Analyzing Patent Data

Alessia Berarducci & Federica Marini

Analysis of Social Networks

Project Report Final Version

January 15, 2025

Contents

1 Introduction **1**

2 Research Question **1**

3 Data Description and Data Preprocessing **2**

3.1 Co-Inventor Network 2

3.1.1 Statistics 3

4 Bipartite Graphs of Inventors Through the Lens of ICL Classification **3**

4.1 First configuration 4

4.1.1 Interpretation of the first configuration 6

4.2 Second configuration 7

4.2.1 Interpretation of the second configuration 7

4.3 Third Configuration: RHEM 8

4.3.1 Interpretation of the third configuration 9

5 Coevolution **10**

5.1 Interpretation of the Results 12

6 Conclusion **13**

A Supplementary Figures **16**

B Configuration-Level Analysis **17**

B.1 Detailed Explanation of Attributes and Statistics for Configuration 1 17

B.2 Detailed Explanation of Attributes and Statistics for Configuration 2 18

B.3 Detailed Explanation of Attributes and Statistics for Configuration 3 19

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, underlying 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. Finally, the CoxPH model is used to look at this coevolution model. It shows how social and technological forces work together to affect the bigger 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.

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?

[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`

- 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 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 5 in the Appendix.

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

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

3.1.1 Statistics

The total number of nodes is 381,442 and the number of edges 954,224. 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, figure A.0.2, 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). 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 134,691 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, figure A.0.3, 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 exists 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.

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

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.

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 4.1, visualizes the relationships between pairs of ICL classes, where the intensity of the color represents the frequency of co-occurrences between two classes.

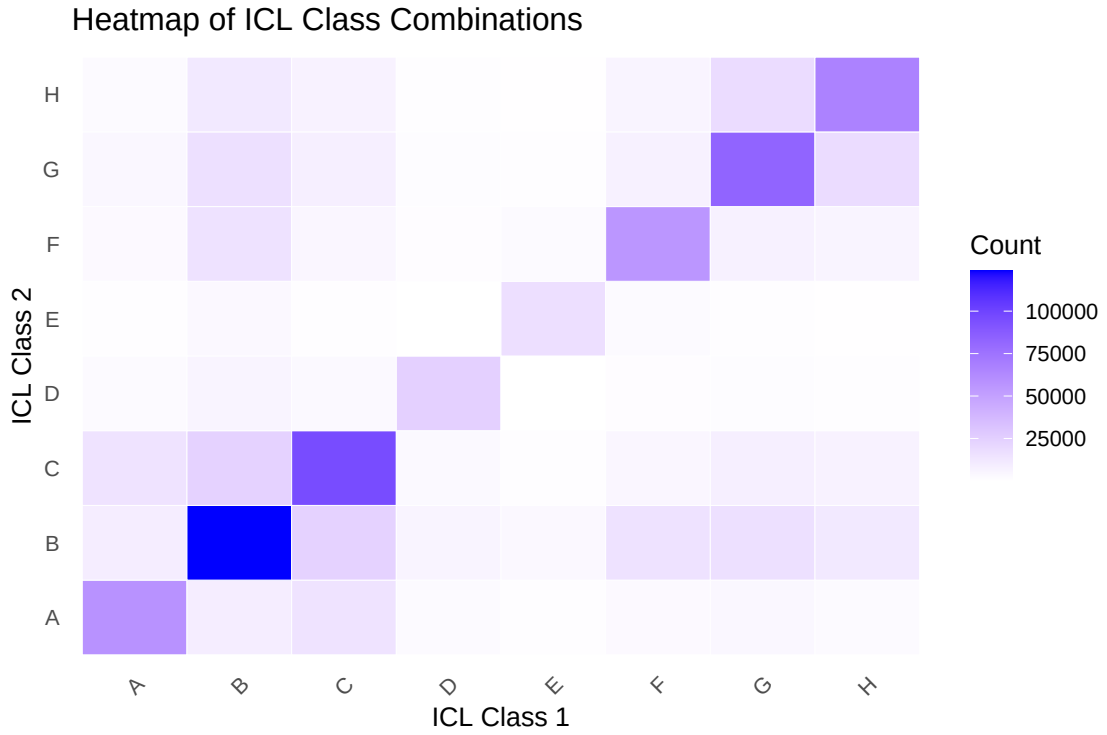


Figure 4.1: Heatmap of technological classes combinations

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 4.2, 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.

[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.

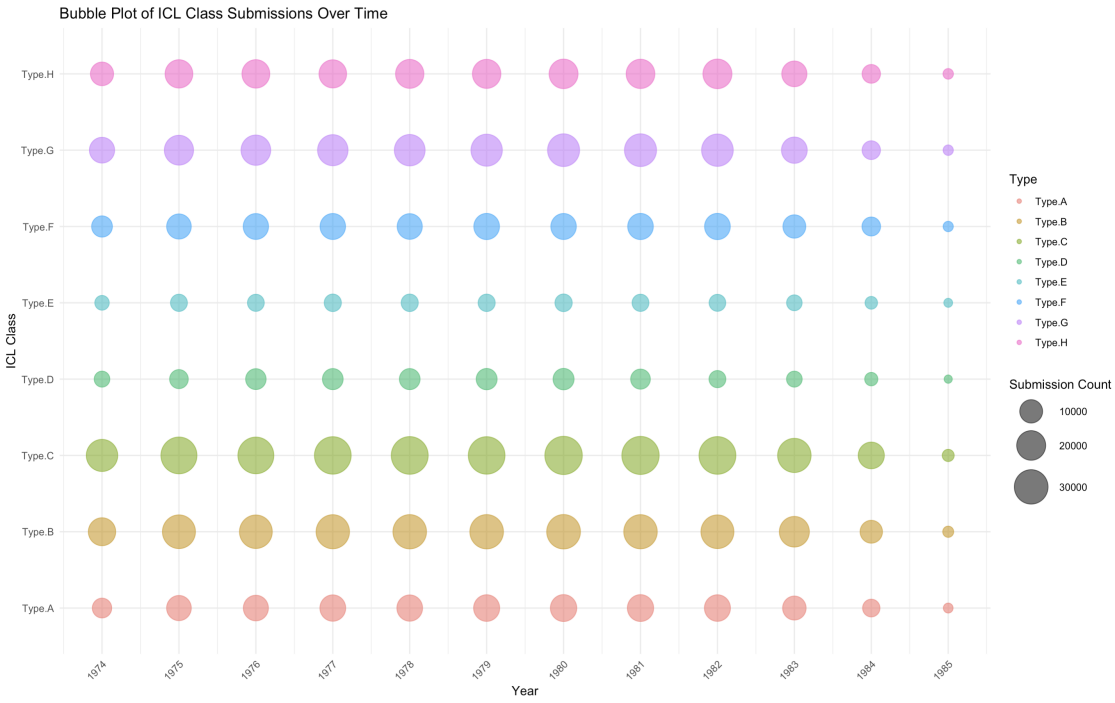


Figure 4.2: Trends in ICL Class Submissions (1974–1985)

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 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[7], 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).

The defined attributes are as follows:

- The author.hyed attribute, defined as a directed hyperedge-level, tracks inventors' activity by incrementing its value each time an inventor participates in a patent event within an ICL class.
- The *has.authored* attribute belongs to the node-level class, it tracks whether an inventor has submitted specific patents using the update mechanism 'SET_VALUE_TO'.

Additionally, various statistics have been computed using the defined attributes:

[7]More information about EventNet configuration can be found in Appendix B.1.

- *avg.patents.per.inventor*: is categorized as a Directed Hyperedge Node Statistic, it measures the average activity level of inventors across all technological classes.
- *avg.submissions.per.class*: is categorized as a Subset Repetition Statistic, it measures the average number of submissions per ICL class.

Lastly, the 'COND_SIZE_DHE_OBS' observation type has been set to analyze relationships based on the size of directed edges, with non-events paired one-to-one with events. This configuration has been applied equally to all setups.

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 2,887,806 and the number of events 1,443,903. The model assesses the hazard[8] 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 1, where the hazard ratio is expressed as the exponential of the coefficient 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 1: 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%. This could imply that highly active inventors stabilize their patenting activity, reducing the need for further submissions.

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%.

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

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.

The activity of inventors within specific ICL classes is tracked using attributes that function as counters. For example, the attribute *icl.class.A* is designed to monitor the frequency of interactions between inventors and ICL Class A. Every time an inventor contributes to a patent classified under ICL Class A, the attribute is incremented by one, ensuring it reflects the cumulative submissions over time. This is achieved through the 'INCREMENT_VALUE_BY' update mechanism, which triggers whenever an event of 'Type.A' (corresponding to ICL Class A) occurs. Similar attributes are used to track submissions for other classes, such as Class B and Class C.

To understand patterns of repeated activity, the framework employs statistics like *icl.class.a.consistency*. This statistic, categorized as a Subset Repetition Statistic, uses the aggregation function MAX to identify the highest level of activity recorded for a single edge within each class. Specifically, it quantifies the consistency with which inventors engage with a specifc ICL Class by measuring the maximum number of times an inventor has previously contributed to this class. For instance, if an inventor frequently submits patents to ICL Class A, the statistic highlights their focused and consistent behavior. Conversely, if their submissions are distributed across multiple classes, the repetition counts will be lower, indicating a less concentrated activity pattern. Furthermore, combining this analysis with the CoxPH allows us to explore how repeated interactions influence the outcomes over time.

4.2.1 Interpretation of the second configuration

The analysis [9] 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 statistic *icl.class.*.consistency*[10] represents the degree of consistency or involvement of inventors with a specific ICL class. Specifically, these statistics capture the maximum number of prior submissions an inventor-class pair has made, reflecting sustained engagement within that classification. The CoxPH model examines how these consistencies influence the hazard. Results are summarized in table 2.

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.

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

[9] More information about EventNet configuration can be found in Appendix B.2.

[10] * in icl.class.*.consistency stands for the A-H ICL Classes

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 2: Cox proportional hazard models for event intensities associated with ICL Class consistency.
Note: *** indicates $p < 0.001$.

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* [11]. 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 4.3, is a stylized representation of the analyzed network.

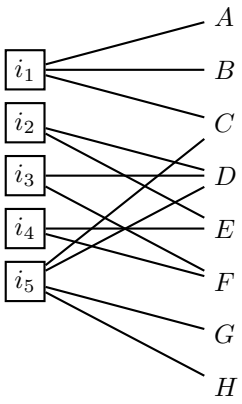


Figure 4.3: An example of a two-mode network RHEM.

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.

For Configuration 3, we utilized attributes and statistics to capture and analyze inventor activity across multiple event types.

The primary attribute, author.hyed, is a directed hyperedge-level attribute that tracks participation in hyperedge-related activities. It is incremented by one for each occurrence of specified event types (‘Type.A‘ to ‘Type.H‘), ensuring a cumulative measure of contributions over time.

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

The main statistic, sub.rep.1, calculates the average number of repeated submissions for single source-to-target interactions based on the *author.hyed* attribute. Similar statistics, *sub.rep.2* to *sub.rep.5*, measure repetition consistency for interactions involving increasing target sizes (from 2 to 5), with all other configurations remaining consistent with *sub.rep.1*.

4.3.1 Interpretation of the third configuration

The analysis[12] examines how often inventors tend to repeat associations with ICL Classes, focusing on both individual classes and subsets of classes. The table below, figure 3, 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.

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 3: 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.

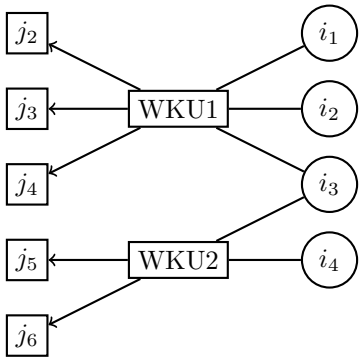
The results for the largest subset size are not available, likely because associations with five ICL Classes simultaneously are either rare or statistically negligible in the dataset. This could suggest that inventors rarely form high-order connections.

[12]More information about EventNet configuration can be found in Appendix B.3.

These findings align with the idea that networks of inventors and ICL Classes tend to form tightly connected groups, or clusters. At higher levels, repeated interactions and overlapping collaborations strengthen these clusters, creating dense networks of associations. This process leads to recurring patterns of collaboration, where the same groups or subsets of individuals frequently work together.

5 Coevolution

The coevolution of co-inventoring and technological classification networks can be modeled with relational hyperevent models (RHEM) using eventnet. An illustrative empirical setting is patent networks, which comprise two types of nodes: inventors and technological classes of patents they submitted, connected by two types of relations. Inventors are connected to the patents they co-invent, and patents are associated with specific technological classes based on their classifications. Patent networks can be viewed as mixed two-mode networks, where the inventor-co-invents-patent relation constitutes a one-mode network and the patent-is-classified-in-technological-class relation constitutes a two-mode network that is interrelated with the one-mode network. Observations in patent networks are sequences of time-stamped hyperevents, where each event corresponds to the publication of a patent, involving a varying number of co-inventors and technological classifications.



The data used for this analysis are stored in a CSV file containing information about inventors, their collaborations, and the technological classes associated with patents. Each row in the dataset represents an event, characterized by several key attributes:

- **Source:** The initiating inventor involved in the event.
- **Target:** The counterpart in the event, which could be:
  - Another inventor (indicating co-authorship in a patent).
  - A technological class (indicating the category of a patent submitted by the source inventor).
- **App_Date:** The date of the patent's application.
- **type:** Specifies the nature of the event:
  - *inv.auth.inv* for inventor collaborations.

  - *inv.sub.class* for patent submissions involving a technological class.

- **WKU_Standardized:** A unique identifier for each event.

The network is modeled as a multimodal structure with two distinct node sets: inventors (Source) and either other inventors or technological classes (Target), depending on the event type. The event types serve for exploring both social relationships (inventor collaboration) and technological trends (engagement with patent classes).

Due to the constraints posed by the dataset's size and complexity, we were unable to use the entire dataset. Including both types would have doubled its size, resulting in over 2 million rows. To address this, we employed stratified random sampling on the 'inv.sub.class' type, stratified by the years 1981 to 1984. Rather than using a fixed sample size, we weighted the samples according to the number of submissions for each year. These years were not chosen at random but aligned with a peak observed after 1980, particularly in September 1982. Furthermore, using WKU matching, we incorporated the 'inv.auth.inv' type by merging rows into a single dataset. Ultimately, the input file for EventNet was reduced to 50,324 events, of which 4,998 are of type inv.sub.class and 45,326 are of type inv.auth.inv.

To capture the dynamic nature of the network, attributes are defined at three different levels: node-level, dyad-level, and directed hyperedge-level.

At the node-level, the attributes analyze the inventors' individual behaviors and interactions. The *has.inventored* attribute tracks whether an inventor has participated in collaboration events ('inv.auth.inv') and updates with each new collaboration. Another key attribute, *inv.submission.activity*, measures the frequency of patent submissions by an inventor, reflecting their overall engagement with technological classes ('inv.sub.class'). Similarly, *inventor.submission.popularity* assesses an inventor's popularity based on how frequently they are listed as co-authors in patents. Finally, the *num.refs.of.patents* attribute records the number of references cited in the patents submitted by an inventor.

At the dyad-level, the attributes focus on relationships between inventors and their technological contexts. The *coinventor.dyadic* attribute quantifies the strength of co-inventor relationships by monitoring collaborative interactions over time. Another attribute, *inv.sub.iclclass.dyadic*, tracks the connections between inventors and the technological classes they engage with during patent submissions.

At the directed hyperedge-level, the attributes model collaborative and technological relationships more abstractly. The *inventor.hyed* attribute tracks directed hyperedges that represent inventor collaborations, while the *icl.class.hyed* attribute captures hyperedges connecting inventors to the technological classes associated with their patents.

The analysis employs a combination of node-level, dyad-level, and hyperedge-level statistics to capture and explain trends in the network.

At the node-level, several statistics are analyzed to quantify the activity and engagement of inventors and technological classes. For instance, the *has.submission.avg* metric provides information about the average activity levels of inventors by calculating their participation in collaborations ('inv.auth.inv'). Similarly, *avg.inv.subm.activity* measures the average number of technological classes an inventor interacts with ('inv.sub.class'). To complement these observations, *icl.outdegree.pop* evaluates the popularity of technological classes by assessing their connections to inventors.

The analysis also includes subset repetition statistics, which reveal patterns of repeated participation by subsets of inventors or technological classes in events. For instance, the *inv.sub.rep.1* metric examines the propensity of individual inventors to collaborate repeatedly, while *collaborate.with.inventor* focuses on whether specific groups of inventors repeatedly co-author patents.

Another critical component is triadic closure statistics, which explore the likelihood of inventors forming connections based on shared collaborators. For instance, the *inv.closure.by.coinv* statistic tests whether co-inventors of a particular source inventor are more likely to collaborate directly with one another.

Finally, the analysis considers directed closure statistics, which evaluate whether inventors linked through specific events, such as submitting the same patents, display a tendency to collaborate further.

The analysis incorporates a case-control sampling approach to balance event observations and non-events. A 'DEFAULT_DHE_OBS' configuration is used, generating one non-event per observed event.

5.1 Interpretation of the Results

Many statistics have been computed, but only a subset of them are statistically significant. The table below presents the results of the Cox proportional hazard models, with significant values highlighted using asterisks.

Covariate	**Hazard Ratio (HR)**	**Interpretation**
source.size	0.2420 (-1.419)***	Hazard decreases by 75.8% per unit increase
target.size	1.003 (0.003)	
has.submission.avg	0.997 (-0.003)**	Hazard decreases by 0.3% per unit increase
avg.inv.subm.activity	1.008 (0.008)***	Hazard increases by 0.8% per unit increase
diff.inv.subm.pop	0.991 (-0.009)	
icl.outdegree.pop	1.013 (0.012)***	Hazard increases by 1.3% per unit increase
inv.sub.rep.1	1.142 (0.133)	
inv.sub.rep.2	1.126 (0.118)	
collaborate.with.inventor	1.133 (0.125)	
inv.collaboration	0.874 (-0.134)	
inv.closure.by.coinv	1.059 (0.057)	
inv.closure.by.coinv.patent	1.134 (0.125)	
inv.closure.by.coinv.from.inv	1.337 (0.290)***	Hazard increases by 33.7% per unit increase
inv.closure.by.coinv.to.inv	0.868 (-0.142)***	Hazard decreases by 13.2% per unit increase
subm.patent.of.coinv	0.885 (-0.122)	
inv.subm.rep	1.000 (0.0002)	
inv.subm.reciprocation	0.998 (-0.002)*	Hazard decreases by 0.2% per unit increase

Table 4: Cox proportional hazard models for Coevolution Model.
Note: *** indicates $p < 0.001$, ** indicates $p < 0.01$, * indicates $p < 0.05$. Interpretation is blank for non-significant coefficients.

At the node level, the results highlight the role of individual inventors' activity and their connections to technological classes. The *source.size* attribute, which measures the size of the initiating group in

collaborations, shows a significant decrease in the hazard (75.8% per unit increase). This suggests that larger inventor groups that collaborate are less likely to initiate additional events, possibly due to coordination challenges. The *has.submission.avg* indicates a small but significant decrease in hazard (0.3% per unit increase). This may be due to a saturation effect, where highly active inventors run out of potential collaboration opportunities. In contrast, *avg.inv.subm.activity*, has a positive effect on the hazard (0.8% increase per unit). This suggests that inventors involved in a wide range of technological domains are more likely to be involved in future collaborations, likely due to the need to diversify technological classes over time to protect against issues like copying or infringement.

Additionally, the *icl.outdegree.pop* statistic suggests that highly popular technological classes (defined as those that are connected to multiple inventors) play a central role in the network, since a unit increase in popularity increases the hazard by 1.3%.

At the dyad level, statistics capturing patterns of repeated collaborations, do not show significant effects. This suggests that, while repeated partnerships may be common, they are not strong predictors of the likelihood of future events in this model.

The role of triadic closure in shaping the network is partially supported by the outcome.
In *inv.closure.by.coinv.from.inv*, a 33.7% increase in the hazard suggests that shared connections significantly facilitate the formation of new direct collaborations. This reinforces the idea that inventors who are already linked indirectly through a common collaborator tend to formalize their connection by collaborating directly. Conversely, the *inv.closure.by.coinv.to.inv* attribute, which captures directional dynamics in triadic relationships, shows a significant decrease in hazard (13.2% per unit). This suggests that directional relationships may reduce the likelihood of direct collaboration, potentially due to hierarchical dynamics.

Hyperedge-level attributes show limited significance in predicting collaboration likelihood. For instance, *subm.patent.of.coinv*, capturing direct patent submissions by co-inventors, does not significantly influence the hazard. Similarly, *inv.subm.rep*, reflecting repeated submission patterns among inventors, has no significant effect. However, *inv.subm.reciprocation*, which measures reciprocal submission behavior among inventors, shows a small but significant decrease in hazard (0.2% per unit). This could suggest that inventors engaged in reciprocal behaviors might be focusing their efforts within established networks, thereby limiting their likelihood of initiating new collaborations.

6 Conclusion

This project provides an analysis of the co-inventor network and its relationship with the technological classifications of patents. Advanced network modeling techniques, such as Relational Event Models and Relational Hyperevent Models,have been used to explain the characteristics of patent collaboration networks. The co-inventor network is characterized by a sparse structure, with a large number of isolated nodes and small communities. Also, there are some highly connected inventors, that are central promoting innovation and collaboration. The interaction between inventors and International Classification of Patents classes has been analyzed and a high degree of interdisciplinarity was observed, but certain classes like Physics and Electricity, constantly act as hubs for innovation. Inventors tend to specialize in certain ICL classes, repeating their contributions to familiar ICL classes, but occasionally they submit into new technological domains. The analysis of coevolutionary dynamics aims at studying the relationship between the co-inventorship network

and technological classifications. The integration of subset repetition and directed closure statistics in the coevolution model provided information about recurring patterns and trends within the networks, capturing both individual- and group-level behaviors. Future research can expand the period of time taken into account and include more recent patents. However, the use of a larger dataset can lead to computational challenges due to the high processing power required to handle such a vast amount of data. It is necessary to address this limitation by using optimized algorithms to manage the data efficiently. Furthermore, it would be interesting to incorporate other attributes in the model, to have a deeper understanding of how collaboration between inventors is created, such as geographic or organizational affiliations.

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

A Supplementary Figures

A.0.1

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 5: Description of the variables in the dataset.

A.0.2

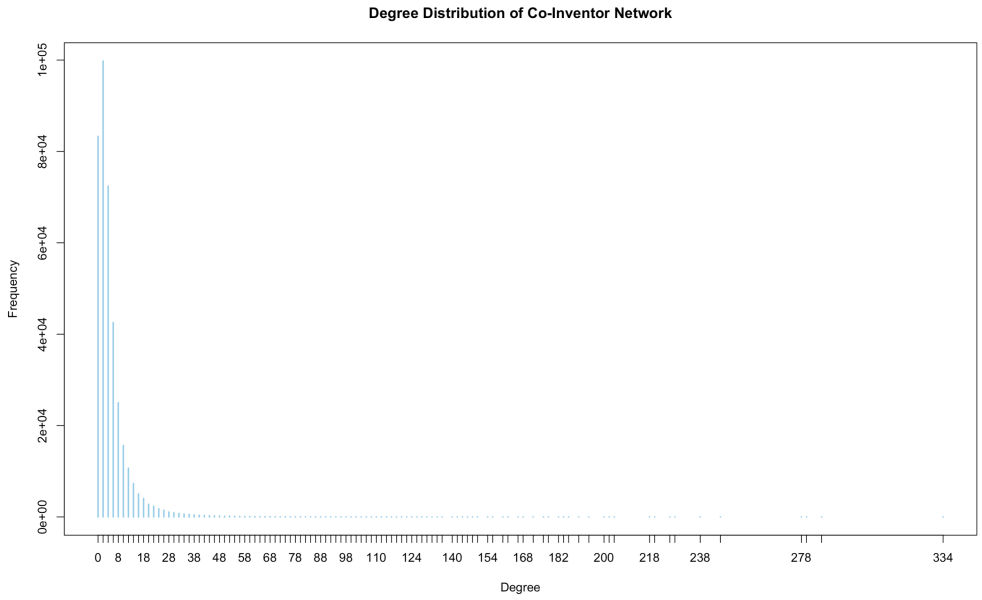


Figure A.1: Degree Distribution

A.0.3

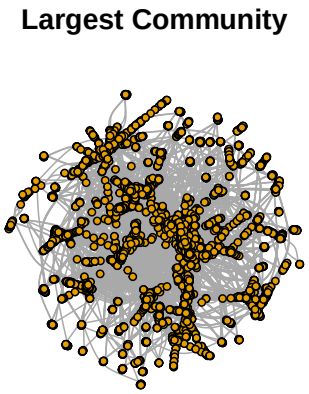


Figure A.2: Example of a Community

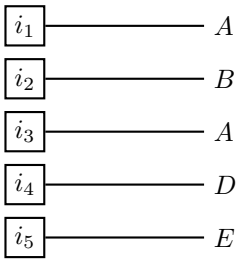


Figure A.3: REM

B Configuration-Level Analysis

B.1 Detailed Explanation of Attributes and Statistics for Configuration 1

B.1.1 Attribute: `author.hyed`

Class: directed hyperedge-level (`DIR_HYPER_LEVEL`)
Type: Default Directed Hyperedge Attribute (`DEFAULT_DHE_ATTRIBUTE`)

Update Mechanism:
`INCREMENT_VALUE_BY`: Every time an inventor participates in a patent event classified under ICL Class, the `author.hyed` attribute increments by one.

Event Response:

Responds to all events of type `EVENT`.

B.1.2 Attribute: `has.authored`

Class: node-level (`NODE_LEVEL`)
Type: Default node-level Attribute (`DEFAULT_NODE_LEVEL_ATTRIBUTE`)

Update Mechanism:
`SET_VALUE_TO`: The `has.authored` attribute is updated by setting its value to the specific event's defined property, tracking the inventor's authorship actions.

Event Response:
Responds to all events of type `EVENT`, specifically focusing on outbound (`OUT`) interactions originating from the node.

B.1.3 Statistic: `avg.patents.per.inventor`

Type: Directed Hyperedge Node Statistic (`DHE_NODE_STAT`)
Direction: SOURCE
Node Attribute: `has.authored`
NA Value: `-1.0`
Aggregation Function: AVERAGE
Description: The statistic `avg.patents.per.inventor` calculates the average number of patents authored per inventor by aggregating the `has.authored` attribute values across all relevant nodes in the network.

B.1.4 Statistic: `avg.submissions.per.class`

Type: Subset Repetition Statistic (`DHE_SUB_REPETITION_STAT`)
Direction: OUT
Hyperedge Attribute: `author.hyed`
Source Size: 1
Target Size: 1
Aggregation Function: AVERAGE
Description: The statistic `avg.submissions.per.class` measures the average number of prior submissions an inventor has made within ICL classes, across all relevant events.

B.2 Detailed Explanation of Attributes and Statistics for Configuration 2

We will explain one attribute and one statistic in detail, as the same principles apply to other attributes and statistics across different classes.

B.2.1 Attribute: `icl.class.A`

Class: directed hyperedge-level (`DIR_HYPER_LEVEL`)
Type: Default Directed Hyperedge Attribute (`DEFAULT_DHE_ATTRIBUTE`)
Description: The attribute `icl.class.A` is designed to track the frequency of interactions between inventors and ICL Class A. Specifically, each time an inventor contributes to a patent classified under ICL Class A, this attribute is incremented by one.

Update Mechanism:
`INCREMENT_VALUE_BY`: With each occurrence of a `Type.A` event (i.e., an inventor contributing to ICL Class A), the `icl.class.A` attribute increases by one. This mechanism ensures that the attribute captures the cumulative submission on the ICL Class A over time.

Event Response:
The `icl.class.A` attribute responds exclusively to events of `Type.A`. This means that only interactions where the event type is `Type.A` (corresponding to ICL Class A) updates to this attribute.

B.2.2 Statistic: `icl.class.a.consistency`

Type: Subset Repetition Statistic (`DHE_SUB_REPETITION_STAT`)
Direction: OUT
Hyperedge Attribute: `icl.class.A`
Source Size: 1
Target Size: 1
Aggregation Function: MAX
Description: The statistic `icl.class.a.consistency` quantifies the consistency with which inventors engage with ICL Class A. Specifically, it measures the maximum number of times an inventor has previously contributed to ICL Class A across all events.

B.3 Detailed Explanation of Attributes and Statistics for Configuration 3

B.3.1 Attribute: `author.hyed`

Class: directed hyperedge-level (`DIR_HYPER_LEVEL`)
Type: Default Directed Hyperedge Attribute (`DEFAULT_DHE_ATTRIBUTE`)

Update Mechanism:
`INCREMENT_VALUE_BY`: The `author.hyed` attribute is incremented by one each time an event of the specified types (`Type.A, Type.B, Type.C, Type.D, Type.E, Type.F, Type.G, Type.H`) occurs, capturing participation in hyperedge-related activities.

Event Response:
Responds to the following event types:

- `Type.A`
- `Type.B`
- `Type.C`
- `Type.D`
- `Type.E`
- `Type.F`
- `Type.G`
- `Type.H`

B.3.2 Statistic: `sub.rep.1`

Type: Subset Repetition Statistic (`DHE_SUB_REPETITION_STAT`)
Direction: OUT
Hyperedge Attribute: `author.hyed`
Source Size: 1
Target Size: 1
Aggregation Function: AVERAGE
Description: The statistic `sub.rep.1` calculates the average number of repeated submissions within a single source-to-target interaction based on the hyperedge attribute `author.hyed`. It aggregates these values over all events involving the specified hyperedges.

B.3.3 Statistic: `sub.rep.2 - sub.rep.5`

For all the other subset repetition statistics (`sub.rep.2`, `sub.rep.3`, `sub.rep.4`, `sub.rep.5`), the configuration remains the same as `sub.rep.1`. The only difference is the **Target Size**, which varies as follows:

- `sub.rep.2`: Target Size = 2
- `sub.rep.3`: Target Size = 3
- `sub.rep.4`: Target Size = 4
- `sub.rep.5`: Target Size = 5