Analysis of Patent Citation Trends Over Time

Econ 23050: Final Project

Joyce Zhang

I. Abstract

The relationship between firm size, patenting behavior, and innovation speed is a critical area of study, particularly as large firms continue to dominate global markets. In the paper “Barriers to Creative Destruction: Large Firms and Non-Productive Strategies”[1], the argument is made that overtime, the patents firms apply for become increasingly similar as their speed of innovation slows down: “with the help of textual analysis and machine learning tools, patent applications that are too similar to their predecessors could be singled out and their necessity and applicability could be scrutinized” (15). This study explores how the speed and quality of innovation change over time, specifically in the context of patent citations and content similarity. Using large-scale datasets of patent applications and citations, along with advanced textual analysis techniques, we analyze patent citation patterns and CPC subclass similarity to explore speed and quality of innovation over time. We find that patents tend to become more similar as they mature in the market, and the overall impact and novelty of their patents decrease. This study contributes to the literature on creative destruction by highlighting the challenges firms face in maintaining innovation, offering new insights for patent policy, and suggesting ways to improve patent examination practices.

[1] Baslandze, Salome, Barriers to Creative Destruction: Large Firms and Non-Productive Strategies (September 19, 2021).

II. Introduction

A. Research Problem and Objectives

Innovation is widely regarded as the main driver of economic development and technological progress. This research project explores the question: *Does the speed and quality of innovation for firms decline over time?* To address this, the analysis focuses on two key dimensions of patent data—average citations per firm and the diversity of CPC subclass classifications—over a span of more than four decades. Patent citations are commonly used as a proxy for the quality and influence of innovative output. High citation counts suggest that a patent has significantly impacted subsequent technological developments. By examining average citations per firm over time, this study investigates whether patents have become less influential, which may indicate a decline in the quality of innovation. In parallel, the project evaluates the diversity within the Cooperative Patent Classification (CPC) subclasses associated with patents. The CPC system categorizes patents into distinct technological areas. Measuring the diversity of these classifications offers insight into the breadth of innovation. A reduction in CPC subclass diversity over time might reflect a narrowing of technological exploration, potentially signaling a slowdown in the speed or range of innovative activities.

All data for this project is sourced from patentsview.org, encompassing patent records from until 2024, allowing a comprehensive examination of global innovation trends. The study leverages downloadable datasets that include patent details, citation records, and CPC classifications. Statistical analyses and figure generation were performed using Python. This robust computational environment enabled efficient processing of large datasets and supported the application of regression models and similarity metrics to uncover underlying trends in the data. Through a dual analysis of average patent citations and CPC subclass diversity, this paper seeks to contribute to our understanding of long-term innovation trends. The insights gained may have important implications for policymakers and industry leaders interested in the dynamics of technological advancement and economic development.

B. Organization of the Paper

The paper is organized as follows: Section III reviews relevant literature and theoretical frameworks, including insights from Schumpeterian growth theory and studies on patent thickets. Section IV describes the data sources, variables, and the rigorous cleaning and integration procedures used to merge large-scale datasets. Section V outlines our methodology, including descriptive and regression analyses as well as the application of machine learning tools for textual analysis. Section VI presents our empirical results, highlighting temporal trends and regression outcomes. Section VII discusses the findings, their implications for patent policy and firm strategy, and proposes avenues for improved innovation measurement. Section VIII details the study's limitations and offers directions for future research, and Section IX concludes the paper.

III. Literature Review

Recent literature on innovation and business dynamism underscores the critical role of creative destruction—driven by new entrants—in revitalizing technological progress. Akcigit and Ates document a decline in business dynamism over recent decades, attributing this trend partly to reduced knowledge diffusion and a lower influx of innovative entrants[2]. In line with Schumpeterian growth theory, Aghion, Akcigit, and Howitt argue that the process of creative destruction relies on new firms entering the market to challenge established incumbents, thereby maintaining competitive pressures and spurring innovation[3]. However, barriers such as the non-productive strategies employed by large firms—ranging from leveraging political connections to engaging in non-productive patenting—can inhibit this essential influx of entrants, stifling the disruptive process that underpins long-term growth[4].

[2] Ufuk Akcigit & Sina T. Ates, 2021. "Ten Facts on Declining Business Dynamism and Lessons from Endogenous Growth Theory," American Economic Journal: Macroeconomics, vol 13(1), pages 257-298.

[3] Philippe Aghion*, Ufuk Akcigit, Peter Howitt, 2014. "What Do We Learn From Schumpeterian Growth Theory?" Handbook of Economic Growth Volume 2, pages 515–563.

[4] Baslandze, Salome, Barriers to Creative Destruction: Large Firms and Non-Productive Strategies (September 19, 2021).

Complementing these insights, research linking innovation to income distribution has highlighted how disruptive innovations contribute to top income inequality. Aghion et al. demonstrate that innovation, particularly from new entrants, can lead to disproportionate rewards that elevate the top income share[5]. Moreover, recent work on measuring novelty by Foster, Shi, and Evans proposes that citation data—especially when patents are cited by works classified under non-similar CPC subclasses—serves as a powerful proxy for disruption and relevance[6]. In this framework, a patent that garners citations from a diverse array of CPC subclasses is more likely to signal a novel breakthrough that challenges existing technological paradigms, thereby reinforcing its disruptive impact. Together, these strands of research emphasize that fostering entrant-driven creative destruction and accurately capturing innovation disruption are pivotal for understanding both technological progress and its socioeconomic repercussions.

The above literature is highly relevant to my research project as it provides a robust theoretical foundation for examining whether the speed and quality of innovation decline over time, specifically through the analysis of average citations and CPC subclass diversity.

IV. Data Description and Processing

The analysis draws on four sets of patent data sourced from patentsview.org. The Patent Assignee Data includes 8 million rows detailing each patent and its corresponding assignee (organization) and its assignee type (company, university, government, etc). Note that despite there being other types of assignees, I am using the word "firm" in this paper because they account for the majority of patents produced (Figure 1). Complementing this is the Granted Patents dataset, which offers 9 million rows each patent and the date it was granted, allowing for temporal analysis of innovation trends.

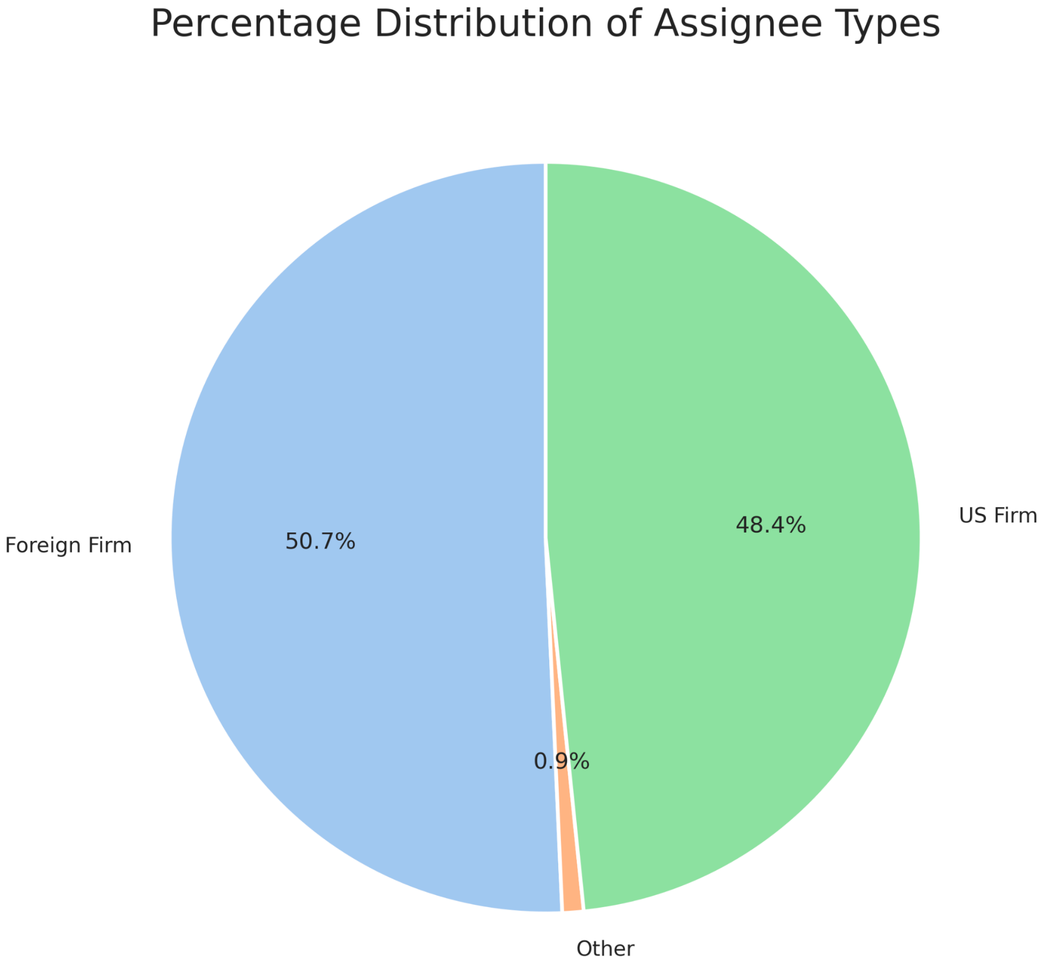
[5] Philippe Aghion, Ufuk Akcigit, Antonin Bergeaud, Richard Blundell, David Hemous, Innovation and Top Income Inequality, The Review of Economic Studies, Volume 86, Issue 1, January 2019, Pages 1–45.

[6] Foster, J. G., Shi, F., & Evans, J. (2021, April 13). Surprise! Measuring Novelty as Expectation Violation.

Table 1

Dataset	Patent Grants	Patent Assignee	CPC Class	Citations
Columns Used	Patent_id Patent_date	Patent_id Assignee	Patent_id CPC_Subclass	Patent_id Citation_id Citation_Date
Rows	9 Million	8 million	55 million	145 million

Figure 1



Additionally, the CPC Classification Data, originally comprising 55 million rows, is refined down to 8 million rows (selecting for CPC sequence = 0) containing each patent and its corresponding CPC sections and subclass. I used the second most specific CPC label (subclass) as it was sufficiently diverse without being overly specific. This dataset is crucial for examining the technological focus and diversity within the patents. Lastly, the Patent Citations dataset is expansive, with 145 million rows featuring each patent, each patent it cites, and the date of the cited patent. It serves as the critical dataset for calculating patent citation counts. Note that all the data is limited to post 1973 patents so that the data is relatively consistent. Pre-1973 data points are sparse and interfere with the patterns in the data, particularly those from over a century ago.

Significant efforts were directed towards data cleaning and integration. Each dataset underwent thorough cleaning to correct inconsistencies and standardize key variables across all sources. This facilitated the merging of datasets into a comprehensive framework where each patent's assignee, grant date, classification, and citation history could be analyzed collectively. Handling such large datasets presented inherent technical challenges, notably with memory and computational limitations. To mitigate the risk of kernel crashes and to manage resource constraints, data processing was executed in manageable chunks. This approach, alongside the use of efficient Python programming practices and optimized libraries, ensured that even the most voluminous datasets, like Patent Citations, were processed in a stable and efficient manner. Despite these measures, challenges remained in merging disparate sources and ensuring consistent data quality, necessitating iterative refinements and validation processes to maintain the integrity of the integrated dataset for subsequent analysis.

V. Methodology

The methodology for this project combines descriptive and regression analyses to explore patent citation trends and CPC subclass similarity across firms. Descriptive analysis involves aggregating patent data by merging datasets on patent IDs and extracting application years to compute annual averages per patent. Patents are grouped by firm and year, and cumulative counts are tracked over time. This approach allows us to calculate average citations per patent and CPC similarity scores on an annual basis, thereby establishing baseline trends while accounting for variations in patent volume across firms.

Regression Analysis employs panel data models with dependent variables representing either average citations or CPC similarity scores. Key predictors include linear and quadratic time variables, top firm dummies (reflecting the Pareto distribution of innovation outputs), and interaction terms (e.g., year × top firm). Standardization of variables minimizes multicollinearity, thereby stabilizing coefficient estimates. Furthermore, our models evaluate the influence of firm stature on innovation quality, even though low R^2 values in some models indicate that unobserved factors—such as R&D specialization or industry-specific dynamics—may also play significant roles.

Technological diversity is quantified using Word2Vec embeddings trained on CPC subclass co-occurrence data (window size = 5, vector size = 100). For each firm, the dominant CPC subclass is compared with other patents' subclasses via cosine similarity. A patent that is cited by patents from diverse CPC subclasses is indicative of disruptive potential, whereas high similarity suggests incremental innovation. Data wrangling and visualization were managed using Python's pandas, numpy, and Seaborn libraries, while regression analysis was conducted with statsmodels and textual embeddings generated via gensim. All the code for the analysis and visualizations are in the submitted notebook file.

VI. Empirical Results

A. Descriptive Trends

The total number of patents granted is increasing over the years (Figure 2). Despite the growing number of patents, however, the average citations each year per patent per firm reveals a different temporal pattern (Figure 3). From 1977 to 2010, citations per patent remained relatively stable, fluctuating between 2.5 and 3.0 (Figure 2). A notable spike occurred in the late 1970s, potentially linked to the 1976 U.S. Copyright Act or increased R&D investments during economic shifts. Post-2010, a steady decline emerged, dropping to approximately 1.5 citations per patent by 2024. This decline aligns with broader concerns about diminishing patent quality and "patent thickets" crowding out impactful innovations.

CPC subclasses were mapped into semantic space using Word2Vec. Figure 4 is a visualization of the top 100 most popular subclasses. The size of the text on the graph is based on the frequency of the subclass code in the dataset. Using the model of mapped distances, CPC subclass cosine similarity score between each patent and its assignee's top patent subclass was calculated. Then the average patent similarity score for each year was calculated as a measure of technological focus, with average similarity ranging from 0.64 to 0.74 over the decades (Figure 5). The trend reveals that firms become increasingly focused with their core IP. Additionally, average similarity between patents and their citations (Figure 6) rose slightly—from 0.68 to 0.72—indicating a potential decrease in disruptive innovation over time.

Figure 2

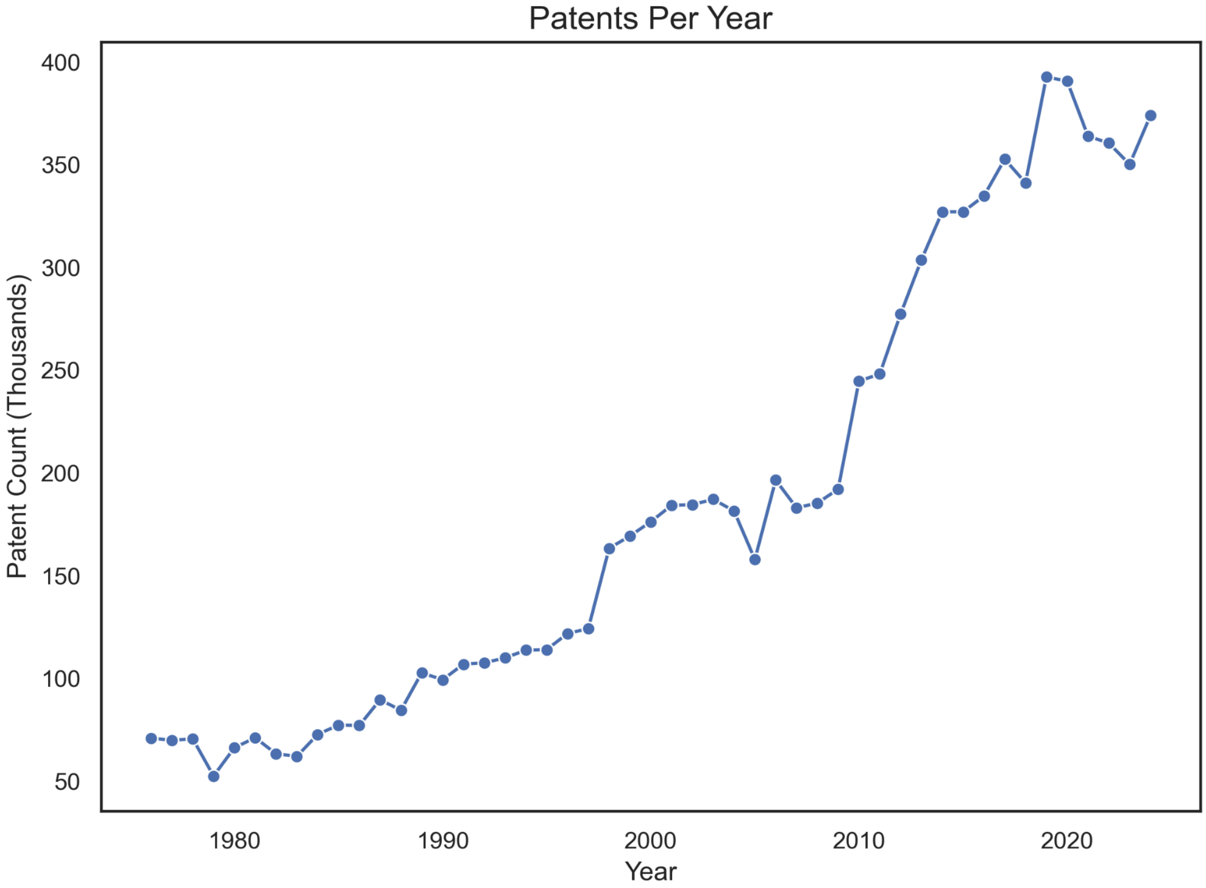

Patents Per Year
Patent Count (Thousands)
400
350
300
250
200
150
100
50
1980
1990
2000
2010
2020
Year

Figure 3

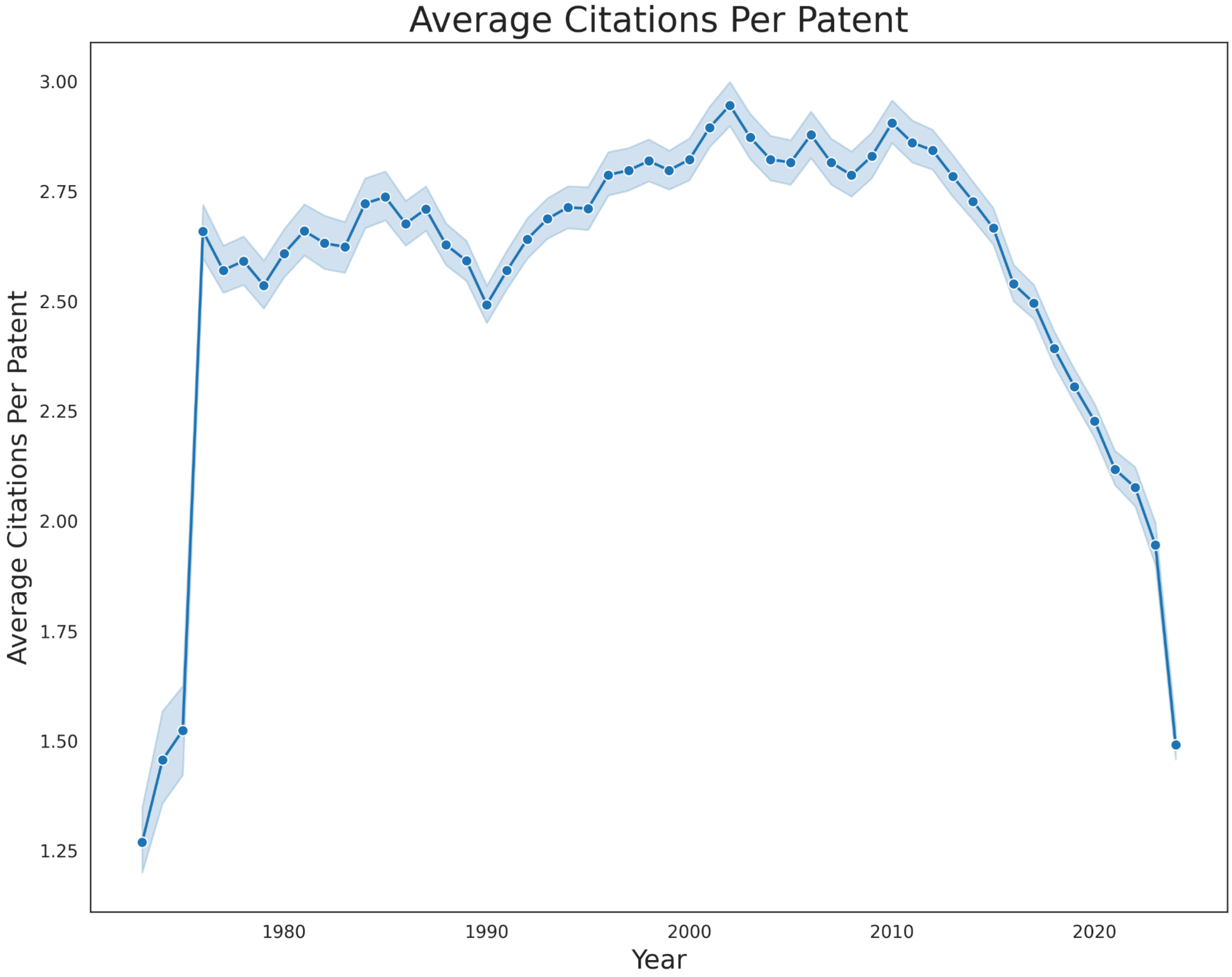
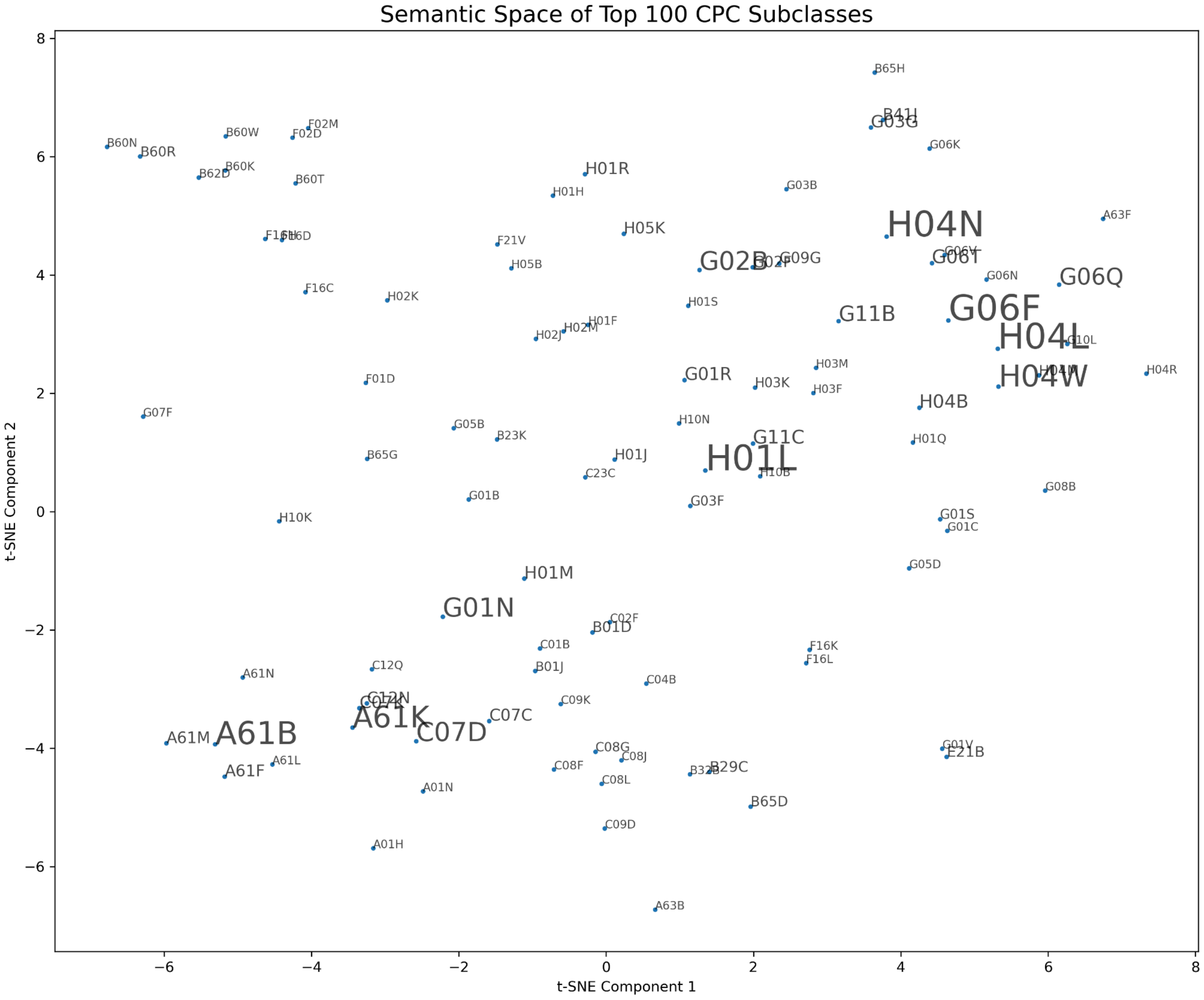

Average Citations Per Patent
Average Citations Per Patent
3.00
2.75
2.50
2.25
2.00
1.75
1.50
1.25
1980
1990
2000
2010
2020
Year

Figure 4



Semantic Space of Top 100 CPC Subclasses
t-SNE Component 1
t-SNE Component 2
8
6
4
2
0
−2
−4
−6
−6
−4
−2
0
2
4
6
8
B65H
B41J
G03G
G06K
F02M
F02D
B60W
B60N
B60R
B60K
B62D
B60T
H01R
H01H
G03B
A63F
H04N
F16H
F16D
F21V
H05K
G06V
G06T
H05B
G02B
G02F
G09G
G06N
G06Q
F16C
H02K
H01S
G11B
G06F
H04L
G10L
H01F
H02M
H02J
H03M
H04M
H04R
F01D
G01R
H03K
H03F
H04W
H04B
G07F
H10N
G05B
B23K
G11C
H01Q
B65G
H01J
H01L
H10B
C23C
G08B
G01B
G03F
H10K
G01S
G01C
G05D
H01M
G01N
C02F
B01D
F16K
C01B
F16L
C12Q
B01J
A61N
C04B
C12N
C07K
C09K
C07C
A61K
A61B
C07D
A61M
G01V
E21B
C08G
C08J
A61L
B29C
B32B
C08F
A61F
C08L
A01N
B65D
C09D
A01H
A63B

CPC Section Descriptions:

CPC Section	**Description**
A	Human necessities
B	Performing operations; transporting
C	Chemistry; metallurgy
D	Textiles; paper
E	Fixed constructions
F	Mechanical engineering; lighting; heating; weapons; blasting engines or pumps
G	Physics
H	Electricity

Figure 5

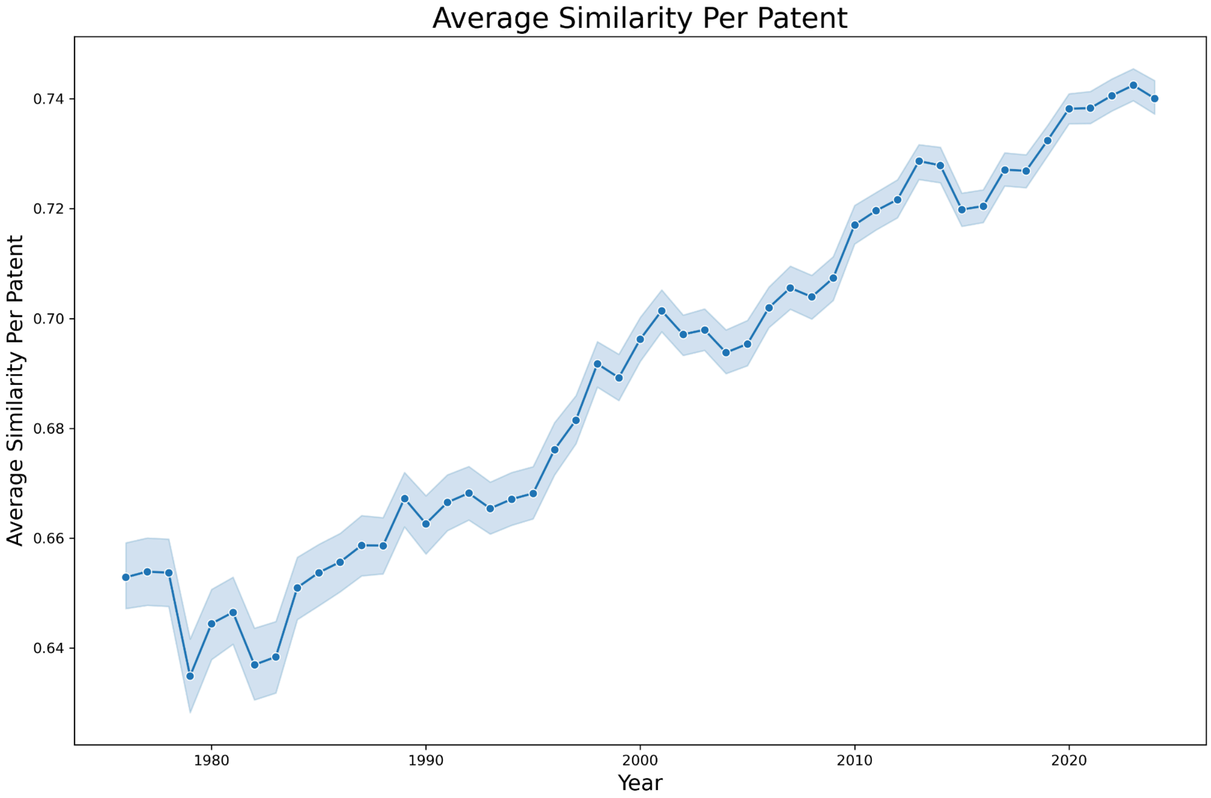
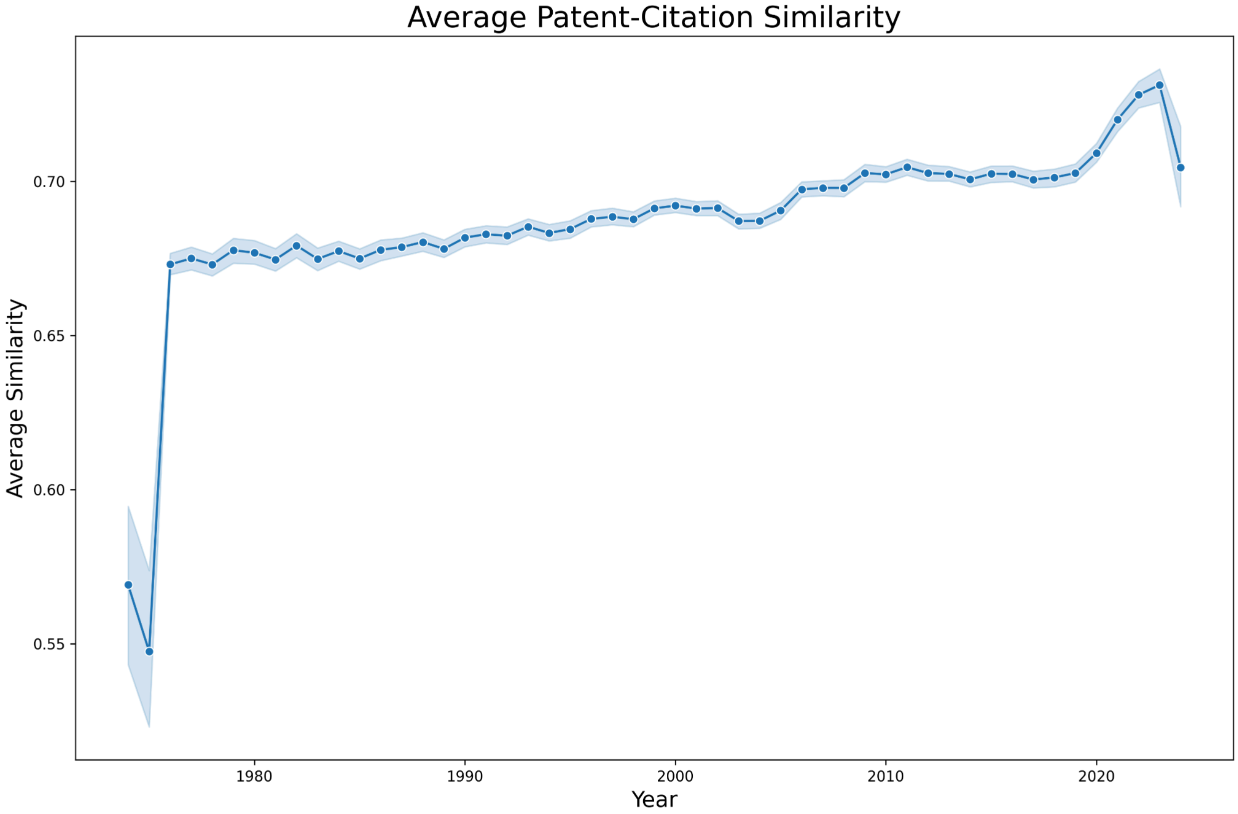


Figure 6



B. Regression Analysis

The citation regression model highlights significant relationships between time, firm, and citation count (Table 2). Top firm dummy has a positive coefficient (37.27, $p < 0.001$) indicating that being a top firm means receiving significantly more citations on average, likely due to their market dominance. However, standardized year (−0.16, $p < 0.001$) and year-squared (−0.17, $p < 0.001$) coefficients are both negative, meaning for non top firms there is a non-linear decline in average citations received over time. There is a positive interaction coefficient between top firm and year but negative coefficient between top firm and year squared, suggesting top firms initially resist the decline but face sharper long-term drops.

For the top CPC similarity model (Table 3), top firm dummy has a negative coefficient (−0.06, $p < 0.001$), implying that top firms' patents are less similar to their dominant CPC subclass, signaling broader innovation efforts. A slight positive trend (0.03, $p < 0.001$) for the year variable suggests gradual homogenization across all firms, though top companies amplify this marginally (dummy_comp:year_std = 0.0035, $p < 0.001$). The low R^2 value in the model underscores unobserved factors, such as R&D specialization or sector-specific dynamics. The first regression model, while stronger, still leaves 64.5% of variance unexplained, pointing to omitted variables like patent litigation strategies or geopolitical influences.

Table 2

Variable	Coef	Std. Err	t-value	P>|t|	[0.025	0.975]	Metric	Value
Intercept	2.6643	0.004	650.769	0	2.656	2.672	No. Observations	1150157
dummy_comp	37.2753	0.056	664.845	0	37.165	37.385	Df Model	5
year_std	-0.1576	0.003	-50.932	0	-0.164	-0.152	R-squared	0.355
dummy_comp:year_std	0.4037	0.041	9.933	0	0.324	0.483	Adj. R-squared	0.355
year_sq_std	-0.1721	0.003	-59.138	0	-0.178	-0.166	F-statistic	3978
dummy_comp:year_sq_std	-5.5604	0.034	-163.892	0	-5.627	-5.494	Prob (F-statistic)	0

Table 3

Variable	Coef	Std. Err	t-value	P>|t|	[0.025	0.975]	Metric	Value
Intercept	0.7144	0	2156.499	0	0.714	0.715	No. Observations	830038
dummy_comp	-0.0601	0.001	-86.203	0	-0.061	-0.059	Adj. R-squared	0.022
year_std	0.0292	0	88.692	0	0.029	0.03	F-statistic	6264
dummy_comp:year_std	0.0035	0.001	5.013	0	0.002	0.005	Prob (F-statistic)	0

Table 4

Variable	Coef	Std. Err	t-value	P>|t|	[0.025	0.975]	Metric	Value
Intercept	0.6826	0	2885.439	0	0.682	0.683	No. Observations	923480
dummy_comp	0.049	0.001	91.901	0	0.048	0.05	Adj. R-squared	0.013
year_std	0.0083	0	35.129	0	0.008	0.009	F-statistic	3978
dummy_comp :year_std	0.0096	0.001	18.184	0	0.009	0.011	Prob (F-statistic)	0

For citation CPC similarity (Table 4), top firm dummy has a positive coefficient (0.05, $p < 0.001$), indicating that top firms' patents receive citations with more similar CPC subclass, affirming our hypothesis and previous literature that patents become *decreasingly* innovative over time. This is perhaps also due to an increase in self citation practices by firms. A slight positive trend (0.008, $p < 0.001$) for the year variable suggests a very slight increase in similarity across citations, with top companies amplifying this effect marginally.

VII. Findings and Implications

A. Interpretation of Findings

The steady decline in average citations per patent after 2010 and steady increase patent-citation similarity aligns with concerns about "patent thickets"—overlapping patents that stifle follow-on innovation—and non-productive patenting by large firms to block competitors rather than advance technology. This trend mirrors previous observations of slowing knowledge diffusion, suggesting that incumbents may prioritize defensive strategies over disruptive R&D. The sharper decline among top firms (dummy_comp:year_sq_std = −5.56) supports the hypothesis that scaling R&D efforts yields diminishing returns, as larger portfolios risk redundancies and incrementalism.

While top companies initially receive more citations (coefficient = 2.66), their long-term decline in citation impact underscores a “creative accumulation trap.” Large firms may leverage existing technological dominance to attract early citations but struggle to sustain breakthrough innovation due to bureaucratic inertia or risk aversion. This aligns with Schumpeter’s later work, which cautioned that monopolistic firms might stagnate without competitive pressure.

Top firms’ lower CPC similarity scores (−0.06) suggest strategic diversification to hedge against disruption. By spreading innovations across subclasses, incumbents may potentially block entrants’ entry points while exploring adjacent markets. However, the slight homogenization trend over time implies broader industry-wide convergence. The positive correlation between top firm and citation similarity (coefficient = 0.05) similarity demonstrates the challenges with continuous high quality innovation over time.

The minimal explanatory power of CPC similarity models ($R^2 = 0.022$) highlights potential unobserved factors such as sector-specific R&D cultures or internal knowledge silos. For instance, firms may specialize in niche technologies not fully captured by CPC subclasses, or cross-industry collaborations could blur classification boundaries.

B. Implications for Policy and Firm Practice

The findings validate calls for the USPTO to prioritize patent quality scrutiny. AI tools, as proposed in the literature, could flag low-novelty patents by analyzing the text itself and/or CPC similarity and citation patterns. Additionally, policies encouraging open innovation (e.g., patent pools) might mitigate thickets while preserving incumbents’ incentives to invest in R&D. In addition, the coexistence of top firms’ CPC diversity and citation decline raises a paradox: Are these firms truly innovating broadly, or are they fragmenting innovations to game the patent system? Further text analysis of patents could disentangle whether diversity reflects genuine exploration or strategic obfuscation.

Firms should strike a balance between technological diversification and specialization to sustain innovation quality. While top companies exhibit broader average CPC subclass diversity—suggesting strategic patenting across domains—their long-term decline in citation impact signals diminishing returns from scaled R&D. To avoid overextension, firms should prioritize core technological IPs (e.g., investing in foundational innovations like AI hardware or biotech platforms) while selectively diversifying into adjacent fields to capture more emerging markets. For example, a semiconductor firm might anchor R&D in advanced chip design while exploring quantum computing applications. Regular portfolio audits can prune low-impact patents, freeing resources for high-value filings.

Collaboration and adaptability are equally critical. Engaging in open innovation—via partnerships with startups, academia, or patent pools—can inject fresh ideas and mitigate the risks of patent thickets. Proactive regulatory engagement, such as advocating for clearer USPTO novelty standards, ensures compliance while shaping policies that reward breakthrough innovations. Additionally, firms should leverage acquisitions or spin-offs to integrate disruptive technologies (e.g., acquiring a robotics startup to enhance automation capabilities). By aligning IP strategies with agile R&D practices and market trends, firms can navigate the tension between scale and innovation, transforming patent portfolios from defensive shields into engines of sustained growth.

VIII. Limitations and Future Research

Despite offering robust insights into patent trends and innovation dynamics, this study is not without limitations. The massive datasets required processing in chunks, which introduced challenges in merging disparate sources and ensuring consistent data quality. In addition, our regression models and similarity metrics do not fully capture certain nuances such as self-citations or within-firm citation dynamics. These limitations, combined with potential biases in data cleaning and integration, suggest that the current analysis may not account for all factors influencing the speed and quality of innovation over time.

Another limitation stems from my current knowledge and experience in Natural Language Processing (NLP) and the use of Word2Vec for generating CPC embeddings. While the visualization outputs and similarity metrics indicate that the current approach is fairly adequate for capturing trends in technological focus, there remains uncertainty regarding the optimal tuning of parameters such as window size, vector dimensions, and alternative embedding models that might improve the quality of the similarity analysis. Further exploration and advanced expertise in NLP could potentially enhance these measures; however, based on the present visualizations and consistency of the results, the analysis provides a reliable foundation.

Building on these findings, future studies should also explore the influence of external factors—such as geopolitical shifts, regulatory changes, and patent litigation strategies—on innovation trends. Longitudinal studies across diverse industries and regions (e.g US vs. foreign firms) could elucidate how evolving market conditions affect both patent quality and citation dynamics. Sensitivity analyses to assess the robustness of the observed trends under different data cleaning protocols and threshold settings would further solidify the empirical conclusions. It would also be interesting to investigate how mergers and acquisitions reshape innovation trajectories. For example, does acquiring startups boost patent diversity, or does integration dilute disruptive potential? Such extensions would provide a more comprehensive understanding of the mechanisms driving creative destruction and innovation decline.

IX. Conclusion

This study reveals that while top firms initially dominate patent citations, their long-term innovation impact wanes as portfolios grow more diverse yet less disruptive. The decline in post-2010 citations and modest CPC similarity trends underscore systemic risks of non-productive patenting and R&D inertia. This trend suggests that defensive patenting strategies may lead to a gradual erosion of breakthrough innovation, with patents increasingly converging in content and impact. The findings underscore concerns over patent thickets and reduced knowledge diffusion, emphasizing the challenges firms face in sustaining disruptive innovation.

For policymakers, the results validate calls for AI-enhanced patent scrutiny to prioritize novel applications, while firms must rebalance portfolio breadth with deep-tech investments. Future research, particularly text-based novelty scoring and firm life cycle analysis, could further decode innovation's "black box." As patent ecosystems evolve, bridging empirical rigor with strategic adaptability will remain key to fostering economies where scale and disruption coexist.

Ultimately, this study attempts to contribute to our understanding of innovation dynamics in a rapidly evolving technological landscape. The insights gained have significant implications for patent policy and firm strategy, suggesting a need for more rigorous patent examination practices and the promotion of open innovation frameworks. Future research that refines these analytical methodologies will be essential in addressing the complex interplay between firm behavior, patenting practices, and the long-term trajectory of technological innovation.