the number of new invention keywords in a trademark's identification. A larger number of new invention keywords in a trademark implies more use of frontier inventions and a higher degree of trademark novelty.

Novelty ratio is derived by dividing the number of new invention keywords by the total number of invention keywords in a trademark's identification. In contrast to *novelty counts*, *novelty ratio* measures

Table 3: Identified invention keywords in No.87561314

No.87561314: CLIMATECOIN, filed in 2017.
Indentification: Computer software platforms for purchasing and managing cryptocurrency; Downloadable mobile applications for purchasing and managing cryptocurrency
Invention keywords: 6
uni-grams: cryptocurr **bi-grams:** download mobil, softwar platform, manag cryptocurr, cryptocurr download **tri-grams:** comput softwar platform
New invention keywords (2013-2017): 3
uni-grams: cryptocurr (2013) **bi-grams:** cryptocurr download (2014), manag cryptocurr (2016)

Notes. We present in this table all the identified invention keywords and new invention keywords in trademark No. 87561314. The trademark contains 6 invention keywords, 3 of which are new invention keywords (i.e., "cryptocurr," "cryptocurr download," and "manag cryptocurr"). The creation year of each invention keyword is shown in parentheses.

the weight of frontier inventions in a trademark.

As an example, we use trademark No. 87561314, which was filed in 2017. Table 3 presents invention keywords and new invention keywords for trademark No. 87561314. The original identification for this trademark is "Computer software platforms for purchasing and managing cryptocurrency; Downloadable mobile applications for purchasing and managing cryptocurrency." Among the 6 invention keywords, 3 new invention keywords were first created between 2013 and 2017: "cryptocurr," "cryptocurr download," and "manag cryptocurr." Thus, *novelty counts* of No.87561314 is 3, and *novelty ratio* of No.87561314 is $3/6 = 0.5$.

The lag between the filing year of a trademark and the creation years of invention keywords that it contains also serves as a measure of the trademark's novelty. For example, No. 87812751, filed in 2018, includes an identification for "Letter bulletin boards for household use; Picture frames," with all invention keywords created before the 2010s. In contrast, No. 86582468, filed in 2015, includes an identification for "Dashboard cameras." Because its invention keyword, "dashboard camera," was created in the 2010s, No. 86582468 can be regarded as more novel than No. 87812751. We thus propose the indicator *Freshness*.

Freshness is defined as 41 - the average lag (i.e., filing year of the trademark - creation year of invention keyword).[13] The third column of Table 4 lists the difference between every invention keyword's creation year in No.87561314. No. 87561314 has the filing year of 2017, and the average creation year of its 6 invention keywords is 2002.67, resulting in a *freshness* of 26.67. In contrast, No. 90383155 has the filing year of 2020, and the average creation year of its 4 invention keywords is 1985, resulting in a *freshness* of 6. This comparison implies that No. 87561314 incorporates newer inventions compared to No. 90383155.

We note that not all trademarks have positive novelty values. For instance, trademark No. 78689193

[13]The upper bound of this measure is set to the full range of our data coverage (i.e., 41).

Table 4: Creation years and lags of invention keywords in No. 87561314 and 90383155

Panel A: No. 87561314:		
Keywords	Creation Year	Lag (Filing Year - Creation Year)
cryptocurr	2013	4
download mobil	2000	17
softwar platform	1990	27
manag cryptocurr	2016	1
cryptocurr download	2014	4
comput softwar platform	1990	27
Panel B: No. 90383155:		
Keywords	Creation Year	Lag (Filing Year - Creation Year)
sock polo	1985	35
t-shirt hat pant	1988	32
sock polo shirt	1985	35
polo shirt pullov	1982	38

Notes: The first column lists invention keywords, the second column displays their creation year, and the third column displays the difference between the filing year and the creation year of invention keywords. A lower value indicates that the invention keyword is relatively newer compared to the given trademark.

of APTIV DIGITAL (see Table 5), which describes "Interactive video software applications for use on set-top boxes and equipment connected to those set-top boxes in the corresponding digital network, for cable TV, satellite, and telephone networks," has 7 invention keywords, which we use to calculate its *freshness* of 22.75. However, it has zero values for *novelty counts* and *novelty ratio*, as it does not contain any new invention keywords.[14]

3.2 Survival duration

To empirically examine if the novelty measures we proposed are related to the value of trademarks, we also calculate the conventional, retrospective measure *survival duration*.

Survival duration is the difference between a trademark's registration and cancellation dates.[15] If a trademark lacks a cancellation record, we use its latest recorded date (e.g., renewal date) instead.[16] For instance, trademark No. 78194886 was registered on October 21, 2003, and it was renewed on October 23, 2013, which represents a survival duration of 10.01 years.[17] Trademark No. 85758173 was registered on January 21, 2014, and it was cancelled on August 28, 2020, which represents a survival duration of 6.6 years.[18] As long as *survival duration* indicates a trademark's vintage, No. 78194886 is

[14]Overall, 57.08% of trademarks have zero values for *novelty counts* and *novelty ratio*, while only 148 of trademarks have zero values of *freshness*.

[15]The lasting time of a non-registered trademark is zero.

[16]We use the date of its latest record if the status code is in the following list: $400-406, 411, 412, 414-417, 600-614, 618, 622, 626, 632, 709-715, 781-782, 900-901$, and 970.

[17]No. 78194886 was renewed on October 23, 2013, with a cfh status code of 800. For cfh status code definitions, see Table 1 of the Trademark Applications Documentation (TAD).

[18]No. 85758173 was cancelled on August 28, 2020, with cfh status code 710.

Table 5: Examples of resulting data

Serial No.	Year	Novelty Counts	Novelty Ratio	Freshness	Survival Duration
78194886	2002	5	0.71	32	10.01
76330388	2001	2	0.33	14.2	9.99
77721751	2009	4	0.09	25.69	6.6
78211892	2003	22	0.29	19.53	9.96
78689193	2005	0	0	22.75	6.6
75301493	1997	1	1	20.5	22.12

Notes: This table shows example data for measures of trademark novelty. The first column lists the trademark number, while the second column indicates the year when the trademark was filed. The remaining columns contain our constructed measures, including *novelty counts, novelty ratio, freshness*, and *survival duration*.

more valuable than No. 85758173.[19]

While *survival duration* can accurately evaluate whether a trademark is successful, it takes decades for this measure to be calculated. Thus, it is not available to researchers and investors for real-time analyses. In contrast, our novelty measures may capture trademark value in a more timely matter. In fact, *novelty counts*, *novelty ratio*, and *freshness* can measure a trademark's value at the time of its filing.

3.3 Novelty and survival duration

In Table 6, we report summary statistics for our measures for trademark-level novelty. Table 6 Panel A clearly presents the right-skewed nature of novelty. For the *novelty counts*, its minimum, 25^{th} percentile, and median are all zero, the 75^{th} percentile is 2, and the maximum is 1,616. For the *novelty ratio*, its minimum, 25^{th} percentile, and median are all zero; in addition, the 75^{th} percentile is 0.18, and the maximum is 1.

We also report the summary statistics of *freshness*, which is not as skewed as *novelty counts* or *novelty ratio*. The mean and the standard deviation of *freshness* are 15.10 and 6.65. The minimum, 25^{th} percentile, median, 75^{th} percentile, and maximum are 0, 10.33, 14.45, 19.42, and 41, respectively. Finally, with respect to *survival duration*, the mean and standard deviation are 5.23 and 5.58. The minimum, 25^{th} percentile, median, 75^{th} percentile, and maximum are 0, 0, 6.52, 6.76, and 40.66, respectively.

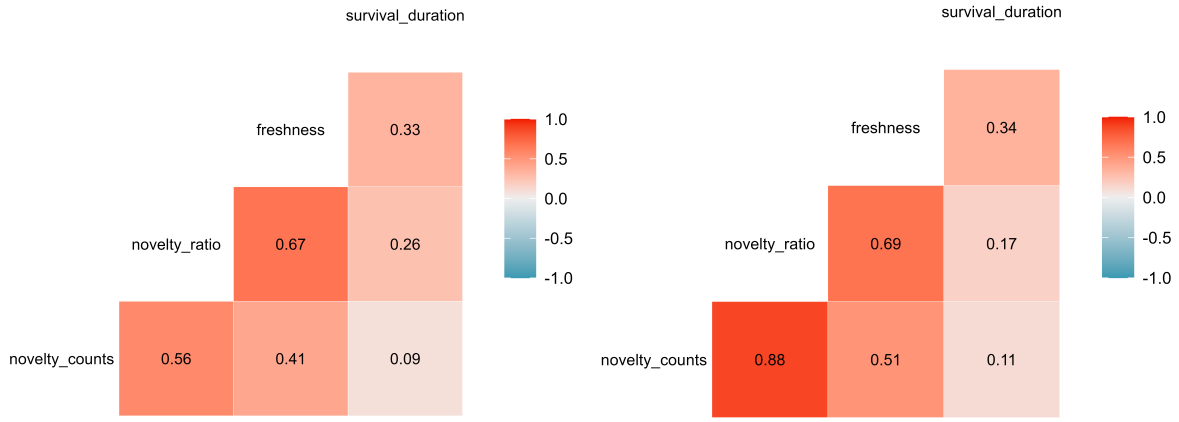
To compare our measures with *survival duration*, we use a simple correlation analysis on trademarks by using at least one invention keyword. Figure 3 displays the Pearson coefficients and Spearman correlation coefficients of our measures to each other and *survival duration* after a log transformation. We find that *novelty counts* has a positive correlation with *survival duration*, resulting in a 0.09 Pearson correlation coefficient and a 0.11 Spearman correlation coefficient. *Novelty ratio* also has a

[19] We exclude 81 trademarks with a negative survival duration, as they may have been abandoned following an inter partes decision by the Trademark Trial and Appeal Board, or were cancelled due to International Registration being cancelled in whole or in part, among other reasons.

Table 6: Summary statistics of trademark novelty

Panel A: Summary statistics for all trademarks							
Variable	Mean(sd)	Min.	1st Qu.	Median	3rd Qu.	Max.	Count
novelty counts	2.19(10.59)	0	0	0	2	1616	3646085
novelty ratio	0.15(0.26)	0	0	0	0.18	1	3646085
freshness	15.10(6.65)	0	10.33	14.45	19.42	41	3646085
survival duration	5.23(5.58)	0	0	6.52	6.76	40.66	3646085
Panel B: Summary statistics for trademarks with non-zero values							
Variable	Mean(sd)	Min.	1st Qu.	Median	3rd Qu.	Max.	Count
novelty counts	5.09(15.70)	1	1	2	5	1616	1056492
novelty ratio	0.34(0.31)	0.0008	0.11	0.22	0.5	1	1056492
freshness	15.1(6.65)	0.08	10.33	14.45	19.41	41	2216243
survival duration	8.77(4.59)	0.003	6.61	6.7	10.03	40.66	3646085

Notes: Overall, 57.08% of trademarks have zero values for *novelty counts*, *novelty ratio*, and *freshness*. For each variable, the sample average and standard deviation are presented as Mean (sd) in parentheses. Additionally, the minimum is presented as Min., the 25^{th} percentile as 1st Qu., the 50^{th} percentile as median, the 75^{th} percentile as 3rd Qu., and the maximum as Max. The last columns reports the number of valid trademarks.



(a) Pearson correlation

(b) Spearman correlation

Figure 3: Correlation coefficients. We present the Pearson correlation coefficients and Spearman correlation coefficients of our measures after log-transformation.

positive correlation with *survival duration*, resulting in a 0.26 Pearson correlation coefficient and a 0.17 Spearman correlation coefficient. Moreover, *freshness* exhibits a positive correlation with *survival duration*. The Pearson and Spearman correlation coefficients between *freshness* and *survival duration* are 0.33 and 0.34, respectively. These correlation coefficients suggest that our novelty measures could predict trademark value, especially given that our novelty measures are known on the actual filing dates themselves.

4 Disambiguation of Assignees

In this section, we report how we harmonize trademark assignees to solve inconsistent and confusing company names that result from the lack of standardization in trademark datasets. For example, Tesla holds trademark No. 88702372 and No. 85541117, but the assignee for No. 88702372 is listed as "Tesla,

Table 7: Examples of original and standardized words

Standardized words: corp **Original words:** corporation, corporaton, cororation, corporations, cooperative.
Standardized words: acq **Original words:** acquisition, acquisitions.
Standardized words: ind **Original words:** industry, industries, industreis.
Standardized words: invest **Original words:** investment, invesetment, investments, invesetments.
Standardized words: edu **Original words:** education, educational, educatioal.

Note. This table shows part of the standardized words and their corresponding original words. **Standardized words** shows standardized words themselves, and **Original words** lists their corresponding original words. The entire list contains 416 corrected elements, which can be accessed in our code.

Inc." and the assignee of No. 85541117 is "Tesla Motors, Inc." We construct unique firm identifiers and fix spelling errors by using a four-step process: 1) standardization, 2) blocking, 3) matching, and 4) combining with the DISCERN dataset.

4.1 Name standardization

We first convert owner names into lowercase. Next, we create a list of standardized names and then correct all typos and different representations. For example, as shown in Table 7, "acquisition" and "acquisitions" are standardized to "acq." Also, "corporation," "corporaton," "cororation," "corporations," and "cooperative" are standardized to "corp." The entire list contains 416 corrected elements, which can be accessed in our code. Third, we remove punctuation.

4.2 Blocking

We implement two blocking rules to optimize our disambiguation process: assignment records and serial numbers. Our blocking rules aim to capture as many potential matches as possible, which not only enhances the efficiency of disambiguation but also allows us to track changes in companies' ownership across different regions and periods.

4.2.1 Assignments

We start our process by gathering information on name changes and corrections for misspellings through assignment data. This step involves extracting assignor-assignee pairs from assignment records in the categories of name change and correction conveyance.[20] Table 8 shows examples of assignor-assignee pairs in different conveyance groups.

4.2.2 Serial number

Second, we group owners by serial number. **Owner** data records the name and location of owners at different stages (e.g., application, publication, registration, new owners after registration). For

[20] See Graham et al. (2018) for details.

Table 8: Blocking by assignment records

Conv Group	Assignor	Assignee
name change	medassets hsca inc	medassets supply chain sys inc
name change	steiner co inc	steiner corp
correction	autobytelcom inc	autobytel corp
correction	vinco hld sa	vinco mt

Notes: The "Conv group" presents the type of conveyance in the assignment record, which can be a name change, merger, assignment, or correction. The "Assignor" presents the name of the assignor, while the "Assignee" presents the name of the assignee.

Table 9: Blocking by serial number

Serial No.	City	Postal Code	Owner Name
73419517	trubschachen	3555	kambly sa spcl de biscuits suisses
73419517	trubschachen berne		kambly sa spcl de biscuits
87080898	palo alto	94304	tesla inc
87080898	palo alto	94304	tesla motors inc

Notes: "kambly sa spcl de biscuits suisses" and "kambly sa spcl de biscuits" are in the same block; "tesla inc" and "tesla motors inc" are also in the same block. Serial no. presents the trademark number. City is the city of the owner. Postal code is the postal code of the owner. Owner name is the owner's name listed in the given trademark number.

instance, the original registrant of No. 73419517 is "kambly sa spcl de biscuits suisses." The subsequent owners after registration are "kambly sa spcl de biscuits," with recorded locations of "trubschachen" and "trubschachen berne." These records refer to the same company but include different spellings and locations. The use of the serial number allows us to group different spellings from different locations. As in these examples from Table 9, No. 73419517 is a block that contains "kambly sa spcl de biscuits suisses" and "kambly sa spcl de biscuits," and No. 87080898 is another block that contains "tesla inc" and "tesla motors inc."

4.3 Matching

To determine whether two names belong to the same company, we calculate similarities from their assignee names and create matched pairs. We then cluster the pairs to represent unique companies and adjust the thresholds until the correct elements are included in the top 1000 clusters. Within each block, we calculate Jaro-Winkler similarity scores for each pair of names, and consider pairs with scores above the pre-determined threshold as those that belong to the same company. The threshold is set at 0.85 for assignor-assignee pairs in the "name change" and "correction" groups. For the serial number blocks, we also set the threshold of the Jaro-Winkler similarity to 0.85.[21]

We link all matched pairs and assign them a unique identifier called "tm_dedup_id." We call the outcome from this process the "deduplication result" because our objective is to find records in a dataset that belong to the same entity. The data consists of 2,343,686 unique tm_dedup_id with

[21] We use the Jaro-Winkler algorithm as it is known for managing typos and short strings. Furthermore, this algorithm has efficient computational speed, making it suitable for millions of comparisons.

an average of 1.02 names per tm_dedup_id and a maximum of 33 forms of names. On average, a tm_dedup_id holds 2.66 trademarks. The largest company, "Mattel Inc.," possesses 9,989 registered trademarks.

4.4 Combining with DISCERN

We further improve the accuracy of our matching process by combining our deduplication results with the DISCERN dataset, which also provides a foundation for linking trademark data to patent data for future research. We use an exact matching method to combine two datasets (trademark and DISCERN), and develop an identifier called "tm_id_name." To develop our "tm_id_name" identifier, we use standardized names within DISCERN as the key for connecting trademark data owners. If a company name in the DISCERN data matches the owner name in our dataset exactly, then we use "id_name"(an identifier in DISCERN) as the "tm_id_name." Additionally, if there are other company names that share the same "tm_dedup_id" with the matched name in the prior case, then we also use the matched "id_name" as "tm_id_name." If these cases do not apply, then we use "-tm_dedup_id" as the "tm_id_name." For example, "tesla motors inc" matches id_name 9249, so we assign the tm_id_name as 9249. The name "tesla inc" shares the same tm_dedup_id of 16142 as "tesla motors inc," so we also assign "tesla inc" to the tm_id_name as 9249. However, "givaudan corp" does not match any id_name, so we assign its tm_id_name as -51.

This step corrects incorrect matches made during the matching process in Section 4.3. In total, we successfully match 0.74 million records, which represents 6% of the 12 million owner records in this step. These matched companies are listed in Computstat, allowing for the integration of our trademark data and any dataset compatible with Computstat. The resulting data consists of 2,056,458 tm_id_name with an average of 1.02 names per tm_id_name and a maximum of 33 forms of names. On average, a tm_id_name holds 2.66 trademarks. The largest company in terms of trademark portfolio size is "Mattel Inc.," which possesses 9,989 registered trademarks.

4.5 Disambiguation outcome and firm-level measures

In Table 10, we present a sample company TurboChef Technologies (tm_id_name = 9519, GVKEY = 30008) that holds seven trademarks (No. 74157013, 74485101, 74485252, 74485274, 77399372, 78397378, and 78704703). Our disambiguation exercise contributes to the literature in two ways. First, our identifier tm_id_name enables researchers to obtain a more precise and comprehensive list of trademarks owned by each firm. Since trademarks contain rich information about an assignee's product lines, our disambiguated data help researchers track the development, evolution, and life cycles of firms' product lines, which facilitates research on product inventions, competition dynamics, industry landscapes, marketing strategies, and more.

Second, our trademark-level novelty and impact measures can be converted into firm-level measures, which will enhance researchers' tools used to capture firms' innovation performance from the perspec-

Table 10: Examples of disambiguated assignees

Serial No.	Owner Name	tm_id_name	id_name	Year	Novelty Counts	Novelty Ratio	Freshness	Survival Duration
74157013	turbochef tech llc	9519	NA	1991	0	0	0	19.64
74485101	turbochef inc	9519	9519	1994	1	1	41	6.78
74485252	turbochef inc	9519	9519	1994	1	1	41	6.76
74485274	turbochef inc	9519	9519	1994	1	1	41	10.77
77399372	turbochef tech llc	9519	NA	2008	1	0.14	28	10.04
78397378	turbochef tech llc	9519	NA	2004	0	0	0	9.97
78704703	turbochef tech llc	9519	NA	2005	0	0	0	10.79

Notes: The "serial no." column represents the trademark numbers. The "owner name" column displays the names of the trademark owners. The "tm_id_name" column shows the constructed firm identifier. The "id_name" column shows the firm identifier used in Arora et al. (2021a). The "year" column displays the year of filing for the trademark. The last four columns present the constructed novelty metrics: *novelty counts, novelty ratio, freshness*, and *survival duration*.

tive of commercialized products and services. For example, one can choose to measure a company's product novelty/impact using the average of its live trademarks' novelty. In our example (tm_id_name = 9519, GVKEY = 30008), the firm-level *novelty counts* is $(1 \times 4 + 0 \times 3)/7 = 0.57$, *novelty ratio* is $(1 \times 3 + 0.14 + 0 \times 3)/7 = 0.45$, and *freshness* is $(41 \times 3 + 28 + 0 \times 3)/7 = 21.57$.

4.6 Public firms

The unique identifier we construct not only allows researchers to obtain more precise and comprehensive knowledge of a company's trademark portfolio from DISCERN, but also enables researchers to link public firms to their trademark portfolios using the GVKEY through DISCERN. Furthermore, by using the information provided by DISCERN, we can track public firms' name changes, merges, and/or acquisitions. For example, in our trademark data, the trademark assignee "turbochef tech inc" has the id_name 9519 and the corresponding permanent number, permno_adj, 80482. Using the id_name 9519, we link the company to DISCERN and determine its GVKEY, which is 30008. According to DISCERN, "turbochef tech inc" was acquired by "middleby corp," resulting in a permno_adj change from 80482 to 75470, and a GVKEY change from 30008 to 13570.

We successfully linked 234,125 trademarks (out of all 3,646,085 trademarks) to 8,370 unique public firms (out of 1,294,881 trademark assignees). Table 11 reports the summary statistics of live trademarks, survival duration, and the number of new trademarks per year of public firms (Panel A) and non-public firms (Panel B).

For the 8,370 linked trademark assignees, the mean of the number of live trademarks is 7.63, and the standard deviation is 35.10. The minimum, $25th$ percentile, median, $75th$ percentile, and maximum are 0, 0, 0, 3, and 976, respectively. The mean of the survival duration is 9.90, and the standard deviation is 4.02. The minimum, $25th$ percentile, median, $75th$ percentile, and maximum are 0.07, 6.76, 8.77, 11.40, and 40.42, respectively. The mean of the number of new trademarks is 2.34, and the standard deviation is 4.08. The minimum, 25^{th} percentile, median, 75^{th} percentile, and maximum are, 1, 1, 1.64, 2.43, and 254.72, respectively.

Table 11: Summary statistics of the characteristics of firms

Panel A: Summary statistics for public firms						
Variable	Mean(sd)	Min.	1st Qu.	Median	3rd Qu.	Max.
live trademarks	7.63(35.10)	0	0	0	3	976
survival duration	9.90(4.02)	0.07	6.76	8.77	11.4	40.42
new trademarks	2.34(4.07)	1	1	1.64	2.43	254.72
Panel B: Summary statistics for non-public firms						
Variable	Mean(sd)	Min.	1st Qu.	Median	3rd Qu.	Max.
live trademarks	0.79(4.67)	0	0	0	1	1216
survival duration	8.42(4.03)	0.003	6.6	6.68	9.22	40.57
new trademarks	1.32 (0.89)	1	1	1	1.25	128

Notes: The first column, tm_id_name, represents the trademark ID. The "survival duration" column presents the lifespan of trademarks in years. The "new trademarks" column presents the count of new registered trademarks per year. For each variable, the sample average and standard deviation are presented as Mean (sd) in parentheses. Additionally, the minimum is presented as Min., the 25^{th} percentile as 1st Qu., the 50^{th} percentile as Median, the 75^{th} percentile as 3rd Qu., and the maximum as Max.

For the 1,286,511 non-linked trademark assignees, the mean of the number of live trademarks is 0.79, and the standard deviation is 4.67. The minimum, $25th$ percentile, median, $75th$ percentile, and maximum are 0, 0, 0, 1, and 1,216, respectively. The mean of the survival duration is 8.42, and the standard deviation is 4.03. The minimum, $25th$ percentile, median, $75th$ percentile, and maximum are 0.003, 6.6, 6.68, 9.22, and 40.57, respectively. The mean of the number of new trademarks is 1.32, and the standard deviation is 0.89. The minimum, 25^{th} percentile, median, 75^{th} percentile, and maximum are 1, 1, 1, 1.25, and 128, respectively. On average, public firms possess more live trademarks and register more new trademarks per year than non-public firms. The survival durations of trademarks owned by public firms are also longer.

5 Discussion

5.1 Summary

In this paper, we propose new measures based on trademark data for researchers to evaluate the novelty at the product level and firms' inventiveness in product markets. We use NLP to identify the first appearance and subsequent occurrences of invention keywords (and product inventions that they represent) in trademark documents. We develop several measures to evaluate the novelty of trademarks, and we find that our measures can explain the future survival of trademarks. These measures provide an *ex ante* perspective, reflecting the value of inventions in products and services.

In addition, we "harmonize" trademark assignees' names to construct a more precise and comprehensive trademark portfolio for each assignee. This allows for a possible link to GVKEY through DISCERN, which enables researchers to connect trademark data (and associated novelty measures) to public firms. We plan to make our codes and datasets publicly available to enhance the research community's future exploration of trademark documents.

5.2 The advantages of our measures

The data and measures we construct and propose enable researchers to capture the evolution of various inventions in products and services. Firms may intentionally not file patents for their inventions because of the business secret, patentability, market strategy, or legal considerations. However, firms will still launch new product lines and file trademarks to protect these inventions. Therefore, when studying innovations, trademarks can be more advantageous than patents and other intellectual property forms. These arguments are supported by the literature: trademarks are the most widely used form of IP protection (Hall et al., 2014),[22] and have been used by firms to protect their new products or services from imitation (Millot, 2009), to market their new products or services (Gao and Hitt, 2012; Flikkema et al., 2019), or to maintain their market power and customer loyalty (Block et al., 2015). Thus, several prior studies have used the number of (new) trademarks as a proxy for firm-level product innovations.[23] However, different from prior studies that use the simple count of trademarks, our measures not only track the evolution of each specific invention but also reflect the novelty and quality of each trademark.

It is also noteworthy that trademark data stand out from various other product and patent data in the following ways: 1.) Trademark data provide a broader scope than 10-K's product descriptions (Hoberg and Phillips, 2010), which are limited to public firms and influenced by their strategic disclosures. 2.) Trademark data cover a much wider range of categories than retail sales data like Nielsen's Retail Scanner (Argente et al., 2020; Aparicio et al., 2021), which focuses on consumer product sales and prices. 3.) Trademark data are free from marketing strategies and media preferences biases in new product announcements covered in media (Mukherjee et al., 2017). 4.) Trademark data mainly reflects new products or services instead of technologies because products competing for the same market may be based on different technologies.

5.3 Possible extensions

Our research itself can be improved or extended in several ways. First, one could use keyword combinations to extract more comprehensive information about product inventions. Second, one could use new ways to extract the information in trademark documents, either to improve our novelty measures or to develop new measures for different purposes. Finally, one could explore alternative matching algorithms and/or utilize external information sources (e.g., the Bayes classifier, PermID, M&A data) to enhance matching accuracy.

This research note not only demonstrates how trademark documents can be used to extract information about product inventions, but also highlights the untapped potential of trademark data for subsequent investigations. We list some possible research directions that call for future research. First, one could

[22] According to the U.S. National Science Foundation's new Business R&D and Innovation Survey (BRDIS) survey, among all firms in the survey, 15% firms answer trademarks being important in the protection of their IP (utility patents 5%, design patents 6%, copyrights 12%, and trade secret 14%).

[23] See Mendonça et al. (2004); Hipp and Grupp (2005); Greenhalgh and Rogers (2006); Sandner and Block (2011); Flikkema et al. (2014); Schautschick and Greenhalgh (2016); Faurel et al. (2017); Crass et al. (2019); Crown et al. (2020); Nasirov (2020); Hsu et al. (2022a). Hsu et al. (2022b), on the other hand, use new trademarks to measure firms' market entry.

associate our firm-level novelty with assignee characteristics by linking our tm_id_name to other data sources. For example, one could link our tm_id_name to CRSP/Compustat to collect assignee firms' financial and accounting information; by doing so, researchers could determine which firm characteristics better explain the outcome of product inventions. Second, using appropriate identification strategies for particular firm characteristics or external factors, one could revisit the determinants of firms' product inventions in the literature or even propose new determinants. Third, one could connect patent information to trademarks through products and examine the interesting relation between technology innovation and product inventions. Fourth, and more importantly, one could investigate whether particular patent features can better explain the novelty and impact of product inventions.

References

Aparicio, Diego, Zachary Metzman, and Roberto Rigobon (2021). *The Pricing Strategies of Online Grocery Retailers*. Tech. rep. 28639. National Bureau of Economic Research.

Argente, David, Salome Baslandze, Douglas Hanley, and Sara Moreira (2020). "Patents to Products: Product Innovation and Firm Dynamics (April, 2020)". In: *SSRN*.

Arora, Ashish, Sharon Belenzon, and Lia Sheer (2021a). "Knowledge Spillovers and Corporate Investment in Scientific Research". In: *American Economic Review* 111(3), pp. 871–98.

Arora, Ashish, Sharon Belenzon, and Lia Sheer (2021b). "Matching patents to compustat firms, 1980–2015: Dynamic reassignment, name changes, and ownership structures". In: *Research Policy* 50(5), p. 104217.

Arts, Sam, Jianan Hou, and Juan Carlos Gomez (2021). "Natural language processing to identify the creation and impact of new technologies in patent text: Code, data, and new measures". In: *Research Policy* 50(2).

Arts, Sam, Nicola Melluso, and Reinhilde Veugelers (2023). *Beyond Citations: Text-Based Metrics for Assessing Novelty and its Impact in Scientific Publications*.

Block, Jörn H., Christian O. Fisch, Alexander Hahn, and Philipp G. Sandner (2015). "Why do SMEs file trademarks? Insights from firms in innovative industries". In: *Research Policy* 44(10), pp. 1915–1930.

Crass, Dirk, Dirk Czarnitzki, and Andrew A. Toole (2019). "The Dynamic Relationship Between Investments in Brand Equity and Firm Profitability: Evidence Using Trademark Registrations". In: *International Journal of the Economics of Business* 26(1), pp. 157–176.

Crown, Daniel, Alessandra Faggian, and Jonathan Corcoran (2020). "Foreign-Born graduates and innovation: Evidence from an Australian skilled visa program". In: *Research Policy* 49(9), p. 103945.

Dinlersoz, Emin M, Nathan Goldschlag, Amanda Fila, and Nikolas Zolas (2021). "An Anatomy of U.S. Firms Seeking Trademark Registration". In: *NBER Working Paper*(25038).

Faurel, Lucile, Qin Li, Devin M Shanthikumar, and Siew Hong Teoh (2017). "CEO incentives and new product development: Insights from trademarks". In: *Available at SSRN*.

Flikkema, Meindert, Carolina Castaldi, Ard-Pieter de Man, and Marcel Seip (2019). "Trademarks' relatedness to product and service innovation: A branding strategy approach". In: *Research Policy* 48(6), pp. 1340–1353.

Flikkema, Meindert, Ard-Pieter De Man, and Carolina Castaldi (2014). "Are Trademark Counts a Valid Indicator of Innovation? Results of an In-Depth Study of New Benelux Trademarks Filed by SMEs". In: *Industry and Innovation* 21(4), pp. 310–331.

Gao, Guodong (Gordon) and Lorin M. Hitt (2012). "Information Technology and Trademarks: Implications for Product Variety". In: *Management Science* 58(6), pp. 1211–1226.

Graham, Stuart J.H., Galen Hancock, Alan C. Marco, and Amanda Fila Myers (2013). "The USPTO Trademark Case Files Dataset: Descriptions, Lessons, and Insights". In: *Journal of Economics & Management Strategy* 22(4), pp. 669–705.

Graham, Stuart J.H., Alan C. Marco, and Amanda Fila Myers (2018). "Monetizing marks: Insights from the USPTO Trademark Assignment Dataset". In: *Journal of Economics & Management Strategy* 27(3), pp. 403–432.

Greenhalgh, Christine and Mark Rogers (2006). "The value of innovation: The interaction of competition, R&D and IP". In: *Research Policy* 35(4), pp. 562–580.

Hall, Bronwyn, Christian Helmers, Mark Rogers, and Vania Sena (2014). "The Choice between Formal and Informal Intellectual Property: A Review". In: *Journal of Economic Literature* 52(2), pp. 375–423.

Hipp, Christiane and Hariolf Grupp (2005). "Innovation in the service sector: The demand for service-specific innovation measurement concepts and typologies". In: *Research Policy* 34(4), pp. 517–535.

Hoberg, Gerard and Gordon Phillips (2010). "Product Market Synergies and Competition in Mergers and Acquisitions: A Text-Based Analysis". In: *The Review of Financial Studies* 23(10), pp. 3773–3811.

Hsu, Po-Hsuan, Dongmei Li, Qin Li, Siew Hong Teoh, and Kevin Tseng (2022a). "Valuation of New Trademarks". In: *Management Science* 68(1), pp. 257–279.

Hsu, Po-Hsuan, Kai Li, Xing Liu, and Hong Wu (2022b). "Consolidating product lines via mergers and acquisitions: Evidence from the USPTO trademark data". In: *Journal of Financial and Quantitative Analysis* 57(8), pp. 2968–2992.

Kelly, Bryan, Dimitris Papanikolaou, Amit Seru, and Matt Taddy (2021). "Measuring Technological Innovation over the Long Run". In: *American Economic Review: Insights* 3(3), pp. 303–20.

Li, Guan-Cheng, Ronald Lai, Alexander D'Amour, David M. Doolin, Ye Sun, Vetle I. Torvik, Amy Z. Yu, and Lee Fleming (2014). "Disambiguation and co-authorship networks of the U.S. patent inventor database (1975–2010)". In: *Research Policy* 43 (6), pp. 941–955.

Marchant, Neil G., Andee Kaplan, Daniel N. Elazar, Benjamin I. P. Rubinstein, and Rebecca C. Steorts (2021). "d-blink: Distributed End-to-End Bayesian Entity Resolution". In: *Journal of Computational and Graphical Statistics* 30(2), pp. 406–421.

Mendonça, Sandro, Tiago Santos Pereira, and Manuel Mira Godinho (2004). "Trademarks as an indicator of innovation and industrial change". In: *Research Policy* 33(9), pp. 1385–1404.

Mermin, Jonathan (2000). "Interpreting the Federal Trademark Dilution Act of 1995: The Logic of the Actual Dilution Requirement Note". In: *Boston College Law Review* 42, pp. 207–238.

Millot, Valentine (2009). "Trademarks as an Indicator of Product and Marketing Innovations". In: *OECD Science, Technology and Industry Working Papers* 2009/06.

Missirian, David E. (2019). "The Death of Moral Freedom: How the Trademark Dilution Act Has Allowed Federal Courts to Punish Subjectively-Defined Immoral Secondary Use of Trademarks". In: *Chicago-Kent Journal of Intellectual Property* 18, pp. 396–428.

Morrin, Maureen, Jonathan Lee, and G.M. Allenby (2006). "Determinants of Trademark Dilution". In: *Journal of Consumer Research* 33(2), pp. 248–257.

Mukherjee, Abhiroop, Manpreet Singh, and Alminas Žaldokas (2017). "Do corporate taxes hinder innovation?" In: *Journal of Financial Economics* 124(1), pp. 195–221.

Nasirov, Shukhrat (2020). "Trademark value indicators: Evidence from the trademark protection life-cycle in the U.S. pharmaceutical industry". In: *Research Policy* 49, p. 103929.

Roe, Jeremy M. (2008). "The Current State of Antidilution Law: The Trademark Dilution Revision Act and the Identical Mark Presumption Symposium: Challenges to the Attorney-Client Relationship: Thirteenth Annual Clifford Symposium on Tort Law and Social Policy: Note". In: *DePaul Law Review* 57, pp. 571–606.

Sandner, Philipp G. and Joern Block (2011). "The market value of R&D, patents, and trademarks". In: *Research Policy* 40(7), pp. 969–985.

Schautschick, Philipp and Christine Greenhalgh (2016). "Empirical studies of trade marks – The existing economic literature". In: *Economics of Innovation and New Technology* 25(4), pp. 358–390.

Simon, Ilanah (2006). "The Actual Dilution Requirement in the United States, United Kingdom and European Union: A Comparative Analysis". In: *Boston University Journal of Science & Technology Law* 12, pp. 271–309.

Steorts, Rebecca C. (2015). "Entity Resolution with Empirically Motivated Priors". In: *Bayesian Analysis* 10(4), pp. 849–875.

Steorts, Rebecca C. (2016). "A Bayesian Approach to Graphical Record Linkage and Deduplication". In: *Journal of the American Statistical Association* 111(516), pp. 1660–1672.

Internet Appendices

A Trademarks overview

A.1 Trademark introduction

A trademark consists of a symbol (e.g. word, phrase, or symbol) and several identifications that describe products and services covered by the trademark and distinguish one from other trademarks. Several requirements must be met when registering a trademark to the USPTO. First, the applicant must specify one or more trademark classes in 45 international classes (associated with one or more U.S. classifications). The exclusive use of the trademark is limited in the registered classes in principle.[24] Second, the applicant must send a specimen to the USPTO to prove that the trademark is currently being used or is going to use in commerce in *all* the specified classes. To maintain the ownership of a registered trademark, if they have to pay maintenance fees and to prove their current use of the trademarks in the sixth year from registration dates and to pay renewal fees and to prove their use of the trademarks every 10 years from registration dates. We provide more details about the laws and procedures for trademark applications, registrations, and renewal in the Appendix A.2.

[24]The pattern "Trademark dilution" denotes a situation in which a trademark or its modified form is used by other entities in product or service areas that are unrelated to those covered by the original trademark (Mermin, 2000; Morrin et al., 2006). The Federal Trademark Dilution Act (FTDA) was enacted in 1996 to prevent such trademark dilution. On particular change made by the FTDA is to remove the requirement of proven actual loss for trademark owners to claim damage. Nevertheless, the implementation of the FTDA is subject to different courts' interpretations (Mermin, 2000; Roe, 2008). For example, Simon (2006) reports that the Fourth Circuit still required the actual dilution evidence, and Missirian (2019) argues that the FTDA may be unnecessary to protect mark holders from actual damage.

A.2 Trademark basics

The first legal system of trademarks was created in France in 1857, with the "Legislation Relating to Commercial Marks and Product Marks" that justifies the laws and enforcements of trademarks and infringements (Millot, 2009). In Britain, the trademark system was established in 1862, with the "Merchandise Marks Act" that made it a criminal offense to imitate another's trademarks. In the U.S., the trademark system was first attempted to establish a federal trademark regime in 1870. The Agreement on Trade-Related Aspects of Intellectual Property Rights in 1994 is the latest attempt to standardize the trademark procedures across countries. Overall, the procedure and protection of trademarks are largely similar in most developed countries (Millot, 2009).

The modern U.S. federal trademark registration system was established with the Lanham Act in 1946[25]. The USPTO defines a trademark as "any word, name, symbol, device, or any combination, used or intended to be used to identify and distinguish the goods/services of one seller or provider from those of others, and to indicate the source of the goods/services."[26] Article 15 of the Agreement on Trade-Related Aspects of Intellectual Property Rights defines trademarks as "any sign, or any combination of signs, capable of distinguishing the goods or services of one undertaking from those of other undertakings, shall be capable of constituting a trademark."[27]

A firm may file a trademark application to the USPTO for a new trademark in some particular product/service classes[28]. The applicant needs to provide proof of the actual use of the trademark in commerce, such as a specimen, or file an Intent-to-Use statement to agree to provide proof within the next six months (Graham et al., 2013).[29] After an application serial number is assigned, the application is forwarded to an examining USPTO attorney for review, which includes a search for conflicting marks and an examination of the written application, and the submitted drawing and specimen. The attorney's job is to ensure that the trademark is novel and reasonably distinct from existing trademarks, and can be easily identifiable by the public. The attorney may reject the application if the proposed trademark has been commonly used by the public (e.g., "Police"), only descriptive of the product or of its quality (e.g., "Cheese" and "Delicious"), has no distinctive characters, has a scandalous connotation, or else refers to specific official emblems (e.g., "California") (see, e.g., Millot, 2009; Graham et al., 2013).[30]

If the examining attorney in the USPTO raises no correction requests or objections, or if the applicant

[25] Although the Act has been amended several times since, it remains the primary federal trademark statute in providing nationwide regulation and protection for trademark registration (Graham et al., 2013).

[26] See https://www.uspto.gov/trademarks-getting-started/trademark-basics

[27] See http://www.wipo.int/wipolex/en/other_treaties/details.jsp?group_id=22&treaty_id=231

[28] There are 45 product/service classes: http://www.wipo.int/classifications/nice/nclpub/en/fr/home.xhtml. A trademark can be filed in one or multiple classes. 86.5% of trademark applications are registered in single classes (Graham et al., 2013).

[29] It is noteworthy that 45.9% of intent-to-use applications are abandoned without being registered.

[30] 8.3% of trademark applications were rejected by examining attorneys (Graham et al., 2013). If the applicant decides that minor corrections are required, he/she will issue a letter (Office Action) to request corrections. If the attorney decides that the proposed trademark should not be registered, he/she will issue a letter (Office Action) explaining any substantive reasons for refusal, and any technical or procedural deficiencies in the application. The applicant needs to respond to the Office Action within six (6) months of the mailing date of the Office action, or the application will be declared abandoned.

has addressed all concerns and overcome all objections raised by the attorney, the examining attorney will approve the trademark to be published in the Official Gazette, a weekly Tuesday publication by the USPTO. After the mark is published in the Official Gazette, a third party may file a notice of opposition to the trademark's registration during this 30-day period after publication.[31] If no opposition is filed or if the opposition is unsuccessful, the application enters the next stage of the registration process.

Before the official registration of the trademark, the applicant will need to file statement of use to prove the actual use of the trademark in commerce if such a proof has not been provided in initial application. After all these necessary conditions are met, the trademark can be officially registered.[32] After a trademark is registered, the firm can use the ® symbol with their trademark and can now enjoy legal trademark protection.

After a firm successfully registers a trademark, it can claim for incontestability by filing the Declaration of Incontestability in the completion of the fifth year from the registration date. Such a claim shields the firm from challenges based on descriptiveness such as (1) the trademark merely describes the goods or services, (2) the mark is descriptive because it is primarily merely a surname, and (3) the mark is descriptive because it is a geographic place name. Firms have strong incentives to file the incontestability claim so that they can use incontestability as a defense against an action for trademark infringement in federal courts.

Firms can hold permanent ownership of their trademarks if they can maintain the trademarks in the sixth year from registration dates and to renew the trademarks every 10 years from registration dates.[33] Failure to file the required maintenance and renewal documents in the specified time periods will result in the cancellation of the trademark or invalidation of legal protection. Between the fifth and sixth year after registration, the owner must file the Declaration of Use of Mark in Commerce to show the continued use of the trademark and pay fees to maintain the registration.[34] In particular, the owner needs to present a specimen that is currently used for each class of goods or services in which the trademark has been registered for.[35] Further, on the date between the ninth and tenth years after the registration (and each successive ten-year period thereafter), the owner needs to renew the trademark registration by filing the Application for Renewal of Registration of a Mark, together

[31] When a notice of opposition is filed, the owner of the opposed application has 30 days to file an answer with the Trademark Trial and Appeal Board (TTAB), which is a body within the USPTO responsible for hearing and deciding certain kinds of trademark-related cases. 98.1% of published applications were registered (Graham et al., 2013).

[32] As shown in Graham et al. (2013), 78.8% of all applications were eventually granted. The median time from application to registration is 1.2 years for all registrations filed with actual use and is 1.9 years for all those filed based on intent-to-use.

[33] The relevant procedures for maintaining and renewing trademarks can be found on the USPTO website:https://www.uspto.gov/trademarks-maintaining-trademark-registration/keeping-your-registration-alive and https://www.uspto.gov/trademarks-application-process/filing-online/registration-maintenancerenewalcorrection-forms. The renewal frequency was 20 years before November 1989 and reduced to 10 years after the enactment of Trademark Law Revision Act of 1988 [Title 1 of Pub. L. 100-667, 102 Stat. 3935 (15 U.S.C. 1051)]. Registrations can be renewed within one year before the end of every 10-year period after the registration date or within the 6-month grace period thereafter.

[34] The owner can still file extension for six months after the sixth year from registration.

[35] Other materials such as the promotion documents or advertisements that demonstrate that the trademark is in use are also acceptable. According to Graham et al. (2013), 47.1% of trademarks registered were maintained after the sixth year.

with the Declaration of Use, by proving the continued use of the trademark and pay fees.[36]

B Overview of data collection

We start with two datasets provided by the USPTO: the USPTO case file dataset and the USPTO trademark assignment dataset. The USPTO case file dataset contains 15 data files, our process involves using the following four: **case_file**, **statement**, **prior_mark**, and **owner**. The USPTO trademark assignment dataset contains 7 data files, we use the following two: **tm_assignor** and **tm_assignee**. To obtain our processed data, we first merge, link, and clean the raw data.

B.1 The USPTO case file dataset

Our analysis focuses on registered trademarks filed after 1980, as noted by Graham et al. (2013) due to possible incomplete coverage of trademarks in the USPTO prior to this period. However, we include trademarks filed prior to 1980 in establishing the baseline words and disambiguation. The USPTO case file dataset comprises 11.6 million trademarks where 6.6 million of them are registered trademarks between 1870 and 2020. We use the 2021 release version.

We use the **case_file** data file, which records basic information for trademark applications or registrations (see Internet Appendix C.1 for a screenshot and more details). The **case_file** data includes information such as the current status, location, filing date, registration date, renewal date, cancellation date, and cfh status date of each trademark. We use it to construct a current status indicator to determine if the trademark is live or not.[37] Our indicator reveals that 32% of trademarks are live registrations as of 2020. This indicator is crucial when estimating the survival function of trademarks, which we calculate based on the registration date, cancellation date, and case file header (cfh) status date.[38]

The **statement** data file contains information about the goods and services that are associated with trademark applications or registrations (see Internet Appendix C.2 for a screenshot and more details). We select observations with a statement type code starting with "GS,"[39] which indicates that they are good and service identifications. Each observation represents an identification of an international class, which the applicant specifies. The applicant must provide evidence to justify that their products and services are related to each registered class. There are 45 international classes, 34 of them are product classes and 11 of them are service classes. The exclusive right using the logo or word is only valid in the applied classes.

We gather information on the names and locations (street address, postal code, city, and state) of the

[36] The owner can still file extension for six months after each successive ten-year period after registration. Among patents that were maintained in the sixth year, 68.9% were renewed in the tenth year (Graham et al., 2013).

[37] To identify if the trademark is live or not at the time of data construction.

[38] See Section 3.2 for the calculation of survival times. The cfh status date records the date of the last recorded status event. Each status event is recorded as a three-digit code from 121 unique values, indicating whether an application was abandoned or pending, or registration was live, canceled, or expired.

[39] This data file records various text statements within trademark applications or registrations. These statements include goods and services, descriptions of the claimed colors, and descriptions of the trademark. Each type of statement is recorded with a distinct statement type code.

assignee recorded for each trademark application or registration in the **owner** data file (see Internet Appendix C.3 for a screenshot and more details). This data file comprises 9.8 million unique trademark serial numbers and holds a total of 23.6 million records. Out of these, 2.6 million serial numbers have only one record, while the remaining 7.2 million serial numbers have multiple records.[40]

B.2 The USPTO trademark assignment dataset

With USPTO assignment records, we can track the full history of interests in a trademark, i.e., ownership, and obtain additional information about the forms of company names. The **tm_convey** data file classifies each assignment into several conveyance types. Conveyance refers to a category that indicates the method of the transaction of interest. For example, the assignment ID 73470366, "koninklijke philips elec n v" changed its name and address to "koninklijke philips n v." To construct assignor-assignee pairs with conveyance types, we merge the **tm_convey**, **tm_assignor**, and **tm_assignee** data. We use the 2021 release version of trademark assignment dataset. Section 4 details the use of assignor-assignee pairs.

The **tm_convey** includes assignments, name changes, mergers, corrections, security interests, releases, and others (see Internet Appendix D.1 for a screenshot and more details). The **tm_convey** performs the classification based on descriptions such as "entire interest," "security interest," and "merger" found in the trademark assignment cover sheet. For more information, refer to Graham et al. (2018). There are 1,199,497 observations in this data file.

The **tm_assignor** file records data for the assignor of each *rf_id*. This data includes the assignor's name, location, type of legal entity, date of execution, and other relevant information. There are 1,322,998 observations in this data file. The **tm_assignee** file similarly records data for the assignee of each *rf_id*, including the assignee's name, location, legal entity type, and other relevant information. There are 1,261,730 observations in this data file. The screenshots and details of these two files are provided in Internet Appendix D.2.

We merged **tm_convey**, **tm_assignor**, and **tm_assignee** using *rf_id*, resulting in a constructed dataset with 1.4 million observations. These observations were classified into eight groups: the assignment group (45.9%), name change group (18.9%), merger group (6.2%), correction group (2.6%), security interest group (16.5%), release group (8.1%), other group (1.6%), and no conveyance recorded (0.3%). We use records from the name change and correction groups to construct potential matching pairs, covering 300,433 (21%) records in the assignment data.

B.3 The DISCERN dataset

We combine company names in the trademark dataset to the company names provided by Arora et al. (2021a). This dataset connects patent assignees with Compustat firms, and its matching results

[40] Multiple recordings of trademark assignees may occur in a trademark due to different owner types or during different stages of the application process.

in dynamic matching, company name changes, and ownership structures are superior to other similar datasets. We combine this dataset with the trademark dataset to improve our disambiguation procedure. We use their name list, which keeps track of both the company names and their corresponding company IDs, referred to as "id_name," within the data files **DISCERN_UO_name_list.dta** and **DISCERN_SUB_name_list.dta**. The **DISCERN_UO_name_list.dta**, contains 10,348 firm names, while the **DISCERN_SUB_name_list.dta**, contains 50,570 firm names. Appendix E provides examples of these two data files and their screenshots.

C The USPTO case file dataset

This section introduces the "USPTO trademark case files dataset" and provides descriptions, screenshots, and examples for the following data files: **case_file, statement**, and **owner**.

C.1 Case file

The case_file data file contains information on trademark registrations and applications. The 2021 release includes 11,560,910 observations of 79 variables. We primarily use the serial number, the current status code, current status dates, filing dates, registration dates, and cancellation dates.

Figure C1 is the screenshot for the case_file data file. The filing date records the submission date of the trademark application to the USPTO. The registration date specifies when the trademark application has successfully passed the examination and obtained registration. Additionally, the cancellation date records when the registration has been canceled.

The case file header (cfh) status date records the date of the last recorded cfh status event, indicating the time of status event happened. Each status event is recorded using a three-digit code from 121 unique values, providing information on whether an application has been abandoned or is pending, or whether the registration is live, canceled, or expired. For example, 600 to 609 represent dead and abandoned applications, 700 represent registered trademarks, and 710 represent canceled trademarks. The definition for all cfh status codes can be found in TAD Table 1.

We use the filing date of the trademark in which an invention keyword first appeared as the creation time of the keyword. And we use the registration date, the cancellation date, and the status date to determine the survival time of the trademark.

C.2 Statement

The statement data file records various text statements in the trademark application or registration. It includes information about goods and services, descriptions of claimed colors, and descriptions of the trademark. We collect identifications of goods and services from this data file and use serial numbers to link with other data files.

Figure C2 is the screenshot of the statement data file. The column "statement type code" indicates the type of statement. The column "statement text" records the content of each type of statement.

serial_no	reg_cancel_dt	filing_dt	mark_id_char	registration_dt	cfh_status_cd	cfh_status_dt	registration_no
87797083	NA	NA		NA	622	2018-02-17	0
73741005	1995-09-11	1988-07-20	TRIBUNE	1989-03-07	710	1995-09-11	1528005
87044217	NA	2016-05-20	SKINNY BEVERAGES	NA	602	2017-04-25	0
90556792	NA	2021-03-03	ASKEEP	2022-01-18	700	2022-01-18	6620367
97072973	NA	2021-10-13	ECORITE	NA	630	2021-10-16	0
87193302	NA	2016-10-05	SPICY TUNA	2017-07-18	700	2017-07-18	5247840
77381069	NA	2008-01-25	OFF THE EATEN PATH	NA	601	2008-02-14	0
75672234	NA	1999-03-31	TRIGUETTE	NA	606	2001-01-19	0
87930732	NA	2018-05-22	SIDEFACE	2019-02-26	700	2019-02-26	5684192
73675692	1995-08-07	1987-07-31		1989-01-31	710	1995-08-07	1522432
88761350	NA	2020-01-16	ROAD REPORT	2021-03-23	700	2021-03-23	6298051
88370091	NA	2019-04-03	UGLY CHRISTMAS PARTY	2019-12-17	700	2019-12-17	5935896
75839930	2021-11-12	1999-11-05	POT RACK WORLD	2001-01-23	710	2021-11-12	2423139
74545690	NA	1994-07-05	PARTY TIME	NA	602	1995-11-15	0
87854199	NA	2018-03-28	STRAIGHTRATE	NA	602	2019-03-14	0
60319952	NA	NA		1934-12-11	626	2005-10-27	319952

Figure C1: Screenshot for the case_file data file. This screenshot shows the crucial variables in this paper: current status, current status dates, filing dates, registration dates, and cancellation dates. While the complete case_file data file contains a total of 79 variables.

The column "serial number" shows the serial numbers.

For example, the first row in Figure C2 states that trademark No. 85688078 has a statement type code GS0281 and a statement text "Bags specially adapted for archery and bow-hunting equipment." In the statement type code GS0281, "GS" represents the identification of goods and services, 028 represents the international class (IC) 028, and 1 represents that the text is a goods and services statement with no "less goods" text. In this paper, we only consider trademarks of goods and services.

C.3 Owner

The owner data file contains information such as the owner's name, type of legal entity, address, and nationality. This file records various stages of ownership, including the names of the original applicant, all owners before and after the publication, the original registrant, and all owners after the registration. We collect the names and locations (street address, postal code, city, and state) of trademark owners recorded in trademark applications or registrations to disambiguate the names of trademark owners.

Figure C3 is the screenshot of the owner data file. The first two columns "own_addr_1" and "own_addr_2," contain the first and the second line of the owner's address. The column "own_addr_city" records the city. The column "owner_name" shows the owner's name, and the column "owner_addr_postal" shows the postal code.

D The USPTO trademark assignment dataset

This section includes the description, screenshots, and examples of the USPTO trademark assignment dataset. The USPTO records assignments of trademark applications or registrations, allowing people to see the full history of claimed interest of a trademark. These assignments include the transaction of mergers, name changes, security interest agreements, corrections, and licenses. We use name changes

statement_type_cd	statement_text	serial_no
GS0281	Bags specially adapted for archery and bow-hunting equip...	85688078
GS0281	Plush stuffed toys	87136266
GS0411	Providing on-line publications in the form of articles and ac...	87799060
GS0341	[Ashtrays]	73393184
GS0251	clothing, namely, coats, jackets, gloves, hats, boxers, swimw...	86692885
GS0091	Battery chargers; Bicycle safety lights; Body cameras for use ...	88553139
PM0000	LOOK NY; LOOK NEW YORK	77982957
GS0111	Downlights; Lightbulbs; Luminaries; Roadlights; Desk lamps; ...	90412925
D10000	MONTE CARLO	76537251
GS0051	NUTRITIONAL SUPPLEMENTS	76513287
GS0381	Peer-to-peer photo sharing and video sharing services, nam...	87646951
TR0010	The English translation of ""blu"" in the mark is ""blue"".	85359228
GS0061	Metal garage doors	78148725
D10000	""ARTESAN?A MURO""	88177182
GS0271	Carpets; leather floor coverings; leather wall coverings; cloth...	78515119
DM0000	The mark consists of the letters ""NBS"" separated by variou...	77469553

Figure C2: Statement. The column "statement type code" indicates the type of statement. The column "statement text" records the content of each type of statement. The column "serial number" shows the serial numbers.

own_addr_1	own_addr_2	own_addr_city	own_name	own_addr_postal	serial_no
Suite 195	44190 Mercure Circle	Dulles	Integral Transport Service, Inc.	20166	77001972
9 Langdon Road		Richmond	Vivian Pratte	04357	87711784
	64 VETERANS DRIVE	NEW BRITAIN	SONIC GOLF COMPANY	06050	73598963
	ONE BUSCH PLACE	ST. LOUIS	ANHEUSER-BUSCH, INCORPORATED	63118	73753944
7683 SOUTHFRONT ROAD		LIVERMORE	ACTIVANT SOLUTIONS INC.	94551	76035741
200 Putter Drive		Brentwood	Ruiz, Juan, Jose	94513	87332459
1318 High Street		Williamsport	JM Quizzo Productions, LLC	17701	86204989
2021 Spring Road, Suite 300		Oak Brook	GMAC Home Services, LLC	60523	77092217
/o Corporate Service Bureau Inc.	28 Old Rudnick Lane	Dover	CONOVER 251-253 ASSOCIATES, LLC	19901	87385166
305,Block B,Zhantao Commercial Plaza,	Tenglong Rd., Longsheng Community,Dalang	St.,Longhua,Shenzhen	Shenzhen LiangZiLiXing Industry Co.,Ltd.	518000	88501411
8610 Evergreen Place		Philadelphia	G. W. Jr. Music, Inc.	19118	75903421
1007 Green Acres Mall		Valley Stream	Omri Inc	11581	77699866
1875 Central Park Avenue		Steamboat Springs	Steamboat Software, Inc.	80488	74530356
399 INTERPACE PARKWAY		PARSIPPANY	RB HEALTH (US) LLC	07054	85634998
14141 Di Giorgio Road		Di Giorgio	Grimmway Enterprises, Inc.	93217	74470381
1350 S.E. 17th Street		Ft. Lauderdale	4Y Yacht Sales, Inc.	33316	78574821
10, Haatsmaut		Raanana	International Trauma-Healing Institute	43460	87069716

Figure C3: Owner. The first two columns, "own_addr_1" and "own_addr_2," contain the first and the second line of the owner's address. The column "own_addr_city" records the city. The column "owner_name" shows the owner's name, and the column "owner_addr_postal" shows the postal code.

rf_id	conv_group
21370344	assignment
56470845	assignment
13980070	assignment
35810890	assignment
09250910	correction
58780809	merger
65310363	assignment
02830220	assignment
67250569	release
55990456	name change
23280001	assignment
41940758	assignment
48780922	name change
52280625	assignment
36060389	assignment

Figure D1: Convey groups. The column "rf_id" is the key used to link the tm_convey data file to other data files in assignment data files including tm_assignor and tm_assignee. The column "conv_group" indicates the conveyance types of assignment.

and correction assignments to disambiguate the names of owners.

D.1 Convey

The tm_convey data file contains the conveyance of assignments. Conveyance indicates the category of the transaction of interest. This includes assignments, name changes, mergers, corrections, security interests, releases, and others. The classification of assignments is based on descriptions in the trademark assignment cover sheet, such as "entire interest," "security interest," and "merger." For more information, see Graham et al. (2018).

Figure D1 is the screenshot of the tm_convey data file. The column named "rf_id" is the key used to link the tm_convey data file to other data files in assignment data files including tm_assignor and tm_assignee. The column named "conv_group" indicates the conveyance types of assignment. The conveyance types of assignments could be "assignment," "name change," "security interest," "merger," "release," "correction," "other," and no conveyance recorded.

D.2 Assignor and assignee

The tm_assignor file records data for the assignor of each rf_id. This data includes the assignor's name, location, type of legal entity, date of execution, and other relevant information. The tm_assignee file similarly records data for the assignee of each rf_id, including the assignee's name, location, legal entity type, and other relevant information.

Figure D2 is the screenshot for tm_assignor. It includes the assignor's name, address, city, postal code,

rf_id	or_name	or_address_1	or_address_2	or_city	or_state	or_postcode	exec_dt
63590460	CUMULUS BROADCASTING LLC						2018-06-01
28760496	BLEACH, BARRY M.						2004-06-03
18690831	CARL KARCHER ENTERPRISES, INC.						1999-03-04
62560564	JS&M SALES AND MARKETING, INC.						2018-01-24
54850973	PREMIER PAINT ROLLER MFG. CO., INC.						2013-06-06
01690365	PADGETT, MARVIN J.	P.O. BOX 66	SHIVERLY BRANCH	LOUISVILLE	KENTUCKY		NA
03810831	STEPAN CHEMICAL COMPANY			NORTHFIELD	ILLINOIS	60093	1980-09-15
42220553	PAPER SOURCE, INC.						2010-06-04
06870684	WAYLITE COMPANY, THE			CHICAGO	ILLINOIS		1990-01-03
05730328	PALCO INDUSTRIES, INC.		10880 WILSHIRE BOULEVARD	LOS ANGELES	CALIFORNIA	90024	1987-07-16
62140552	SINTON, THOMAS						2017-11-17
74220802	KHS INCORPORATED						2021-07-22
34330778	E2V TECHNOLOGIES (UK) LIMITED						2004-06-09
02440356	AMSTEL BROUWERIJ N.V.						1973-10-22
13020202	THOSE LITTLE DONUTS INTERNATIONAL, INC.						1994-12-06

Figure D2: Assignor. This data file contains the assignor's name, address, city, postal code, and execution date of the event.

rf_id	ee_name	ee_address_1	ee_address_2	ee_city	ee_state	ee_postcode
31250470	STA-RITE INDUSTRIES, LLC	293 WRIGHT STREET		DELAVAN	WISCONSIN	53115
56540836	MUUTO A/S	?STERGADE 36, 4.		DK-1100 COPENHAGEN K		
53110350	XPEDX, LLC.	6400 POPLAR AVENUE		MEMPHIS	TENNESSEE	38197
28610857	REFOUAH BLUE, INC.	350 LONGWOOD CROSSING		LAWRENCE	NEW YORK	11559
25870741	EPIPHANY SKINCARE, INC.	8501 GUNNER WAY		FAIR OAKS	CALIFORNIA	95628
74330610	WESTERN ALLIANCE BANK	55 ALMADEN BLVD. STE. 100		SAN JOSE	CALIFORNIA	95113
57410358	SIVANTOS GMBH	HENRI-DUNANT-STR. 100		91058 ERLANGEN		
51750766	SELECTIVA S.P.A.	STRADA STATALE PER GENOVA, KM. 98		I-15122 ALESSANDRIA		
09080399	CANDLE CORPORATION OF AMERICA		141 WEST 62ND STREET	CHICAGO	ILLINOIS	60621
27980301	BANK OF AMERICA, N.A., AS ADMINISTRATIVE AGENT	231 S. LASALLE STREET	IL1-231-08-30	CHICAGO	ILLINOIS	60690
74760542	1105 MEDIA INC.	6300 CANOGA AVENUE	ADDRESS: SUITE 1150	WOODLAND HILLS	CALIFORNIA	91367
64690123	RIBA WATCH (SUISSE) S?RL	RUE DES BILLODES 55		2400 LE LOCLE		
37370606	WILL & BAUMER CANDLE COMPANY, L.L.C.	5226 S 31ST PL		PHOENIX	ARIZONA	85040
17150232	CREDIT SUISSE FIRST BOSTON		11 MADISON AVENUE	NEW YORK	NEW YORK	10010
53710858	ZOOM FRANCHISE COMPANY, LLC	915 S. TROOPER ROAD		AUDUBON	PENNSYLVANIA	19403

Figure D3: Assignee. This data file contains the assignee's name, address, city, and postal code.

and the execution date of the event. Similarly, Figure D3 presents the screenshot for tm_assignee, which contains the assignee's name, address, city, and postal code.

We use the "rf_id" to connect tm_convey, tm_assginor, and tm_assignee. We obtain the assignor-assignee pairs along with their respective conveyance types. For example, rf_id 07810445 indicates a name change from "OPAL CORDIAL CO. PTY. LIMITED " to "OPAL BEVERAGES AUSTRALIA PTY. LTD." rf_id 28900680 indicates a correction with the assignor "RUSSELL CORPORATION" and the assignee "RUSSELL ASSET MANAGEMENT, INC." Figure D4 is a screenshot for the matched data file. The matched data file contains the names of the assignor and assignee, along with their respective conveyance, city, and postal code information.

E The DISCERN dataset

This dataset connects patent assignees with Compustat firms, and its matching results in dynamic matching, company name changes, and ownership structures are superior to other similar datasets. We use the **DISCERN_UO_name_list.dta** data file and the **DISCERN_SUB_name_list.dta** data file, which are downloaded from Arora et al. (2021a). The ultimate owner (UO) and subsidiary historical

rf_id	conv_group	or_name	or_city	or_postcode	ee_name	ee_city	ee_postcode
01640766	assignment	COLORTRAN INDUSTRIES; BY: NATURAL LIGHTING CORP., A...	BURBANK		COLORTRAN INDUSTRIES, INC.	BURBANK	
17660861	name change	FAST SOFTWARE SECURITY GMBH & CO. KG			ALADDIN KNOWLEDGE SYSTEMS GMBH & CO. KG	D-82110-GEMERING	
49380622	security interest	NEWGROUND RESOURCES, INC.			BANK OF AMERICA, N.A.	CHICAGO	60603
07810445	name change	OPAL CORDIAL CO. PTY. LIMITED			OPAL BEVERAGES AUSTRALIA PTY. LTD.		
65050359	name change	KATHREIN-WERKE KG			KATHREIN SE	83022 ROSENHEIM	
70870289	assignment	ALLEN, ALEXANDER			ALLEN, ALEXANDER	APALACHICOLA	32320
60370359	name change	COTY GERMANY GMBH			COTY GERMANY GMBH	55116 MAINZ	
28900680	correction	RUSSELL CORPORATION			RUSSELL ASSET MANAGEMENT, INC.	WILMINGTON	19801
59990429	assignment	MULTI-STATE LOTTERY ASSOCIATION			MULTI-STATE LOTTERY ASSOCIATION	URBANDALE	50322
58270442	assignment	PROPARK AMERICA			PROPARK, INC.	HARTFORD	06103
30180212	assignment	RSI HOLDING CORPORATION			RSI HOME PRODUCTS, MANAGEMENT, INC.	NEWPORT BEACH	92660
63280470	name change	THOMASTIK-INFELD GESELLSCHAFT M.B.H.			THOMASTIK-INFELD GESELLSCHAFT M.B.H.	A-1050 WIEN	
75260979	assignment	BANK, JOSEPH ROBERT			VIDON PLASTICS, INC.	LAPEER	48446
03530295	assignment	BERTOIA, MARA LESTA			BERTOIA, VAL O.	TOWNSHIP OF HEREFORD, BERKS	
08200152	name change	AMALGAMATED BANK OF NEW YORK, THE			AMALGAMATED BANK OF NEW YORK		
58930442	name change	ITALPOLLINA SPA			ITALPOLLINA S.P.A.	I-37010 RIVOLI VERONESE (VR)	

Figure D4: Assignor-assignee pairs. This matched data file contains the assignor's name and assignee's name with their conveyance group, city, and postal code.

standardized name lists.

Figure E1 is the screenshot for **DISCERN_UO_name_list.dta**. It includes the ultimate owner (UO) standardized names and the dynamic reassignment of firms. Similarly, Figure E2 presents the screenshot for **DISCERN_SUB_name_list.dta**, which contains subsidiary historical standardized name and the dynamic reassignment of firms. The "id_name" in each data file is the ID of a company name, it corresponds to a standardized company name, "name_std."

F Invention keywords

F.1 Identifying the creation and novelty of invention keywords: Step by Step

In order to capture the development of the product inventions, we collect the goods and services identifications of all U.S. trademarks from 1870 to 2021 (n = 11,178,121).[41] We classify each trademark in the following for decades by its application year: 1980s, 1990s, 2000s, and 2010s,[42] because the coverage of trademarks in the USPTO could be incomplete before 1980 (Graham et al., 2013). While trademarks filed before 1980 are used to build up the baseline words.

For each identification, we tokenize the text to bi-grams and tri-grams, i.e., sequences of two adjacent words and three words of letters and numbers.[43] Next, we remove tokens only composed of numbers, one-character words, and stop words, i.e., commonly used words which can be ignored in the analysis. Our stop words list comprises from Python Natural Language Toolkit (NLTK) stop words,[44] stop words provided by Arts et al. (2021), and introductory words used in the trademarks.[45] We also remove tokens that occur in less than ten times in all trademarks. We then apply stemming to each token. Stemming is a method of reducing words to their word stem.

[41] We drop trademarks without filing date

[42] The 2010s in our paper covers from 2010 to 2021.

[43] Following Arts et al. (2021), we implement tokenization before stemming. We use the regular expression [a-z0-9][a-z0-9-]*[a-z0-9]+[a-z0-9], allowing tokens separated by "-."

[44] Examples of stop words from NLTK: am, are, is, the, of, any.

[45] The USPTO trademark examining procedure suggests to use these definite terms to further define the introductory wording. https://tmep.uspto.gov/RDMS/TMEP/current#/current/TMEP-1400d1e2196.html Introductory words in the USPTO: namely, consisting, particularly.

id_name	sample	name_std
36	U	A C F IND INC
8391	U	SHAREDATA INC
1687	U	CABOT MED CORP
2859	U	DENSE PACIFIC MICROSYS INC
2691	U	D P A C TECH CORP
116	U	A T O INC
3674	U	FIGGIE INTL INC
3673	U	FIGGIE INTL HLDG INC
8267	U	SCOTT TECH INC
4595	U	I F R SYS INC
3021	U	DOCUGRAPHIX INC
54	U	A E P IND INC
5225	U	J & J SNACK FOODS CORP
6632	U	NEUROTECH CORP
10272	U	XIOX CORP

Figure E1: Ultimate owner. This data file contains the ultimate owner and the dynamic reassignment of firms.

id_name	sample	name_std
29	B	INTERSTATE ELECTR CORP
38	B	STI FOREIGN SALES CORP
43	B	VIRGINIA CTR INC
30	B	MOJONNIER DE MEXICO S DE RL DE CV
31	B	MOJONNIER DO BRASIL INDUSTRIA E COMMERCIO DE EQ...
39	B	STI LICENSING CORP
3	B	CAFIG INC
47	B	WIMBLEDON SHIRT CO LTD
2	B	AV OX INC
40	B	STI PROPERTIES INC
6	B	CHAGRIN HQ VENTURE LTD
37	B	STI CANADIAN HLDG LTD
28	B	FP SPORTSWEAR BV
7	B	CUDAHY SELF STORAGE INC
4	B	CHAGRIN HIGHLANDS INC

Figure E2: Subsidiary. This data file contains subsidiary historical standardized names and the dynamic reassignment of firms.

We introduce an example of bi-grams: "satellite telecommunications network services, namely, voice and data transmission services." can be split into eight bi-grams: "satellite, telecommunications," "telecommunications, network," "network, services," "services, namely," "namely, voice," "voice, and," "and, data," and "transmission, services." We then remove bi-grams that contain stop words such as "services, namely," "namely, voice," "voice, and," and "and, data."[46] Finally, we stem each word to get the bi-grams "cell, phone," "cellular, communic," "radio, telecommun," "telecommun, servic," "cellular, telephon."

F.2 Example: No. 77721751

We use Twitter's trademark (No. 77721751, see Figure 2) as an example, this trademark contains 3 identifications, "Telecommunications services, namely, providing online and telecommunication facilities for real-time interaction between and among users of computers, mobile and handheld computers, and wired and wireless communication devices; enabling individuals to send and receive messages via email, instant messaging or a website on the internet in the field of general interest; providing on-line chat rooms and electronic bulletin boards for transmission of messages among users in the field of general interest; providing an online community forum for users to share information, photos, audio and video content about themselves, their likes and dislikes and daily activities, to get feedback from their

[46] A data structure consisting of multiple elements.

Table F1: Summary statistics for uni-grams

	Unique count	Mean	SD	Min	Median	95^{th}	99^{th}	Max
doc_freq_1grams	15702	127.55	850.303	10	22	355	1710	38452
$class\_freq\_1grams$	15702	127.92	861.849	10	22	348	1765	38601
all_freq_1grams	15702	145.35	1178.018	10	23	365	1896	71587
$doc\_freq\_1grams\_1980s$	4904	194.14	1059.206	10	31	632	1991	38452
$doc\_freq\_1grams\_1990s$	5706	128.82	949.131	10	22	334	1705	37066
$doc\_freq\_1grams\_2000s$	4004	62.43	360.087	10	18	200	727	17306
$doc\_freq\_1grams\_2010s$	1088	60.42	370.564	10	16	123	1182	9486
$class\_freq\_1grams\_1980s$	5043	203.25	1100.952	10	32	660	3128	38601
$class\_freq\_1grams\_1990s$	6021	131.23	964.781	10	23	332	1742	38002
$class\_freq\_1grams\_2000s$	4463	61.15	375.464	10	18	195	721	19618
$class\_freq\_1grams\_2010s$	1385	54.36	341.374	10	15	104	495	9573
$all\_freq\_1grams\_1980s$	6723	158.98	1399.408	10	25	367	1982	71587
$all\_freq\_1grams\_1990s$	5370	255.36	1569.468	10	35	787	3938	64567
$all\_freq\_1grams\_2000s$	5631	63.75	435.452	10	19	195	685	25377
$all\_freq\_1grams\_2010s$	2373	51.39	359.448	10	16	93	419	11743

Notes: The mean (median) document frequency, class frequency, and overall frequency of uni-grams are respectively 127.55 (22), 127.92 (22), and 145.35 (23). The 95% (99%) percentile of document frequency, class frequency, and overall frequency are respectively 355 (1710), 348 (1765), and 348 (1765). We also report the minimum, maximum, and standard deviation.

peers, to form virtual communities, and to engage in social networking" (in IC 038), "Providing on-line journals, namely, blogs featuring user-defined content." (in IC 041), and "Online social networking services; providing a website on the internet for the purpose of social networking; providing on-line computer databases and on-line searchable databases in the field of social networking." (in IC 045). The identification in IC 038 mentions "social network" for 1 time, and the identification in IC 045 mentions "social network" for 3 times, then "social network" is used for 1 time in doc_freq_2grams, 2 times in $class\_freq\_2grams$, and 4 times in all_freq_2grams.

F.3 Summary statistics of invention keywords

This section reports the summary statistics and distribution of uni/bi/tri-grams invention keywords.

The mean (median) document frequency, class frequency, and overall frequency of uni-grams are respectively 127.55 (22), 127.92 (22), and 145.35 (23). The 95% (99%) percentile of document frequency, class frequency, and overall frequency are respectively 355 (1710), 348 (1765), and 348 (1765).

Next, we calculate the frequency of each word in each decade in terms of document frequency, class frequency, and overall frequency. These frequencies help us understand when an invention was created and how it prevails and declines. For example, the uni-grams "e-commerc" was first used in the 1990s, and there were 368 trademarks using this term, and it became popular in the next three decades. There are 2,714 trademarks using this word in the 2000s and 6,620 trademarks using this word in the 2010s. Therefore, $doc\_freq\_1grams\_1980s$, $doc\_freq\_1grams\_1990s$, $doc\_freq\_1grams\_2000s$, and $doc\_freq\_1grams\_2010s$ of "e-commerc" are respectively 0, 368, 2,740, and 6,620. We also calculate the class frequency and overall frequency in each decade.

Overall, in the 1980s, 4,904 uni-grams (which we label as "invention keywords") were introduced and

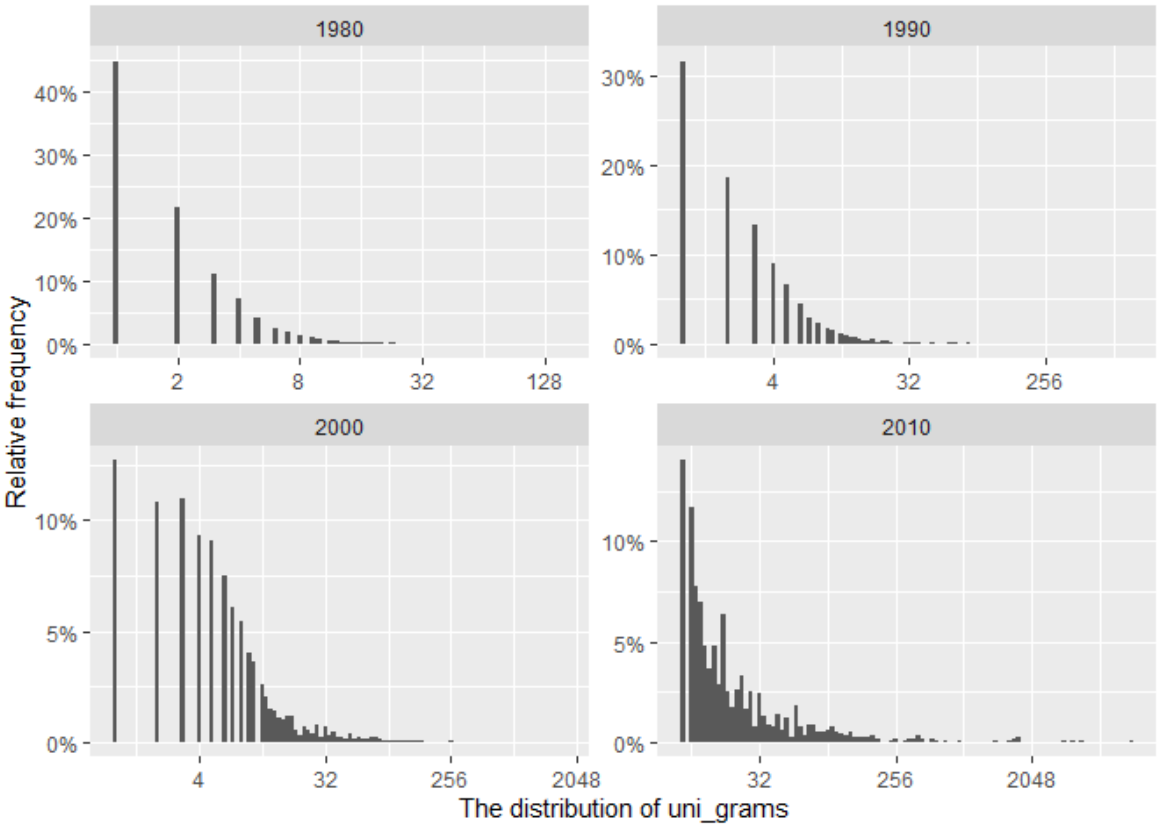


Figure F1: Distribution of uni-grams in the decades they first appear. This figure shows the distribution of document frequency for uni-grams in the decades 1980s, 1990s, 2000s, and 2010s. The vertical axis represents the relative frequency while the horizontal axis displays the document frequency value.

used on average by 194.14 subsequent trademarks, with a median of 31 and a standard deviation of 1059.206. In the 1990s, 5,706 uni-grams were introduced and used on average by 128.82 subsequent trademarks, with a median of 22 and a standard deviation of 949.131. In the 2000s, 4,004 uni-grams were introduced and used on average by 62.43 subsequent trademarks, with a median of 18 and a standard deviation of 360.087. In the 2010s, 1,088 uni-grams were introduced and used on average by 60.42 subsequent trademarks, with a median of 16 and a standard deviation of 370.564.

The mean (median) document frequency in each decade (the 1980s, 1990s, 2000s, and 2010s) are respectively 194.14 (31), 128.82 (22), 62.43 (18), and 60.42 (16). The mean (median) class frequency in each decade are respectively 203.25 (32), 131.23 (23), 61.15 (18), and 54.36 (15). The mean (median) overall frequency in each decade are respectively 158.98 (25), 255.36 (35), 63.75 (19), and 51.39 (16).

Figure F1 shows the distribution of document frequency of uni-grams in the 1980s, 1990s, 2000s, and 2010s. We find that a myriad of invention keywords are rarely used in the first ten years of implementation. More than 50% of keywords were used <= 4 times in their first decade. A similar distribution shows in the 2010s, while around 50% of keywords were used <= 32 times in the 2010s. On the other hand, fewer than 1% of invention keywords were used with frequencies exceeding 128, 256, 256, and 2048 times in the 1980s, 1990s, 2000s, and 2010s, respectively. All these are consistent

Table F2: Summary statistics for bi-grams

	Unique count	Mean	SD	Min	Median	95^{th}	99^{th}	Max
doc_freq_2grams	406437	70.37	524.783	10	20	214	846	160520
class_freq_2grams	406437	70	560.349	10	20	211	834	190927
all_freq_2grams	406437	73.8	740.935	10	21	215	856	284818
doc_freq_2grams_1980s	93380	124.48	944.877	10	28	414	1613	160520
doc_freq_2grams_1990s	139591	70.96	415.715	10	22	215	819	78272
doc_freq_2grams_2000s	123130	44.1	168.34	10	17	133	461	14923
doc_freq_2grams_2010s	50336	32.63	101.648	10	16	90	294	6018
class_freq_2grams_1980s	95452	126.95	1037.545	10	28	419	1625	190927
class_freq_2grams_1990s	144931	71.65	430.882	10	22	216	817	85272
class_freq_2grams_2000s	131712	43.28	166.433	10	18	129	449	16822
class_freq_2grams_2010s	55727	31.34	98.025	10	15	84	278	6057
all_freq_2grams_1980s	99161	143.62	1453.819	10	30	455	1804	284818
all_freq_2grams_1990s	154404	79.12	551.642	10	23	236	884	107901
all_freq_2grams_2000s	154404	44.34	186.927	10	18	130	451	31842
all_freq_2grams_2010s	75795	30.03	100.4	10	16	78	244	9979

Notes: The mean (median) document frequency, class frequency, and overall frequency of bi-grams are respectively 70.37 (20), 70 (20), and 73.8 (21). The 95% (99%) percentile of document frequency, class frequency, and overall frequency are respectively 214 (846), 211 (834), and 215 (856). We also report the minimum, maximum, and standard deviation.

with the long-tail phenomenon in innovation: only a few inventions will succeed, while most others will fail.

Table F2 presents the summary statistics of bi-grams in trademark documents, the mean (median) document frequency, class frequency, and overall frequency of bi-grams are 70.37 (20), 70.63 (20), and 73.8 (21), respectively. The 95% (99%) percentile of document frequency, class frequency, and overall frequency are 214 (846), 211 (834), and 215 (856), respectively. The distributions are extremely skewed to the right.

Next, we calculate the frequency of each word in each decade in terms of document frequency, class frequency, and overall frequency. These frequencies help us understand when an invention was created and how it prevails and declines. For example, the bi-grams "on-lin retail" was first used in the 1990s, and there were 4,454 subsequent trademarks using this word, and it became popular in the next two decades. There are 22,027 subsequent trademarks using this word in the 2000s and 55,408 subsequent trademarks using this word in the 2010s. Therefore, *doc_freq_2grams_1980s*, *doc_freq_2grams_1990s*, *doc_freq_2grams_2000s*, and *doc_freq_2grams_2010s* of "on-lin retail" are respectively 0, 4,454, 22,027, and 55,408. We also calculate the class frequency and overall frequency in each decade.

Overall, in the 1980s, 93,380 bi-grams (which we label as "invention keywords") were introduced and used on average by 124.48 trademarks, with a median of 28 and a standard deviation of 944.877. In the 1990s, 139,591 bi-grams were introduced and used on average by 70.96 trademarks, with a median of 22 and a standard deviation of 415.715. In the 2000s, 123,130 bi-grams were introduced and used on average by 44.1 trademarks, with a median of 17 and a standard deviation of 168.64. In the 2010s, 50,336 bi-grams were introduced and used on average by 32.63 trademarks, with a median of 16 and

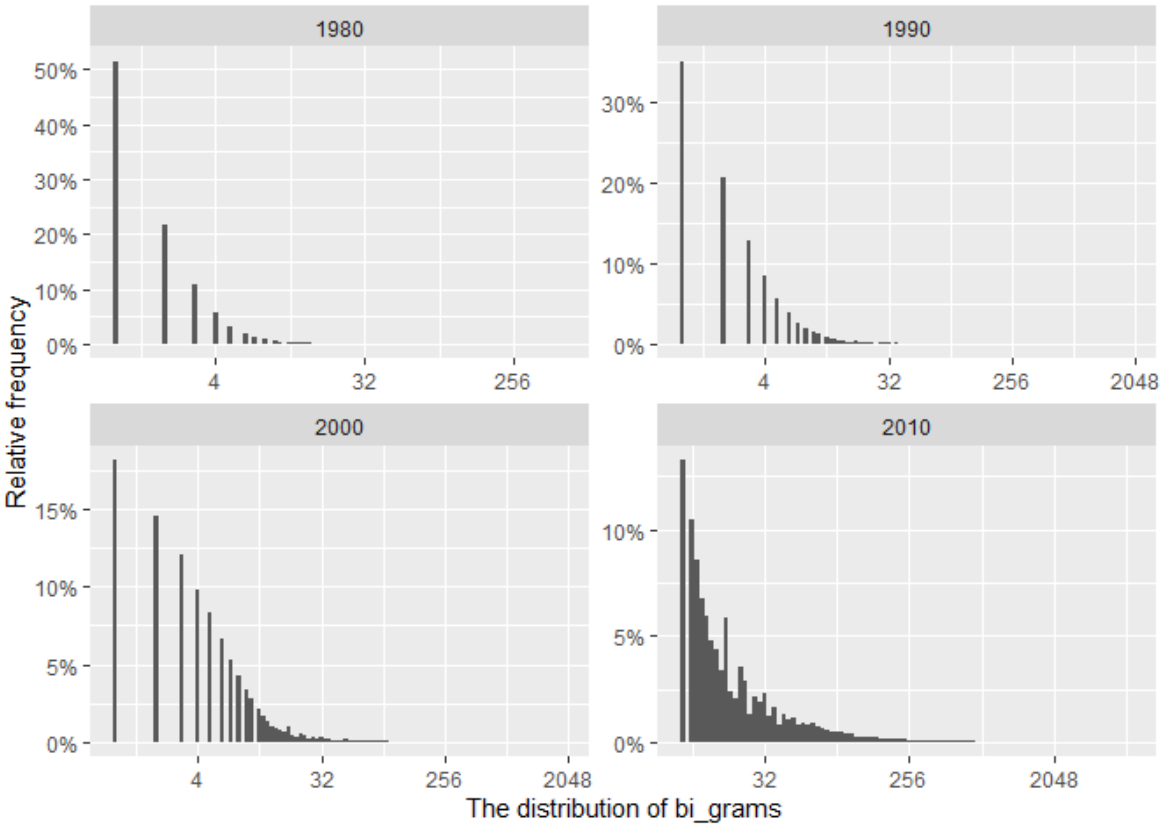


Figure F2: Distribution of bi-grams in the decades they first appear. This figure shows the distribution of document frequency for bi-grams in the decades 1980s, 1990s, 2000s, and 2010s. The vertical axis represents the relative frequency while the horizontal axis displays the document frequency value.

a standard deviation of 101.648.

The mean (median) document frequency in each decade (the 1980s, 1990s, 2000s, and 2010s) are 124.48 (28), 70.96 (22), 44.1 (17), and 32.63 (16), respectively. The mean (median) class frequency in each decade are 126.95 (28), 71.65 (22), 43.28 (18), and 31.34 (15), respectively. The mean (median) overall frequency in each decade are respectively 143.62 (30), 79.12 (23), 44.34 (18), and 30.03 (16).

Figure F2 shows the distribution of document frequency of bi-grams in the 1980s, 1990s, 2000s, and 2010s. We find that a myriad of invention keywords are rarely used in the first ten years of implementation. 80%, 70%, and 50% of keywords were used <= 4 times in the 1980s, 1990s, and 2000s. A similar distribution is shown in the 2010s, 50% of keywords were used <= 32 times in the 2010s. On the other hand, fewer than 1% of invention keywords were used with frequencies exceeding 32, 32, 256, and 1024 times in the 1980s, 1990s, 2000s, and 2010s, respectively. All these are consistent with the long-tail phenomenon in innovation: only a few inventions will succeed, while most others will fail.

Table F3: Summary statistics for tri-grams

	Unique count	Mean	SD	Min	Median	95^{th}	99^{th}	Max
doc_freq_3grams	645579	58.11	366.29	10	19	177	672	209473
class_freq_3grams	645579	51.03	276.022	10	18	156	541	124415
all_freq_3grams	645579	50.96	328.716	10	18	151	535	163309
doc_freq_3grams_1980s	78922	98.4	572.804	10	25	345	1207	109304
doc_freq_3grams_1990s	169314	62.18	290.222	10	21	200	673	48396
doc_freq_3grams_2000s	238870	43.44	149.521	10	18	134	429	20821
doc_freq_3grams_2010s	158473	30.13	71.981	10	16	86	249	6004
class_freq_3grams_1980s	80237	98.87	611.139	10	25	346	1212	124415
class_freq_3grams_1990s	174245	62	291.494	10	21	198	668	48409
class_freq_3grams_2000s	250347	42.72	147.673	10	18	130	421	21052
class_freq_3grams_2010s	169393	29.4	70.192	10	15	83	239	6040
all_freq_3grams_1980s	82556	105.49	784.462	10	26	358	1278	163309
all_freq_3grams_1990s	183136	65.08	351.098	10	22	204	702	55521
all_freq_3grams_2000s	277315	42.86	158.49	10	18	128	421	25965
all_freq_3grams_2010s	215194	28.49	73.872	10	15	77	224	9868

Notes: The mean (median) document frequency, class frequency, and overall frequency of bi-grams are 58.11 (19), 51.03 (18), and 50.96 (18), respectively. The 95% (99%) percentile of document frequency, class frequency, and overall frequency are 177 (672), 156 (541), and 151 (535), respectively. We also report the minimum, maximum, and standard deviation.

Table F3 presents the summary statistics of tri-grams in trademark documents, the mean (median) document frequency, class frequency, and overall frequency of bi-grams are 58.11 (19), 51.03 (18), and 50.96 (18), respectively. The 95% (99%) percentile of document frequency, class frequency, and overall frequency are 177 (672), 156 (541), and 151 (535), respectively. The distributions are extremely skewed to the right.

Similarly, the appearance and frequency of tri-grams were calculated for each decade. For instance, "prerecord comput program" was first introduced in the 1980s, with 1038 trademarks using this term. However, the number of trademarks using "electron mail servic" decreased in the 1990s, 2000s, and 2010s with 298, 168, and 88. Accordingly, *doc_freq_3grams_1980s*, *doc_freq_3grams_1990s*, *doc_freq_3grams_2000s*, and *doc_freq_3grams_2010s* of "prerecord comput program" are respectively 1,038, 298, 168, and 88, respectively. We also calculate the class frequency and overall frequency for each decade in the same table.

Overall, in the 1980s, 78,922 tri-grams were introduced as invention keywords and were used by an average of 98.4 subsequent trademarks, with a median of 25 and a standard deviation of 572.804. In the 1990s, 169,314 tri-grams were introduced and used by an average of 62.18 subsequent trademarks, with a median of 21 and a standard deviation of 290.222. In the 2000s, 238,870 tri-grams were introduced and used by an average of 43.44 subsequent trademarks, with a median of 18 and a standard deviation of 149.521. In the 2010s, 158,473 tri-grams were introduced and used by an average of 30.13 subsequent trademarks, with a median of 16 and a standard deviation of 71.981.

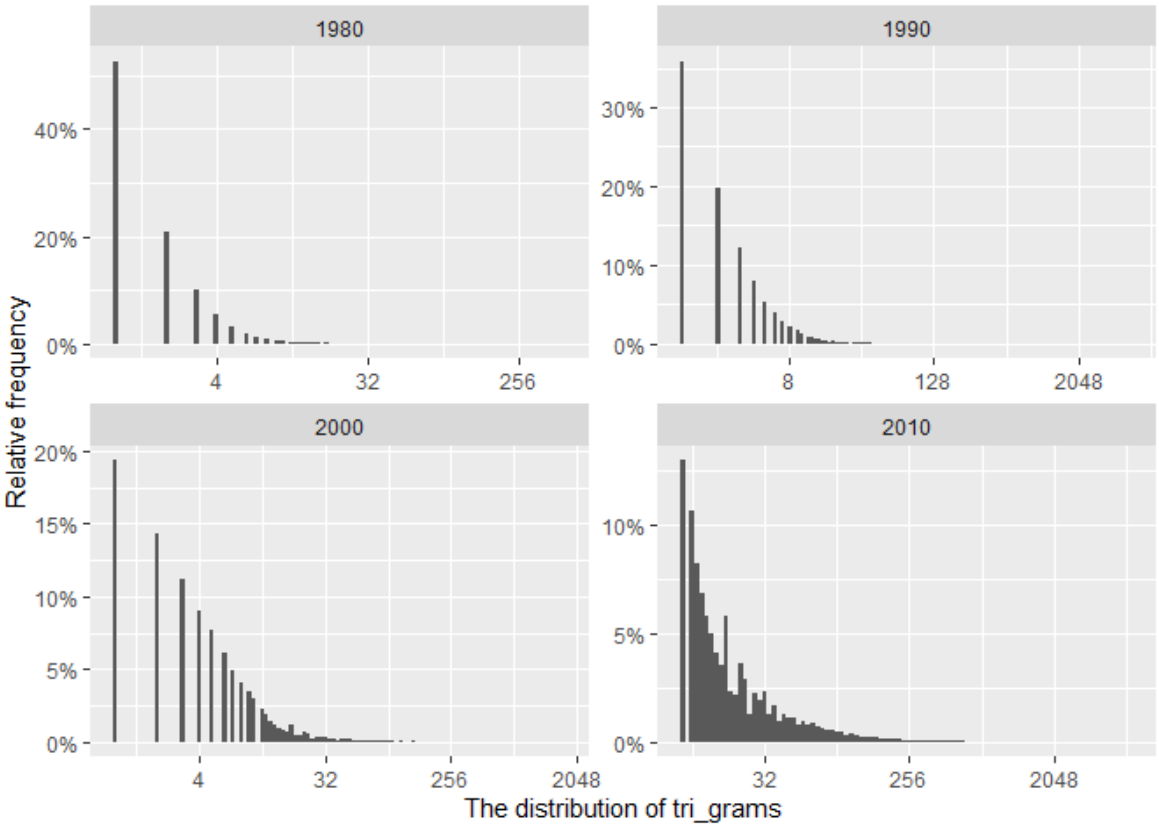


Figure F3: Distribution of tri-grams in the decades they first appear. This figure shows the distribution of document frequency for tri-grams in the decades 1980s, 1990s, 2000s, and 2010s. The vertical axis represents the relative frequency while the horizontal axis displays the document frequency value.

The mean (median) document frequency, class frequency, and overall frequency of tri-grams are respectively 58.11 (19), 51.03 (18), and 50.96 (18). The 95% (99%) percentile of document frequency, class frequency, and overall frequency are respectively 177 (672), 156 (541), and 151 (535). Not surprisingly, the distributions are extremely skewed to the right. The mean (median) document frequency in each decade (1980s, 1990s, 2000s, and 2010s) are 98.4 (25), 62.18 (21), 43.44 (18), and 30.13 (16), respectively. The mean (median) class frequency in each decade are 98.87 (25), 62 (21), 42.72 (18), and 29.4 (15), respectively. The mean (median) overall frequency in each decade are 105.49 (26), 65.08 (22), 42.86 (18), and 28.49 (15), respectively.

Figure F3 presents the distribution of document frequency of tri-grams invention keywords in the 1980s, 1990s, 2000s, and 2010s. We observe that 80%, 70%, and 55% of invention keywords were used <= 4 times in the 1980s, 1990s, and 2000s. A similar distribution in the 2010s, while 50% of keywords were used <= 32 times in the 2010s. On the other hand, fewer than 1% of invention keywords were used with frequencies exceeding 32, 32, 256, and 1024 times in the 1980s, 1990s, 2000s, and 2010s, respectively. All these are consistent with the long-tail phenomenon in innovation: only a few inventions will succeed, while most others will fail.

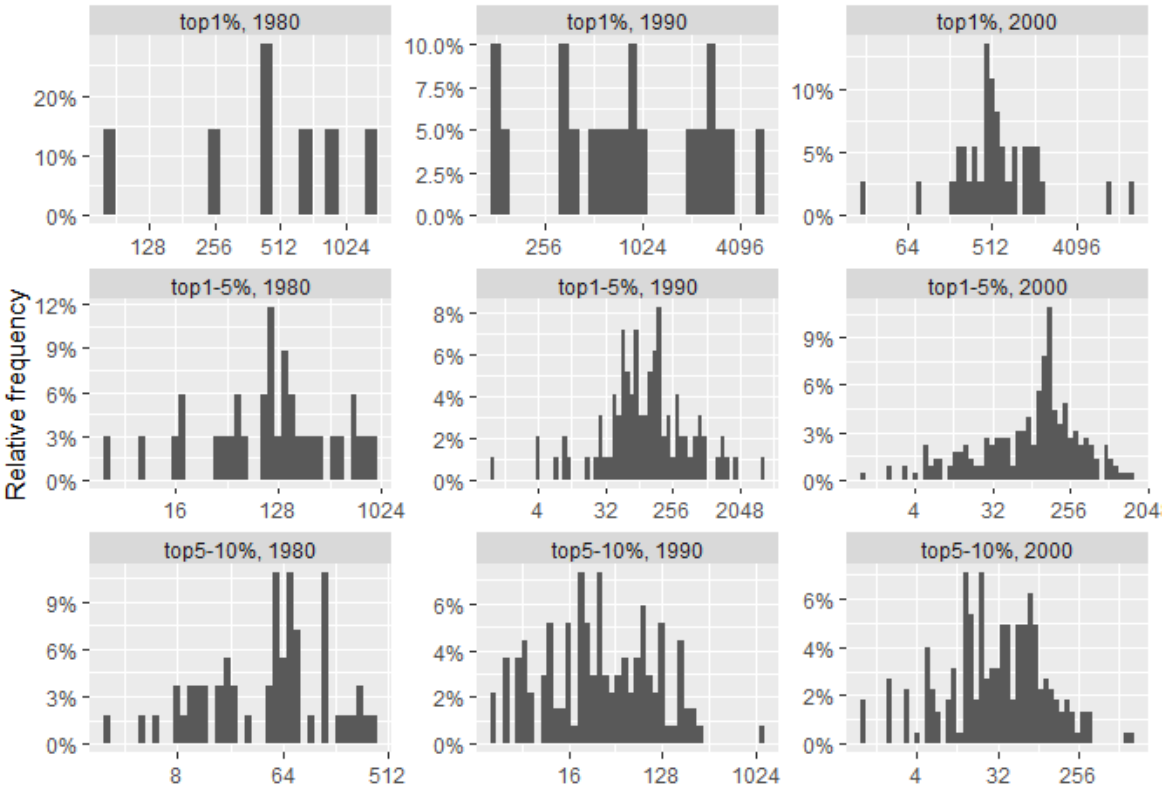


Figure F4: Prevalence of invention keywords uni-grams in the 1980s, 1990s, and 2000s (Log Scale).

F.4 The prevalence of heterogeneous invention keywords

We then analyze the prevalence of heterogeneous invention keywords by examining the frequencies of the top 1%, top 1-5%, and top 5-10% invention keywords. We examine the frequencies of the top 1%, top 1-5%, and top 5-10% invention keywords in the decades right *after* their creation decades.

For uni-grams invention keywords in the 1980s, the top 1% group consists of 7 invention keywords with frequencies in the $(79, +\infty)$ interval; the top 1-5% group consists of 34 invention keywords with frequencies in the $(22, 79]$ interval; the top 5-10% group consists of 60 invention keywords with frequencies in the $(14, 22]$ interval. For uni-grams invention keywords in the 1990s, the top 1% group consists of 20 invention keywords with frequencies in the $(79, +\infty)$ interval; the top 1-5% group consists of 34 invention keywords with frequencies in the $(22, 79]$ interval; the top 5-10% group consists of 139 invention keywords with frequencies in the $(14, 22]$ interval. For uni-grams invention keywords in the 2000s, the top 1% group consists of 38 invention keywords with frequencies in the $(79, +\infty)$ interval; the top 1-5% group consists of 237 invention keywords with frequencies in the $(22, 79]$ interval; the top 5-10% group consists of 233 invention keywords with frequencies in the $(14, 22]$ interval.

Figure F4 presents the frequencies of the uni-grams invention keywords in the decade right after their creation decade with a logarithmic scale on the x-axis (due to the skewness of frequencies). In these plots, we observe that most top 1% invention keywords in the 1980s, 1990s, and 2000s remained popular in their second decade. Among all the top 1% invention keywords born in the 1980s, "workstat" was

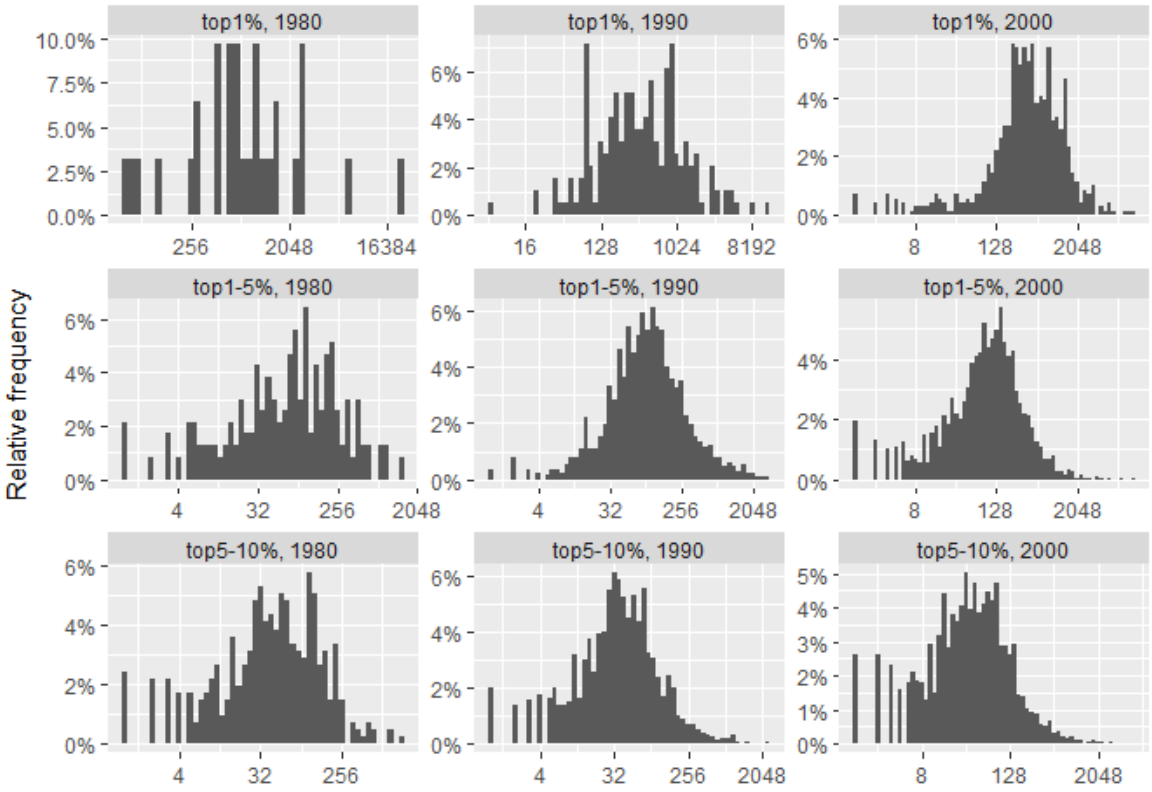


Figure F5: Prevalence of invention keywords bi-grams in the 1980s, 1990s, and 2000s (Log Scale).

the most frequently used one and appeared in 1,423 trademarks in the 1990s. Among all the top 1-5% invention keywords created in the 1980s, "videotex" was the most frequently used one and appeared in 861 trademarks in the 1990s. Among all the top 5-10% invention keywords created in the 1980s, "yoyo" was the most frequently used one and appeared in 371 trademarks in the 1990s.

Among all the top 1% invention keywords born in the 1990s, "e-mail" was the most frequently used one and appeared in 5,469 trademarks in the 2000s. Among all the top 1-5% invention keywords created in the 1990s, "mp3" was the most frequently used one and appeared in 3,964 trademarks in the 2000s. Among all the top 5-10% invention keywords created in the 1990s, "internet-bas" was the most frequently used one and appeared in 1,102 trademarks in the 2000s.

Among all the top 1% invention keywords born in the 2000s, "podcast" was the most frequently used one and appeared in 15,682 trademarks in the 2010s. Among all the top 1-5% invention keywords created in the 2000s, "webisod" was the most frequently used one and appeared in 1,326 trademarks in the 2010s. Among all the top 5-10% invention keywords created in the 2000s, "vegetable-fruit" was the most frequently used one and appeared in 1,001 trademarks in the 2010s.

Overall, most top 1% invention keywords are used 256 to 2048 times in their second decade, top 1-5% groups are used 32 to 256 times in their second decade, and the top 5-10% are used 4 to 256 times. Indeed, all plots show that the distributions of frequencies are highly right-skewed with a long right tail.

For bi-grams invention keywords in the 1980s, the top 1% group consists of 31 invention keywords with frequencies in the $(76, +\infty)$ interval; the top 1-5% group consists of 239 invention keywords with frequencies in the $(24, 76]$ interval; the top 5-10% group consists of 439 invention keywords with frequencies in the $(15, 24]$ interval. For bi-grams invention keywords in the 1990s, the top 1% group consists of 196 invention keywords with frequencies in the $(76, +\infty)$ interval; the top 1-5% group consists of 1330 invention keywords with frequencies in the $(24, 76]$ interval; the top 5-10% group consists of 2177 invention keywords with frequencies in the $(15, 24]$ interval. For bi-grams invention keywords in the 2000s, the top 1% group consists of 707 invention keywords with frequencies in the $(76, +\infty)$ interval; the top 1-5% group consists of 3,611 invention keywords with frequencies in the $(24, 76]$ interval; the top 5-10% group consists of 4,856 invention keywords with frequencies in the $(15, 24]$ interval.

Figure F5 presents the frequencies of the uni-grams invention keywords in the decade right after their creation decade with a logarithmic scale on the x-axis (due to the skewness of frequencies). In these plots, we observe that most top 1% invention keywords in the 1980s, 1990s, and 2000s remained popular in their second decade. Among all the top 1% invention keywords born in the 1980s, "comput networ" was the most frequently used one and appeared in 22,639 trademarks in the 1990s. Among all the top 1-5% invention keywords created in the 1980s, "interact comput" was the most frequently used one and appeared in 1,429 trademarks in the 1990s. Among all the top 5-10% invention keywords created in the 1980s, "unmount photograph" was the most frequently used one and appeared in 1,219 trademarks in the 1990s.

Among all the top 1% invention keywords born in the 1990s, "on-lin retail" was the most frequently used one and appeared in 12,625 trademarks in the 2000s. Among all the top 1-5% invention keywords created in the 1990s, "non-download comput" was the most frequently used one and appeared in 3,174 trademarks in the 2000s. Among all the top 5-10% invention keywords created in the 1990s, "mp3 player" was the most frequently used one and appeared in 2,484 trademarks in the 2000s.

Among all the top 1% invention keywords born in the 2000s, "cloud comput" was the most frequently used one and appeared in 14,814 trademarks in the 2010s. Among all the top 1-5% invention keywords created in the 2000s, "download mobil" was the most frequently used one and appeared in 14,808 trademarks in the 2010s. Among all the top 5-10% invention keywords created in the 2000s, "phone tablet" was the most frequently used one and appeared in 6,024 trademarks in the 2010s.

Overall, most top 1% invention keywords are used 256 to 2048 times in their second decade, top 1-5% groups are used 32 to 256 times in their second decade, and the top 5-10% are used 4 to 128 times. Indeed, all plots show that the distributions of frequencies are highly right-skewed with a long right tail.

For tri-grams invention keywords in the 1980s, the top 1% group consists of 13 invention keywords with frequencies in the $(111, +\infty)$ interval; the top 1-5% group consists of 119 invention keywords

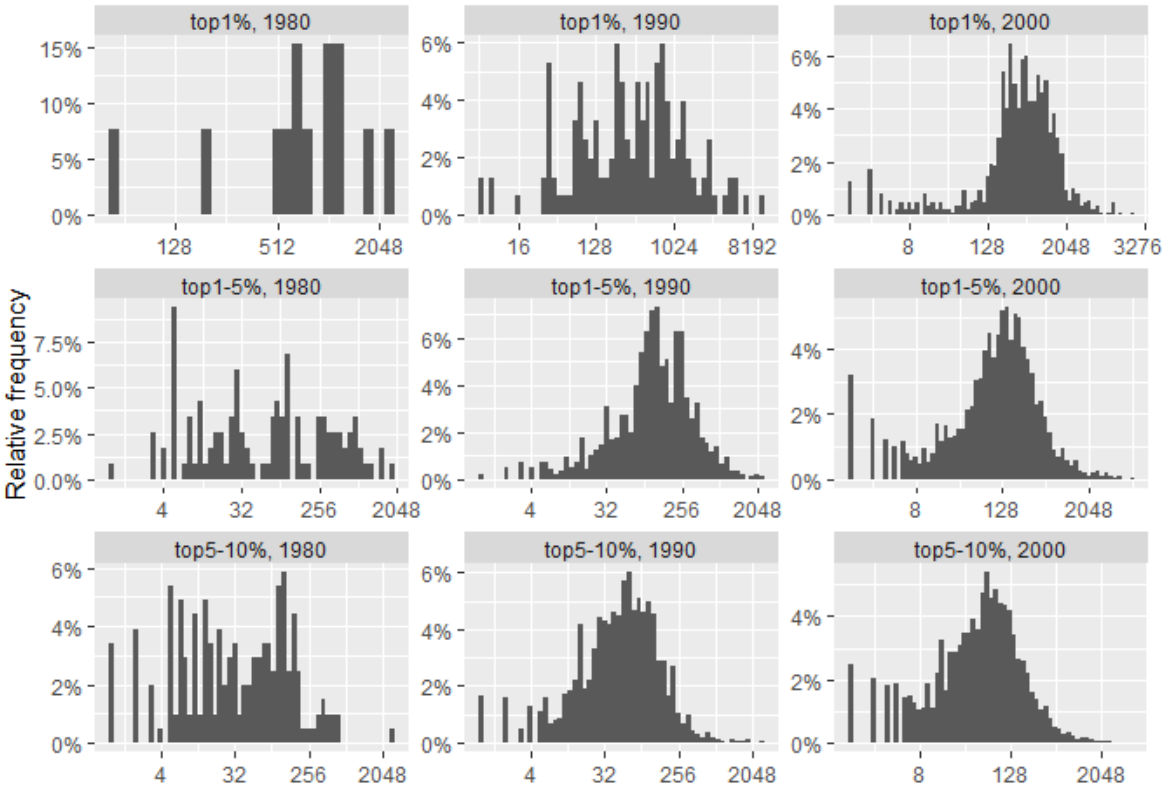


Figure F6: Prevalence of invention keywords tri-grams in the 1980s, 1990s, and 2000s (Log Scale).

with frequencies in the $(37, 111]$ interval; the top 5-10% group consists of 224 invention keywords with frequencies in the $(22, 37]$ interval. For tri-grams invention keywords in the 1990s, the top 1% group consists of 151 invention keywords with frequencies in the $(111, +\infty)$ interval; the top 1-5% group consists of 972 invention keywords with frequencies in the $(37, 111]$ interval; the top 5-10% group consists of 1,577 invention keywords with frequencies in the $(22, 37]$ interval. For tri-grams invention keywords in the 2000s, the top 1% group consists of 927 invention keywords with frequencies in the $(111, +\infty)$ interval; the top 1-5% group consists of 4,489 invention keywords with frequencies in the $(37, 111]$ interval; the top 5-10% group consists of 6,195 invention keywords with frequencies in the $(22, 37]$ interval.

Figure F6 presents the frequencies of the uni-grams invention keywords in the decade right after their creation decade with a logarithmic scale on the x-axis (due to the skewness of frequencies). In these plots, we observe that most top 1% invention keywords in the 1980s, 1990s, and 2000s remained popular in their second decade. Among all the top 1% invention keywords born in the 1980s, "user manual sold" was the most frequently used one and appeared in 2,244 trademarks in the 1990s. Among all the top 1-5% invention keywords created in the 1980s, "compact disc featur" was the most frequently used one and appeared in 1,794 trademarks in the 1990s. Among all the top 5-10% invention keywords created in the 1980s, "comput softwar use" was the most frequently used one and appeared in 2,598 trademarks in the 1990s.

Among all the top 1% invention keywords born in the 1990s, "on-lin retail store" was the most frequently used one and appeared in 10,495 trademarks in the 2000s. Among all the top 1-5% invention keywords created in the 1990s, "websit featur inform" was the most frequently used one and appeared in 2,458 trademarks in the 2000s. Among all the top 5-10% invention keywords created in the 1990s, "non-download comput softwar" was the most frequently used one and appeared in 2,773 trademarks in the 2000s.

Among all the top 1% invention keywords born in the 2000s, "portabl media player" was the most frequently used one and appeared in 20,332 trademarks in the 2010s. Among all the top 1-5% invention keywords created in the 2000s, "social entertain event" was the most frequently used one and appeared in 7,712 trademarks in the 2010s. Among all the top 5-10% invention keywords created in the 2000s, "cloud comput featur" was the most frequently used one and appeared in 5,315 trademarks in the 2010s.

Overall, most top 1% invention keywords are used 512 to 2048 times in their second decade, top 1-5% groups are used 128 to 256 times in their second decade, and the top 5-10% are used 32 to 128 times. Indeed, all plots show that the distributions of frequencies are highly right-skewed with a long right tail.

In this section, we show how the top 1%, top 1-5%, and top 5-10% invention keywords evolve in the next decade, as measured by the frequencies of these invention words in the decade after their creation. The most frequently used invention keywords in the top 1% group are much more than the 1-5% and the 5-10% group, suggesting successful inventions' disproportionate share of trademark use. We also find that on average top 1% invention keywords are used by more trademarks than the top 1-5% and top 5-10% group in the next decade. As a result, invention keywords exhibit a higher frequency in the creation period, which will be used more frequently in the future, which is consistent with the "winners-being-winners phenomenon."

F.5 The long-term development of heterogeneous invention keywords

We further consider eight groups by the frequencies of their use in the decade it was created: top 1%, top 1-5%, top 5-10%, top 10-20%, top 20-30%, top 30-40%, top 40-50%, and bottom 50%. This analysis aims to explore the ongoing evolution of invention keywords from their inception through the 2010s.

Panel 1980 of Figure F7 presents the development of uni-grams invention keywords for each group. It presents the development of invention keywords born in the 1980s. We find that invention keywords with higher initial frequencies tend to become more prominent in the future. For the top 1%, the average frequencies in the 1980s, 1990s, 2000s, and 2010s were respectively 111, 593.86, 1,387.14, and 3,751.43. The increase rates of the lower-frequency groups are lower than the top 1% group. The average frequencies of the top 1-5% group in the 1980s, 1990s, 2000s, and 2010s were respectively 34.56, 209.47, 453.15, and 1128.26. The average frequencies of the top 5-10% group in the 1980s,

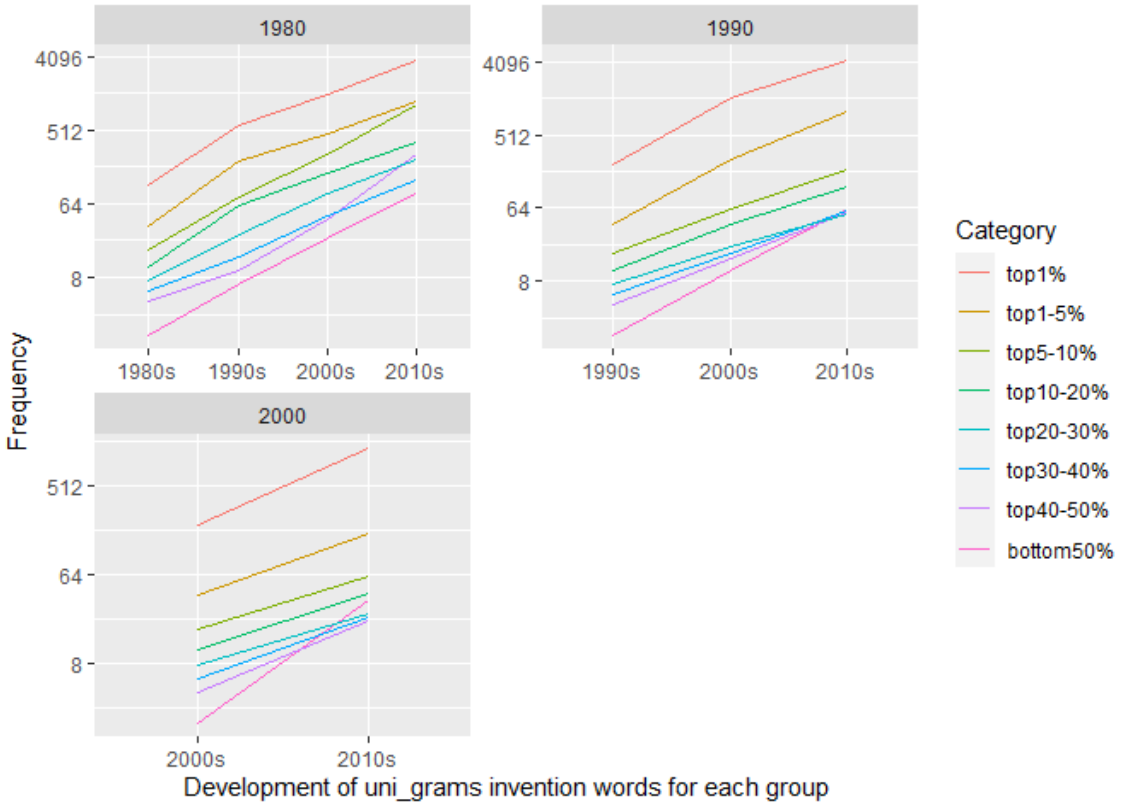


Figure F7: Long-term development of uni-grams Invention Words for Each Group. Panel 1980 presents the development of invention keywords created in the 1980s. Panel 1990 presents the development of invention keywords created in the 1990s. Panel 2000 presents the development of invention keywords created in the 2000s.

1990s, 2000s, and 2010s were respectively 17.53, 78.03, 261.38, and 1033.72. The average frequencies of the top 10-20% group in the 1980s, 1990s, 2000s, and 2010s were respectively 10.6, 61.98, 155.32, and 354.82. The average frequencies of the top 20-30% group in the 1980s, 1990s, 2000s, and 2010s were respectively 7.39, 26.82, 84.22, and 220.12. The average frequencies of the top 30-40% group in the 1980s, 1990s, 2000s, and 2010s were respectively 5.37, 14.5, 45.77, and 124.33. The average frequencies of the top 40-50% group in the 1980s, 1990s, 2000s, and 2010s were respectively 4, 9.62, 40.57, and 256.37. The average frequencies of the bottom 50% group in the 1980s, 1990s, 2000s, and 2010s were respectively 1.56, 6.49, 61.13, and 85.13.

Panel 1990 of Figure F7 presents the development of invention keywords created in the 1990s. Similar to Panel 1980, for the top 1%, the average frequencies in the 1990s, 2000s, and 2010s were respectively 222.6, 1468.05, and 4401.45. The average frequencies of the top 1-5% group in the 1990s, 2000s, and 2010s were respectively 40.95, 254.69, and 1001.38. The average frequencies of the top 5-10% group in the 1990s, 2000s, and 2010s were respectively 17.83, 61.94, and 194.61. The average frequencies of the top 10-20% group in the 1990s, 2000s, and 2010s were respectively 10.96, 41.17, and 118.49. The average frequencies of the top 20-30% group in the 1990s, 2000s, and 2010s were respectively 7.44, 21.2, and 54.17. The average frequencies of the top 30-40% group in the 1990s, 2000s, and 2010s were respectively 5.4, 17.34, and 60.32. The average frequencies of the top 40-50% group in the 1990s, 2000s, and 2010s were respectively 4, 15.07, and 57.22. The average frequencies of the bottom 50% group in the 1990s, 2000s, and 2010s were respectively 1.71, 10.65, and 61.13.

As shown in Panel 2000 of Figure F7, for the top 1%, the average frequencies in the 2000s and 2010s were respectively 200.87, and 1258.32. The average frequencies of the top 1-5% group in the 2000s and 2010s were respectively 39.54 and 165.95. The average frequencies of the top 5-10% group in the 2000s and 2010s were respectively 17.69 and 61.47. The average frequencies of the top 10-20% group in the 2000s and 2010s were respectively 10.85 and 41.13. The average frequencies of the top 20-30% group in the 2000s and 2010s were respectively 7.48 and 25.26. The average frequencies of the top 30-40% group in the 2000s and 2010s were respectively 5.45 and 23.25. The average frequencies of the top 40-50% group in the 2000s and 2010s were respectively 4 and 21.8. The average frequencies of the bottom 50% group in the 2000s and 2010s were respectively 1.95 and 35.41.

Panel 1980 of Figure F8 presents the development of bi-grams invention keywords for each group. It presents the development of invention keywords born in the 1980s. We find that invention keywords with higher initial frequencies tend to become more prominent in the future. For the top 1%, the average frequencies in the 1980s, 1990s, 2000s, and 2010s were respectively 171.03, 1,844.19, 3,839.29, and 5,901.71. The increase rates of the lower-frequency groups are lower than the top 1% group. The average frequencies of the top 1-5% group in the 1980s, 1990s, 2000s, and 2010s were respectively 36.57, 143.42, 356.71, and 822.45. The average frequencies of the top 5-10% group in the 1980s, 1990s, 2000s, and 2010s were respectively 18.97, 76.42, 213.89, and 483.42. The average frequencies of the top 10-20% group in the 1980s, 1990s, 2000s, and 2010s were respectively 12.51, 44.21, 129.13, and 355.69.

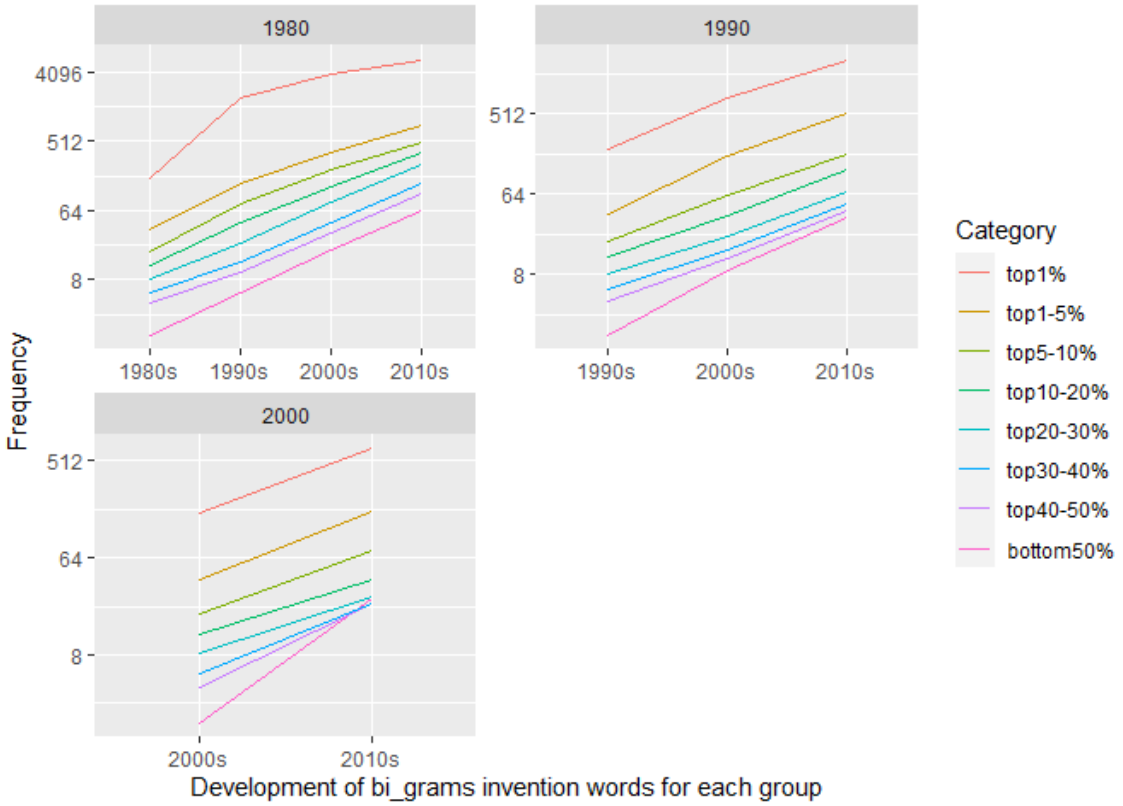


Figure F8: Long-term development of bi-grams Invention Words for Each Group. Panel 1980 presents the development of invention keywords created in the 1980s. Panel 1990 presents the development of invention keywords created in the 1990s. Panel 2000 presents the development of invention keywords created in the 2000s.

The average frequencies of the top 20-30% group in the 1980s, 1990s, 2000s, and 2010s were respectively 8.06, 23.35, 80.97, and 250.73. The average frequencies of the top 30-40% group in the 1980s, 1990s, 2000s, and 2010s were respectively 5.39, 13.83, 45.14, and 140.97. The average frequencies of the top 40-50% group in the 1980s, 1990s, 2000s, and 2010s were respectively 4, 9.98, 33.27, and 107.86. The average frequencies of the bottom 50% group in the 1980s, 1990s, 2000s, and 2010s were respectively 1.51, 5.32, 19.13, and 64.63.

Panel 1990 of Figure F8 presents the development of invention keywords created in the 1990s. Similar to Panel 1980, for the top 1%, the average frequencies in the 1990s, 2000s, and 2010s were respectively 200.5, 774.85, and 2,068.12. The average frequencies of the top 1-5% group in the 1990s, 2000s, and 2010s were respectively 37.66, 169.95, and 517.78. The average frequencies of the top 5-10% group in the 1990s, 2000s, and 2010s were respectively 12.2, 62.1, and 182.68. The average frequencies of the top 10-20% group in the 1990s, 2000s, and 2010s were respectively 12.57, 36.67, and 122.13. The average frequencies of the top 20-30% group in the 1990s, 2000s, and 2010s were respectively 8.17, 21.51, and 68.16. The average frequencies of the top 30-40% group in the 1990s, 2000s, and 2010s were respectively 5.41, 14.95, and 50.43. The average frequencies of the top 40-50% group in the 1990s, 2000s, and 2010s were respectively 4, 12.28, and 41.14. The average frequencies of the bottom 50% group in the 1990s, 2000s, and 2010s were respectively 1.67, 8.82, and 35.35.

As shown in Panel 2000 of Figure F8, for the top 1%, the average frequencies in the 2000s and 2010s were respectively 165.42, and 678.4. The average frequencies of the top 1-5% group in the 2000s and 2010s were respectively 39.44 and 170.96. The average frequencies of the top 5-10% group in the 2000s and 2010s were respectively 19.04 and 74.61. The average frequencies of the top 10-20% group in the 2000s and 2010s were respectively 12.57 and 40.34. The average frequencies of the top 20-30% group in the 2000s and 2010s were respectively 8.24 and 27.43. The average frequencies of the top 30-40% group in the 2000s and 2010s were respectively 5.44 and 24.13. The average frequencies of the top 40-50% group in the 2000s and 2010s were respectively 4 and 23.71. The average frequencies of the bottom 50% group in the 2000s and 2010s were respectively 1.86 and 26.78.

Panel 1980 of Figure F9 presents the development of bi-grams invention keywords for each group. It presents the development of invention keywords born in the 1980s. We find that invention keywords with higher initial frequencies tend to become more prominent in the future. For the top 1%, the average frequencies in the 1980s, 1990s, 2000s, and 2010s were respectively 271.46, 906.46, 1,306.07, and 1,953.31. The increase rates of the lower-frequency groups are lower than the top 1% group. The average frequencies of the top 1-5% group in the 1980s, 1990s, 2000s, and 2010s were respectively 50.84, 193.45, 423.58, and 989.58. The average frequencies of the top 5-10% group in the 1980s, 1990s, 2000s, and 2010s were respectively 28.54, 82.42, 176.48, and 318.01. The average frequencies of the top 10-20% group in the 1980s, 1990s, 2000s, and 2010s were respectively 17.58, 48.88, 121.55, and 245.54. The average frequencies of the top 20-30% group in the 1980s, 1990s, 2000s, and 2010s were respectively 12.15, 27.42, 75.15, and 155.56. The average frequencies of the top 30-40% group in the

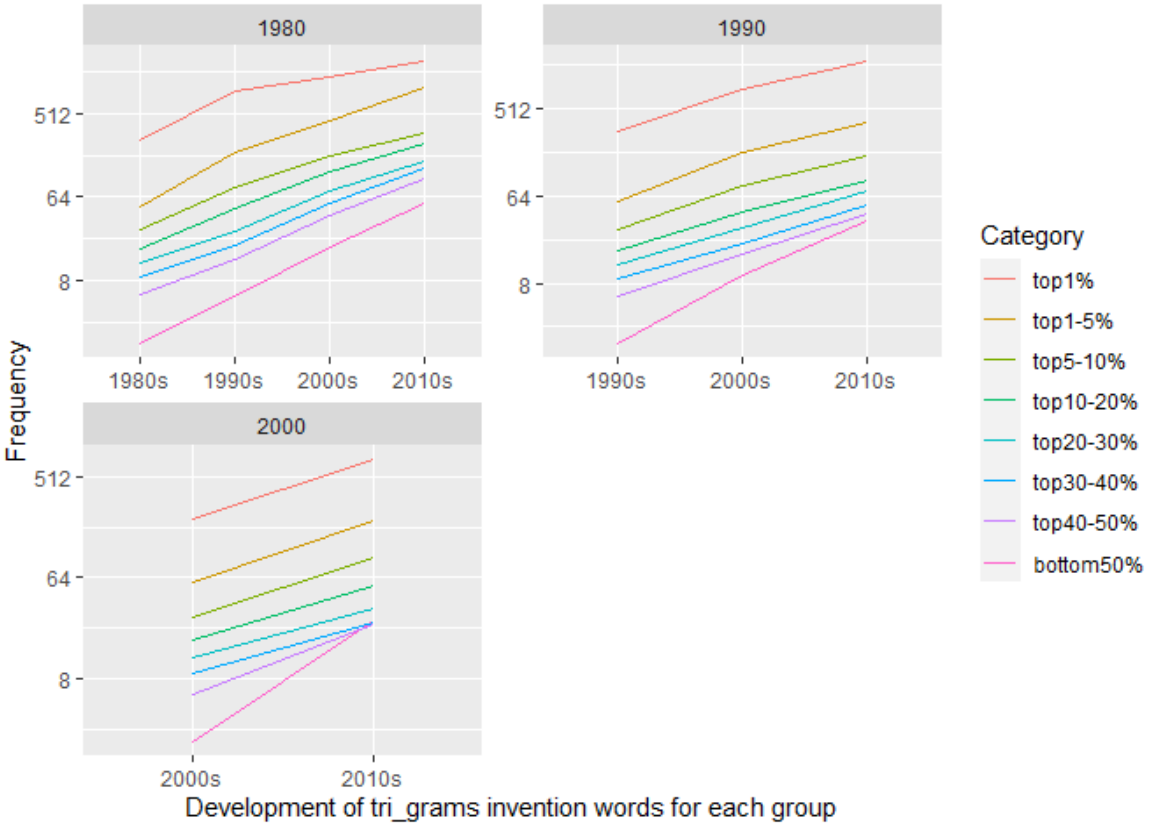


Figure F9: Long-term development of tri-grams Invention Words for Each Group. Panel 1980 presents the development of invention keywords created in the 1980s. Panel 1990 presents the development of invention keywords created in the 1990s. Panel 2000 presents the development of invention keywords created in the 2000s.

1980s, 1990s, 2000s, and 2010s were respectively 8.81, 19.52, 53.75, and 132.94. The average frequencies of the top 40-50% group in the 1980s, 1990s, 2000s, and 2010s were respectively 5.71, 13.59, 39.56, and 99.21. The average frequencies of the bottom 50% group in the 1980s, 1990s, 2000s, and 2010s were respectively 1.64, 5.45, 18.4, and 55.64.

Panel 1990 of Figure F9 presents the development of invention keywords created in the 1990s. Similar to Panel 1980, for the top 1%, the average frequencies in the 1990s, 2000s, and 2010s were respectively 296.09, 816.85, and 1,630.97. The average frequencies of the top 1-5% group in the 1990s, 2000s, and 2010s were respectively 57.16, 184.07, and 367.68. The average frequencies of the top 5-10% group in the 1990s, 2000s, and 2010s were respectively 28.53, 82.62, and 166.66. The average frequencies of the top 10-20% group in the 1990s, 2000s, and 2010s were respectively 17.75, 42.79, and 92.5. The average frequencies of the top 20-30% group in the 1990s, 2000s, and 2010s were respectively 12.29, 29.43, and 73.13. The average frequencies of the top 30-40% group in the 1990s, 2000s, and 2010s were respectively 8.85, 20.43, and 51.01. The average frequencies of the top 40-50% group in the 1990s, 2000s, and 2010s were respectively 5.81, 15.73, and 41.03. The average frequencies of the bottom 50% group in the 1990s, 2000s, and 2010s were respectively 1.89, 9.66, and 34.75.

As shown in Panel 2000 of Figure F9, for the top 1%, the average frequencies in the 2000s and 2010s were respectively 214.7, and 744.84. The average frequencies of the top 1-5% group in the 2000s and 2010s were respectively 58.11 and 206.72. The average frequencies of the top 5-10% group in the 2000s and 2010s were respectively 28.32 and 97.83. The average frequencies of the top 10-20% group in the 2000s and 2010s were respectively 17.7 and 54.13. The average frequencies of the top 20-30% group in the 2000s and 2010s were respectively 12.24 and 34.02. The average frequencies of the top 30-40% group in the 2000s and 2010s were respectively 8.89 and 25.46. The average frequencies of the top 40-50% group in the 2000s and 2010s were respectively 5.85 and 24.66. The average frequencies of the bottom 50% group in the 2000s and 2010s were respectively 2.18 and 26.34.

Overall, these charts indicate that our measurements provide interesting insights into the future development of inventions. The growth of top groups was far superior to bottom groups. In the top 1% group, invention keywords are very likely to continue to grow in the future. However, the growth of bottom groups is the slowest over time.