Analyzing Patent Data

Alessia Berarducci & Federica Marini

Analysis of Social Networks

Project Report 1

December 19, 2024

Contents

1 Introduction

Patents are complex legal documents containing a large amount of information, including details on patent citations, inventors, and technical descriptions. This project will focus on analyzing the inventor network within the United States Patent and Trademark Office (USPTO) dataset.

Beyond the co-inventorship network, the analysis also explores bipartite graphs that link inventors with the International Classification categories of the patents they have submitted.

The study begins with an examination of the inventor-technological class network using the EventNet framework to calculate a range of statistics that capture its structure and evolution. These statistics explain how inventors are connected to each other and to specific technological fields, highlighting patterns of collaboration. To further analyze these dynamics, a Cox proportional hazards model (CoxPH)[1] is applied, utilizing the statistics to explore how the network evolves over time.

The project also adopts a coevolutionary perspective by examining two interconnected layers of the network: the co-inventorship network and the network that links inventors to technological classes. This dual-layered approach captures the interactions between social relationships among inventors and their technological engagements. Using EventNet[2] statistics are computed to track changes in the network over time: these reveal how shifts in co-inventorship relationships and technological associations influence one another. Lastly, this coevolution model is analyzed using the CoxPH model, which provides insights into the ways in which social and technological factors interact to influence the larger patterns of innovation. All our analyses are publicly available on GitHub at `https://github.com/aleberas/SNA_Analyzing_Patent_Data`.

2 Research Question

This study investigates first the structure of the co-inventor network, focusing on how inventors collaborate and connect. Key questions include:

- How are inventors connected within the co-inventor network?
- Are there isolated inventors or distinct sub-communities?

By addressing these questions, the research aims to uncover the overall connectivity of the network and highlight the existence of collaborative hubs or more independent inventors.

[1] The Cox proportional hazards model is a regression model used in survival analysis to assess the impact of explanatory variables on the hazard or event rate.

[2] `https://github.com/juergenlerner/eventnet`

The relationship between inventors and the International Classification of Patents classes is also a focus of this study. Specifically, the research explores the level of interdisciplinarity in the network composed of inventors and the ICL classes associated with their patents. The following questions guide this investigation:

- At what level of granularity is interdisciplinarity most visible?
- Which ICL classes are central to the network?
- Which areas exhibit the highest levels of innovation activity?

The latter study seeks to understand the technological domains that contribute to collaboration and innovation, as well as the role of interdisciplinarity in shaping the network.

3 Data Description and Data Preprocessing

The dataset, spanning from 1976 to 1985[3], contains approximately 500.000 patents. Given the vast amount of patent data submitted weekly and the consequent quantity of information, the first step of the analysis involves reducing the dataset's complexity by focusing on specific columns and removing potential errors, such as missing values.

For each year within the period under study, the corresponding CSV file of patent data has been loaded: to ensure data integrity, the types of the columns were explicitly defined.

The dataset exhibited numerous issues, some of them due to the fact that data were collected across different years, and so methods and standards used to gather information had evolved over time. These changes lead to inconsistencies.

In the data cleaning process, several important adjustments were made to ensure data consistency. Through this process, it became evident that some names were misspelled. This inconsistency can lead to issues, as even minor variations in names result in treating the same individual as two separate entities. To address that problem, an algorithm based on the Levenshtein Distance[4], a string metric that quantifies the difference between two strings, has been implemented. Our function treats one unit of difference between two strings as significant, except when the difference involves a single letter that could represent a middle name or an abbreviation; in such cases, it does not count as a unit of distance, as forcing such differences to be treated as significant is unreasonable. For instance, *michael k hughes* and *michael p hughes* are two different inventors and we don't want to standardize them under a single name. Additionally, we performed a double check before

[3] The years correspond to the issue dates of the patents.

[4] `https://en.wikipedia.org/wiki/Levenshtein_distance`

applying standardization: we ensured that the two WKUs of the patent inventors were different, as an inventor can be associated with a patent only once.

The cleaned dataset contains eight variables, as showed in table 1.

Variable	**Description**
WKU	Patent ID.
Title	The name or description of the invention.
App_Date	The date on which the patent application was officially filed with the patent office.
Issue_Date	The date on which the patent was officially granted by the patent office.
Inventor	The individual (one or more) who contributed to the conception of the invention described in the patent.
Assignee	The entity (individual, organization, or company) that owns the rights to the patent.
ICL_Class	The International Classification (ICL) Code that categorizes the patent according to the nature and field of the invention, based on the IPC system.
References	The citations made by the patent to prior patents or literature, including patents, academic papers, or other publications.

Table 1: Description of the variables in the dataset.

3.1 Co-Inventor Network

One of our purpose is to observe the co-inventor network focusing on inventors who collaborate with different inventors on the same patent. The nodes are the inventors and an edge between two inventors exists if they submitted a patent together.

3.1.1 Statistics

The total number of nodes is 381442 and the number of edges 954224. The average degree of the network, calculated as the mean number of connections per inventor, is approximately 5.21. This indicates that, on average, each inventor collaborated with about five other inventors.

The density of the network, which measures the proportion of possible connections that are actually present, is 1.367283e-05. A density value close to 0 suggests a sparse network, as seen here. Similar to many real-world networks, the degree distribution exhibits a long tail, indicating the presence of a few highly connected nodes that serve as hubs within the network.

The global clustering coefficient, which reflects the level of interconnectedness in the network, was measured at 0.7184537. This high value, close to 1, signifies that inventors who collaborated with a common individual are also likely to collaborate with each other. Additionally, the network contains a significant number of isolated nodes—83,304 to be precise. These nodes, representing inventors with a degree of 0, indicate individuals who did not collaborate with any other inventor.

That network reflects all the characteristics of an empirical network, in which there are a lot of isolated nodes, a lot of reciprocal and triadic structures, the degree distribution has a long tail, the centralization is high and the distances between nodes is high.

To identify the communities present in the network, the Louvain Algorithm[5] has been applied. This algorithm identifies groups of nodes that are more densely connected internally than with the rest of the network. The total number of communities detected is 134691 and that large number indicates that many inventors work in small teams or alone. In fact, 61.85% of them are singletons, meaning that many work independently. The largest community consists of 4,815 nodes and 28,208 edges, demonstrating a high level of interconnectedness within this group. That network reflects all the characteristics of an empirical network, in which there are a lot of isolated nodes, a lot of reciprocal and triadic structures, the degree distribution has a long tail, the centralization is high and the distances between nodes is high.

4 Bipartite Graphs of Inventors Through the Lens of ICL Classification

The aim is to create and study a bipartite graph that represents the relationships between inventors and the International Classification of Patents associated with their patents.

The network is bipartite, meaning that an edge only exist between nodes of different sets and not within the same set, hence there will be edges only between inventors and ICL classes and not within inventors or within the technological classes. An edge connects an inventor node to an ICL class node if the inventor contributed to a patent classified under that technological class. The network is considered directed as the relationship between the two sets of nodes implies a specific direction of association: the inventor contributes to a patent, which is subsequently classified under an ICL class.

In the first two configurations, we used Relational Event Models (REM). A REM captures pairwise interactions between nodes over time, where one source node (e.g., an inventor) connects to one target node (e.g., an ICL class) in a one-to-one relationship.

In the third configuration, we expanded the model to a Relational Hyperevent Model (RHEM), where a single source node (e.g., an inventor) connects simultaneously to multiple target nodes (e.g., several ICL classes). This approach accounts for events involving multiple associations occurring at the same time.

[5] https://en.wikipedia.org/wiki/Louvain_method

4.1 First configuration

In the initial analysis, only the first letter of each ICL Class was considered to begin at a lower level of granularity. An inventor, associated with a single WKU, submits on average in 1.37 different technological classes. The heatmap, figure 1, visualizes the relationships between pairs of ICL classes, where the intensity of the color represents the frequency of co-occurrences between two classes.

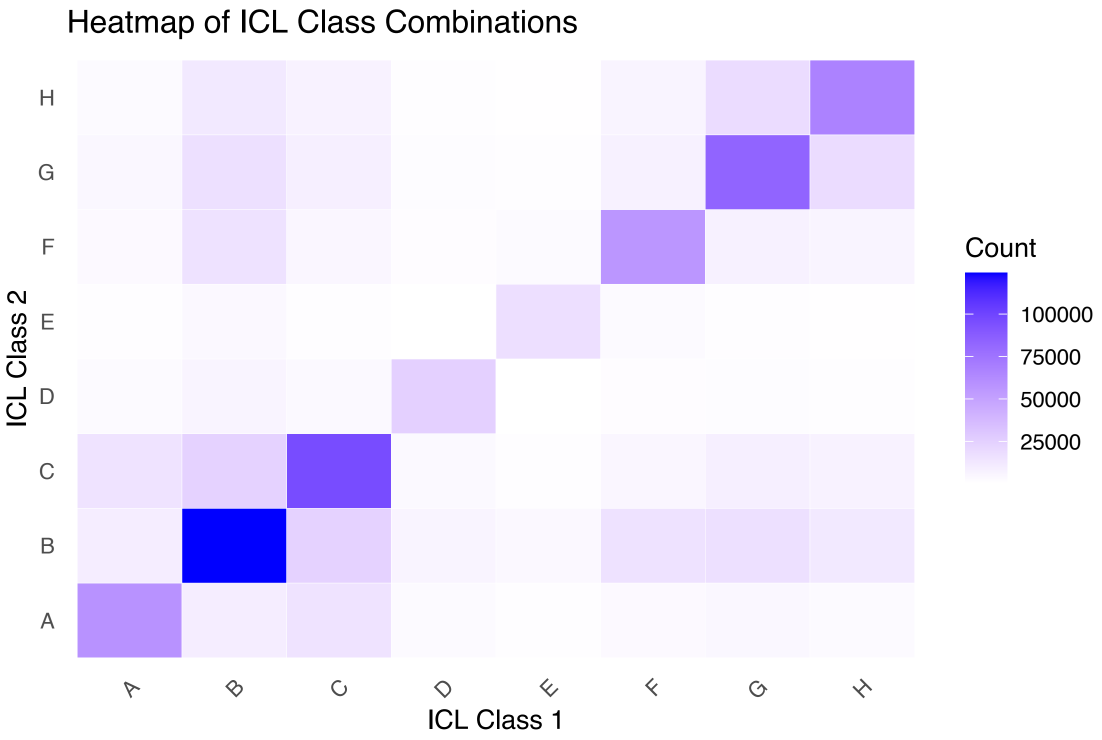

Figure 1: Heatmap of technological classes combinations

The strongest color intensity is along the diagonal, suggesting that most patents are concentrated within the same ICL classes. The most co-occurrent off diagonal matches are B (Performing Operations, Transporting) with C (Chemistry, Metallurgy) and G (Physics) with H (Electricity), with 24196 and 19209 occurrences.

The bubble plot, figure 2, visualizes ICL Class submissions over time[6] with distinct bubble sizes representing submission counts. We can notice a consistent prominence of certain classes (e.g., Type C) and the peak in submissions around 1980. The peak may be due to increased innovation in the late 1970s and early 1980s.

The network is a two mode network where the nodes consist of two types: inventors and technological classes at level of granularity 1. The edges between these nodes indicate that an inventor has

[6] The year on the y-axis of the bubble plot refers to the application date, which is more dispersed compared to the issue date.

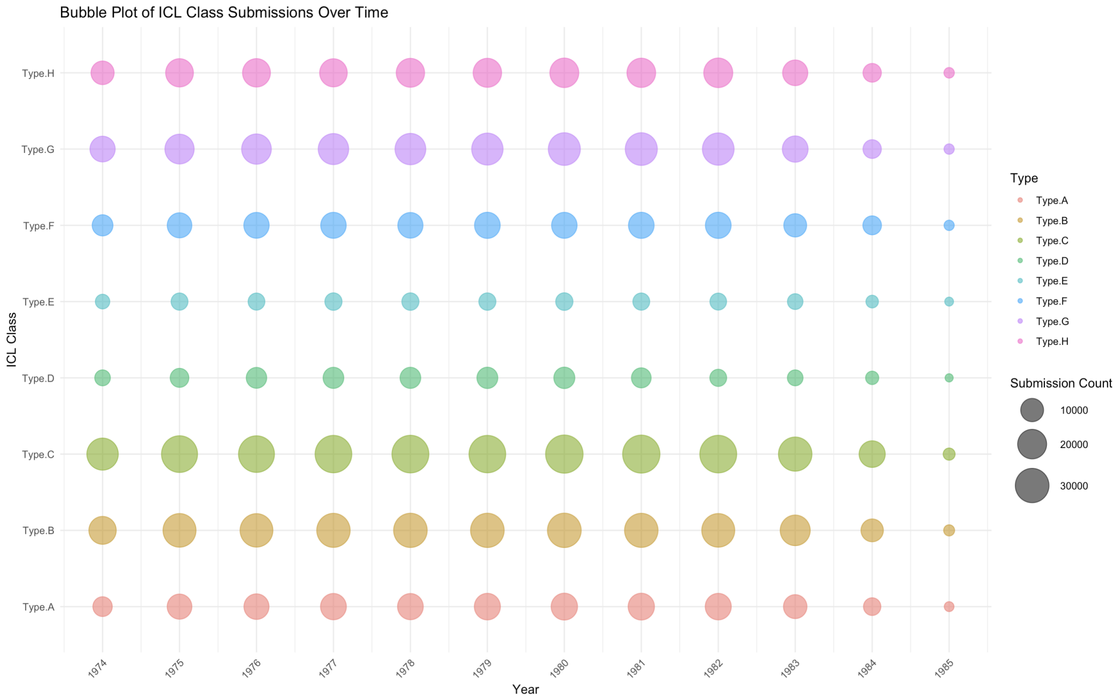

Figure 2: Trends in ICL Class Submissions (1974–1985)

submitted a patent within the corresponding technological class. The temporal aspect of this analysis incorporates the application date of each patent.

To analyze this network, a Relational Event Model has been used. In this framework, edges represent directed, one-to-one connections between a source node and a target node, with no loops permitted, ensuring that a node cannot connect to itself within the event structure.

The analysis has been carried out using the Event Network Analyzer (EventNet) that is a specialized software designed by Jürgen Lerner for the analysis of relational events and hyperevents. The setting has been configured to align with the desired model, defining the source, target, date, and the event ID as the WKU. By construction, the source size and target size are always one, ensuring a strict one-to-one relationship between nodes, which reflects the behavior of the Relational Event Model (REM). Additionally, no loops within the same set of node are permitted.

Different statistics have been calculated using Eventnet:

- **avg.patents.per.inventor**: average activity level of inventors across all technological classes.
- **avg.submissions.per.class**: average number of submissions per ICL class.

The output from EventNet is used to fit a Cox proportional hazards model in R, using the library *Survival*. In fitting the model, the square root transformation was applied to the statistics. This is necessary because the variables represent aggregated measures derived from EventNet, which are often on a larger scale than the original data. The square root transformation is used to rescale these variables closer to the scale of the raw data.

4.1.1 Interpretation of the first configuration

The number of observations in that network is 2887806 and the number of events 1443903. The model assesses the hazard[7] of patent submission in a general way, considering the average activity level of inventors and the average number of submissions per ICL class. This approach is useful for understanding the broad dynamics of the network but does not explain how different types of ICL classes may influence these patterns. Main results are presented in figure 2, where the hazard ratio is expressed as exp(coef), with the corresponding coefficient (coef) shown in parentheses. All the coefficients in the model were found to be statistically significant.

Covariate	**Hazard Ratio (HR)**	**Interpretation**
avg.patents.per.inventor	0.148 (-1.908)***	Hazard decreases by 85% per unit increase
avg.submissions.per.class	1.530 (0.425)***	Hazard increases by 53% per unit increase

Table 2: General Cox proportional hazard model

Note: *** indicates $p < 0.001$.

Inventors who submitted on average in a higher number of ICL classes have a lower likelihood of submitting additional patents over time: for every unit increase in the average number of patents per inventor, the hazard decreases by 85%.
Inventors who submit patents in the most popular ICL classes, are much more likely to file additional patents. The hazard ratio indicates an increase in the likelihood of submitting another patent for each unit increase in this statistic. Specifically, for every unit increase in the average number of submissions per class, the hazard of submitting new patents increases by 53%.

4.2 Second configuration

In this second configuration, the concept of 'type' has been introduced to account for each ICL class separately. A type serves as a label for every submission, allowing us to analyze the behavior of ICL classes individually over time. Here, we are still considering a one-to-one relationship, meaning the REM is used to model interactions between inventors and ICL classes.

Each submission type is tracked using attributes that act like counters. For example, for ICL Class

[7] The hazard represents the risk or likelihood of a specific event

A, the attribute 'icl.class.A' increases by one for every submission labeled under this class. Separate attributes track submissions for other classes, such as Class B, Class C, and so on.

To identify patterns of repeated activity, we use the aggregation function MAX to find the highest level of activity recorded for a single edge within each class. This allows us to measure repetition consistency, which is the maximum number of times an inventor interacts with a specific ICL class. For instance, if an inventor repeatedly submits patents to ICL Class A, this framework highlights that focused and consistent behavior. Conversely, if submissions are spread across multiple classes, the repetition counts will be lower.

Analyzing the behavior of each class separately is useful because different classes represent distinct fields and by isolating the analysis for each ICL class, we can identify patterns of specialization and repeated contributions. Furthermore, combining this analysis with the CoxPH allows us to explore how repeated interactions influence the outcomes over time.

4.2.1 Interpretation

The analysis evaluates the hazard of an event, meaning the likelihood of observing an inventor's contribution to a specific ICL class on a given application date. It depends on how consistently inventors engage with that ICL class over time.

Each statistics *icl.class.*.consistency* represents the degree of consistency or involvement of inventors with a specific ICL class. The CoxPH model examines how these consistencies influence the hazard. Results are summarized in the table 3, where the hazard ratio is expressed as exp(coef), with the corresponding coefficient (coef) shown in parentheses.

Covariate	**Hazard Ratio (HR)**	**Interpretation**
icl.class.a.consistency	1.41 (0.343)***	Hazard increases by 40.98% per unit increase
icl.class.b.consistency	1.56 (0.445)***	Hazard increases by 56.10% per unit increase
icl.class.c.consistency	1.38 (0.320)***	Hazard increases by 37.82% per unit increase
icl.class.d.consistency	1.62 (0.479)***	Hazard increases by 61.58% per unit increase
icl.class.e.consistency	1.54 (0.434)***	Hazard increases by 54.45% per unit increase
icl.class.f.consistency	1.50 (0.408)***	Hazard increases by 50.50% per unit increase
icl.class.g.consistency	1.47 (0.387)***	Hazard increases by 47.38% per unit increase
icl.class.h.consistency	1.64 (0.496)***	Hazard increases by 64.29% per unit increase

Table 3: Cox proportional hazard models for event intensities associated with ICL Class consistency.

Note: *** indicates $p < 0.001$.

Inventors with higher consistency in working with specific ICL classes are more likely to be observed contributing to those classes. The hazard ratios indicate that ICL Class H (Electricity) has the strongest association with the hazard, followed by Class D (Textiles, Paper) and Class B (Performing Operations, Transporting). Consistency in collaborating or engaging with an ICL class increases the likelihood of observing the inventor within that class. Some ICL classes (e.g., Class H, Class D) show stronger associations, suggesting inventors may preferentially focus on these classes.

The shift from REM to RHEM captures multiple ICL classes for a single inventor over time, expanding beyond one-to-one relationships.

4.3 Third Configuration: RHEM

The process of submission of a patent in one or more technological classes by the same inventor is an *hyperevent* [8]. To analyze this network, a Relational Hyperevent Model has been used. A RHEM is a model that aim at describing the sequence of interactions (hyperevents) between sets of nodes, to see whether factors explaining the reasons of these events exists.

The following picture, figure 3, is a stylized representation of the analyzed network.

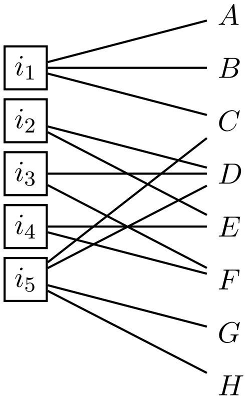

Figure 3: An example of a two-mode network RHEM.

In this new setting, the target size varies depending on the inventor (chosen as the event ID) for a given WKU, while the source size remains one. This is achieved by allowing loops in the model.

4.3.1 Interpretation

The analysis examines how often inventors tend to repeat associations with ICL Classes, focusing on both individual classes and subsets of classes. The table below, figure 4, presents the results obtained after applying a Cox proportional hazards (CoxPH) model. The hazard ratio is expressed as exp(coef), with the corresponding coefficient (coef) shown in parentheses.

[8] A relational hyperevent represents a time-stamped interaction among varying and potentially large numbers of nodes.

Covariate	Hazard Ratio (HR)	Interpretation
sub.rep.1 (Order 1)	1.59 (0.4662371)***	Hazard increases by 59.39% per unit increase
sub.rep.2 (Order 2)	0.85 (-0.1582096)***	Hazard decreases by 14.63% per unit increase
sub.rep.3 (Order 3)	1.33 (0.2880114)***	Hazard increases by 33.38% per unit increase
sub.rep.4 (Order 4)	1.55 (0.4407824)***	Hazard increases by 55.39% per unit increase
sub.rep.5 (Order 5)	NA	Not estimable

Table 4: CoxPH models for event intensities associated with subset-repetition levels of ICL Classes.

Note: *** indicates $p < 0.001$.

With regard to the repetition at the individual level, the results demonstrate a strong tendency for inventors to repeatedly associate with the same ICL Class. This is evident from the significantly positive effect of *sub.rep.1*. In particular, if an inventor has recently published or collaborated in a specific ICL Class, the probability of their being associated with the same class in future events is significantly elevated. This supports the idea of preferential attachment, where inventors are more likely to pursue classes in which they are already familiar with or have previously been active in.

Interestingly, the results of the repetition of order 2 indicate a negative effect for dyadic subset repetition. This suggests that when an inventor has been associated with a pair of ICL Classes recently, the likelihood of repeating this combination decreases. A possible explanation for this finding is saturation: after working across two classes, inventors may move on to explore beyond simple dyadic combinations.

From the results for triadic subset repetition, we observe a positive and significant effect, indicating that inventors are more likely to repeat associations with combinations of three ICL Classes. This aligns with the idea of a familiarity effect, where inventors submit patents consistently in the same subsets of classes they have previously worked in. A similar pattern emerges for higher-order subset repetition, where the results indicate a significant and positive effect. When inventors have previously been associated with four ICL Classes, the likelihood of them revisiting this subset increases substantially.

The results for the largest subset size are not available, likely because associations with five ICL Classes simultaneously are either rare or statistically negligible in the dataset. This could suggest that inventors rarely form high-order connections.

These findings are consistent with the notion of local density and clustering in inventor and ICL Class networks. Particularly at higher levels, repetition and subset-repetition, reinforce dense groups of associations, leading to recurring patterns of collaboration.

5 Coevolution

The next step for the analysis will involve modeling the coevolution of co-inventoring and technological classification networks using Relational Hyperevent Models (RHEM) with eventnet. In this framework, patent networks serve as an illustrative setting, where inventors, patents, and technological classes are interconnected. Specifically, inventors co-invent patents (a two-mode network), while patents are classified into technological classes (a one-mode network). Observations are represented as time-stamped hyperevents, capturing the simultaneous involvement of multiple inventors and technological classifications in the publication of each patent.

References

[1] Lerner, J., Hâncean, M.-G., & Lomi, A. (2024). Relational hyperevent models for the coevolution of coauthoring and citation networks. Journal of the Royal Statistical Society Series A: Statistics in Society. https://doi.org/10.1093/jrsssa/qnae068

[2] Lerner, J., Lomi, A., University of Konstanz, RWTH Aachen, & University of Italian Switzerland. (2023). Relational hyperevent models for polyadic interaction networks. Journal of the Royal Statistical Society Series A: Statistics in Society, 577–600. https://doi.org/10.1093/jrsssa/qnac012

[3] Lerner, J., Lomi, A., Mowbray, J., Rollings, N., Tranmer, M., & The Authors. (n.d.). Dynamic network analysis of contact diaries. In Social Networks (Vol. 66, pp. 224–236).

[4] Bianchi, F., & Lomi, A. (2022). From ties to events in the analysis of interorganizational exchange relations. Organizational Research Methods, 1–42. https://doi.org/10.1177/10944281211058469

[5] Lerner, J. (n.d.). EventNet: Repository for relational hyperevent models and network analysis. GitHub. https://github.com/juergenlerner/eventnet