Social Network Analysis: Analisi brevettuale

Pierluigi Brasile, Marco Mannarà, Alma Stira

Gennaio 2024

1 Introduzione

Prima di raccontare le fasi del progetto è necessario definire cos'è un brevetto. Un brevetto è il diritto conferito a un inventore non per l'uso o la pratica dell'invenzione, ma per impedire agli altri di praticare o utilizzare l'invenzione.
Negli ultimi anni, il ruolo dei brevetti è cambiato radicalmente. I brevetti erano un problema solo per un gruppo di praticanti legali o specialisti e nessun altro ne era veramente preoccupato. Ma con il cambiare dei tempi e la crescente concorrenza nel mondo degli affari, i brevetti sono diventati un fattore chiave per qualsiasi azienda.

Per scopi di questa ricerca, abbiamo utilizzato l'analisi dei brevetti per affrontare la gestione strategica della tecnologia dell'azienda.

Lo step iniziale è stato quello di scaricare dal database di "Lens" tramite una specifica query le informazioni sulle 10 aziende che hanno brevettato di più negli ultimi anni.
Ecco il link ipertestuale: clicca qui!

Nella fase di pulizia dei dati, abbiamo rimosso i record con valori mancanti e duplicati, nonché alcune caratteristiche non rilevanti, fino a ottenere un dataset che può essere riassunto come segue:

- **Publication_Number**: rappresenta il numero unico di pubblicazione assegnato a un brevetto quando viene rilasciato.
- **Publication_Year**: l'anno di pubblicazione del brevetto.
- **Inventors**: l'inventore o gli inventori che hanno sviluppato il brevetto.
- **Applicants**: l'entità o la persona che presenta una domanda per il rilascio di un diritto di proprietà industriale.
- **CPC_Classifications**: un elenco di tutti gli ID di classificazione assegnati a un singolo brevetto.

Il nucleo della nostra analisi si basa sui codici di classificazione CPC, poiché "i brevetti sono orientati alla protezione legale delle tecnologie e quindi la classificazione dei brevetti si basa su tecnologie o prodotti che utilizzano specifiche tecnologie".

In ogni brevetto, vengono assegnati diversi codici di classificazione CPC. La classificazione CPC è un codice alfanumerico gerarchico che rappresenta le caratteristiche di un brevetto. Più precisamente, le prime tre cifre del codice rappresentano la classe, le prime quattro cifre rappresentano la sottoclasse e il codice fino allo slash rappresenta il gruppo.

Dopo questo, abbiamo iniziato la fase di comprensione dei dati per avere una panoramica del nostro dataset (ad esempio, abbiamo cercato attori, paesi, ecc.).

Successivamente, abbiamo deciso di estrarre un elenco di sottoclassi che rappresentano solo i settori tecnologici. Approfondendo l'analisi, abbiamo anche estratto gruppi dalle sottoclassi per identificare tecnologie. Per trovare le sottoclassi, abbiamo seguito tre approcci diversi:

- Estrazione delle sottoclassi più frequenti e più citate.
- Estrazione delle sottoclassi dai brevetti più citati.
- Analisi di rete basata sulle occorrenze dei codici CPC di co-classificazione. La co-classificazione è l'insieme di tutti i codici CPC assegnati a un brevetto.

Sulla base di questo, abbiamo considerato le frequenze delle collaborazioni tra autori per uno stesso brevetto. Questo elemento è stata la chiave principale per realizzare un grafo bipartito non diretto, in cui i nodi rappresentano i brevetti e gli archi rappresentano i collegamenti tra di essi. Ciascun arco è ponderato dal conteggio delle occorrenze. Durante l'analisi, in particolare attraverso l'applicazione della Higher-order Network Analysis, si è ritenuto vantaggioso adottare un approccio diverso nella costruzione del grafo. Invece di focalizzarsi principalmente sugli autori, ci si è concentrati esclusivamente sulle CPC al fine di orientare l'analisi su di esse. Ciò è dovuto al fatto che, durante la creazione dell'ipergrafo, non era possibile preservare integralmente gli elementi chiave del grafo originale.

Di conseguenza, si è deciso di rielaborare il codice al fine di stabilire collegamenti tra i nodi in modo differente. Da questo punto in poi, soprattutto per l'open question, abbiamo focalizzato la nostra attenzione sui codici CPC e quindi sui settori in cui le aziende hanno brevettato.

2 Analisi del network

Il grafo sembra essere abbastanza grande e complesso, con una densa rete di collaborazioni. Esso è rappresentato da 13619 nodi e 46487 archi. Il fatto che non sia completamente connesso potrebbe indicare la presenza di gruppi di autori che non collaborano tra loro. Il grado medio e il coefficiente di clustering suggeriscono che potrebbe esserci una struttura comunitaria nella rete, con gruppi di autori che collaborano più frequentemente tra loro che con altri gruppi. Il livello di densità prossimo allo zero è interpretabile come un ulteriore indice della grandezza del grafo dato che rappresenta la frazione degli archi presenti rispetto al numero totale possibile di archi. Le caratteristiche della rete sono elencate di seguito:

Table 1: Features of the RW Network

Number of nodes	13619
Number of edges	46487
Weighted	Yes
Directed	No
Connected	No
Self-loops	No
Average degree	6.82
Average clustering coefficient	0.63
Network density	0.0005013

Abbiamo analizzato diversi modelli sul nostro grafo indiretto, , confrontando la rete Del Mondo Reale (RW) con reti di tipo Erdos-Renyi (ER), Barabasi-Albert (BA), Watts-Strogatz (WS) e Configuration Model (CM), al fine di ottenere una comprensione più completa della sua struttura, confrontare la realtà con le aspettative e determinare quale modello si adattasse meglio alla nostra rete reale. Per quanto riguarda CM è stato preso in considerazione come parametro la sequenza dei gradi; per BA, il numero di nodi del nostro grafo reale e il parametro di valore '4' come numero di collegamenti da aggiungere a ciascun nuovo nodo nel grafo BA; per WS, il numero di nodi del nostro grado e un parametro 'k' che rappresenta il numero di vicini di ciascun nodo e un parametro 'p' che rappresenta la probabilità di raggiungere un collegamento; mentre per quanto riguarda ER il numero di nodi e archi. Si potrebbe pensare che il modello ER sia quello che meglio si adatti alla nostra rete , tuttavia bisogna sempre ricordare che questo modello non tiene conto del peso degli archi e della loro distribuzione.

Table 2: Nodes and Edges in Various Networks

	CM	ER	WS	BA
Number of Nodes	13619	13619	13619	13619
Number of Edges	46406	46487	27238	54460

2.1 Degree Distribution

Confrontando la distribuzione dei gradi delle nostre reti, abbiamo notato che il Modello di Configurazione (CM) è quello più simile. Questo avviene perché il CM è un modello di rete casuale che si basa completamente su mantenere lo stesso grado della rete RW. Inoltre, la tendenza del grado della rete RW è molto simile anche a un modello BA, il quale segue quindi la formula successiva della distribuzione dei gradi e può essere considerato come una rete scale-free:

P(k) Ck

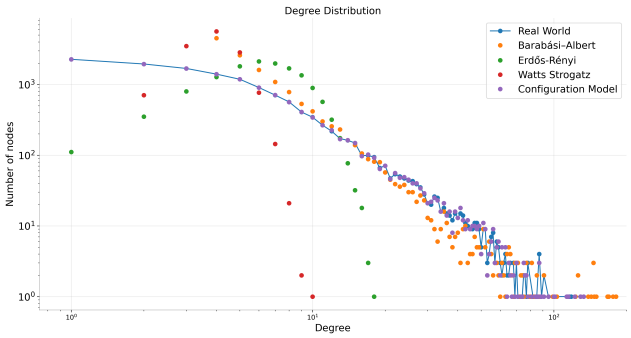
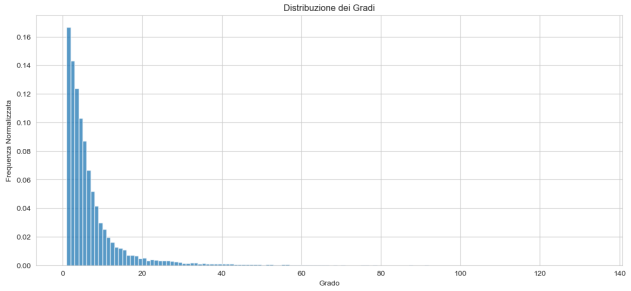


Figure 1: Degree Distribution

Nella rete RW, maggiore è il grado di un nodo minore è il numero di nodi con tale grado. Il nodo con il grado più alto ha un valore di 133 (etichettato con il nome 'Sako Yoichiro'), mentre il grado minimo è di 1, il che significa, non essendoci nodi con grado pari a 0, che non esistono nodi isolati. Il grado medio è invece di 6.826786107643733, ottenuto dalla formula (2 * num_archi) / num_nodi.



2.2 Connected Components

Il nostro grafo analizzato è composto da un totale di 45 componenti connesse, mentre le quattro altre reti dei modelli presi in considerazione sono caratterizzate da un numero diverso di componenti connesse:

- ER network: 16
  Questo modello mostra una struttura in cui il numero di componenti connesse può variare. Il fatto che ci siano 16 componenti suggerisce una distribuzione più dispersa rispetto al nostro grafo.

- BA network: 1
  Il modello BA genera reti con una distribuzione di grado a legge di potenza, spesso con un grande cluster fortemente connesso. Il fatto che ci sia solo una componente connessa indica che il modello BA ha generato una rete relativamente coesa e fortemente connessa.

- CM network: 26
  Il modello configurazionale cerca di riprodurre le distribuzioni di grado osservate in una rete esistente. Il fatto che ci siano 26 componenti connesse suggerisce che il modello ha generato una struttura di rete più frammentata rispetto al nostro grafo.

- WS network: 1
  Il modello WS è noto per la sua capacità di generare reti con elevato clustering locale e brevi percorsi. Il fatto che ci sia solo una componente connessa suggerisce che il modello WS ha generato una rete altamente integrata.

In generale, la diversità nel numero di componenti connesse tra i diversi modelli riflette le diverse caratteristiche strutturali introdotte dai diversi approcci generativi. La tua rete, con le sue 45 componenti connesse, potrebbe rappresentare una rete più frammentata e meno integrata rispetto a modelli come BA e WS. Queste differenze possono essere importanti per comprendere la struttura e la dinamica delle collaborazioni tra autori di brevetti nella tua rete; ad esempio, il modello Watts-Strogatz ha generato una rete altamente integrata con una sola componente connessa. Ciò potrebbe indicare che la tua rete ha una struttura di clustering locale e brevi percorsi simile a quella introdotta dal modello WS, deducibile anche dalla visualizzazione del grafo complessivo e dalle metriche precedentemente discusse.

2.3 Path Analysis

Il percorso medio più breve in un grafo indiretto rappresenta la misura della lunghezza media dei cammini più brevi tra tutte le coppie di nodi nel grafo, di conseguenza per questo scopo abbiamo considerato le componenti connesse di base nei vari modelli. Questa misura è un indicatore importante della "accessibilità" o della "distanza" media tra i nodi all'interno del grafo. Se il percorso medio più breve è basso, significa che, in media, è possibile raggiungere rapidamente un nodo qualsiasi da un altro nodo, il che nel nostro caso vorrebbe dire che le persone all'interno della rete sono strettamente connesse.

Table 3: Avarage Shortest Path

	RW	BA	WS	CM
Number of Nodes	13237	13619	13619	13569
Average Shortest Path	9	4	8	4
Connected Components	45	1	1	26

In base ai dati della tabella, si può notare che:

- Il modello BA mostra il percorso medio più breve più basso (4), suggerendo una maggiore connettività tra i nodi rispetto agli altri modelli.

- Il modello RW ha un percorso medio più breve di 9, indicando una maggiore distanza media tra i nodi rispetto agli altri modelli.

- Il modello WS ha un percorso medio più breve di 8, che si trova tra i valori del modello BA e del modello RW.

- I modelli BA e WS hanno una sola componente connessa, mentre il modello RW ha 45 componenti connesse e il modello CM ne ha 26.

2.4 Clustering Coefficient

Il clustering coefficient in una rete misura la tendenza dei nodi a formare gruppi densamente connessi o 'cluster'. Anche in questo caso abbiamo confrontato la nostra rete con i modelli classici e abbiamo ottenuto i seguenti risultati:

Table 4: Clustering Coefficients

	RW	ER	BA	WS	CM
avg	0.63	0.0004	0.004	0.18	0.002
min	0	0	0	0	0
max	1	0.34	0.34	1	1
mean	0.63	0.0003	0.004	0.19	0.003
stdev	0.39	0.006	0.02	0.2	0.03

Analizzando l'average clustering coefficient per ogni modello possiamo dedurre innanzitutto che la nostra RW ha un alto grado di clustering, il che significa che i nodi tendono a formare gruppi o cluster densamente connessi (nel nostro caso si parla di comunità o di attori

che cooperano strettamente). Questo è in contrasto sia con i modelli casuali (ER e CM) e il modello BA, che hanno coefficienti di clustering bassi. Mentre l'alto valore del coefficiente medio di clustering per il modello WS tende a suggerire che la nostra rete potrebbe avere caratteristiche di piccolo mondo, con un buon equilibrio tra Clustering e distanze corte tra i nodi.

Figure 2: Clustering Coefficients distribution

2.5 Density Analysis

La densità è una misura della quantità di collegamenti effettivamente presenti in rapporto al numero massimo possibile di collegamenti in una rete. Una densità maggiore indica una rete più densamente connessa, ma nel nostro caso la nostra RW presenta lo 0.05% delle possibili connessioni presenti nella rete, quindi ha un numero limitato di connessioni rispetto al numero totale possibile. Anche in questo caso abbiamo eseguito il confronto con i diversi modelli, ne seguono i risultati:

- RW : 0.0005
- ER : 0.0005
- BA : 0.0006
- WS : 0.0003
- CM : 0.0005

La nostra RW presenta risultati simili al modello ER e CM, mentre il modello BA ha una densità leggermente più alta e il modello WS ha una densità più bassa. In altre parole, la densità della rete RW è più vicina a quella delle reti casuali, il che suggerisce che le connessioni nella rete RW non seguono uno schema altamente strutturato o prevedibile come nelle reti BA e WS. Questo potrebbe significare che le connessioni nella rete RW sono più distribuite casualmente o non seguono un modello specifico come nei modelli BA e WS.

2.6 Degree Centrality

Il nodo con il grado più alto è 'SAKO YOICHIRO' (133), che è il nodo più importante, seguito da 'CHENG KANGGUO'(129) e 'BEWLAY BERNARD PATRICK', (123). Dalla figura 4, possiamo vedere che i modelli WS, BA e CM hanno valori molto più alti rispetto agli altri, infatti possiamo appena vedere le altre curve.

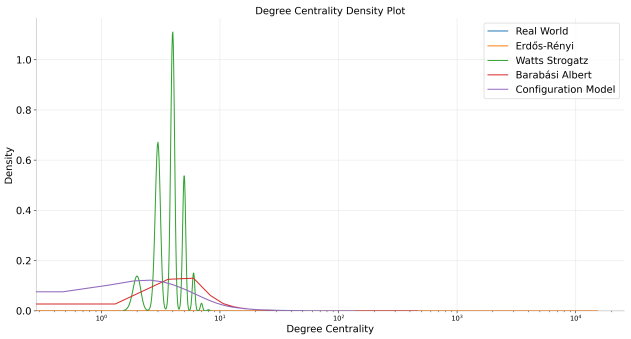


Figure 3: Degree Centrality distribution

2.7 Connectivity - Based Centralities

In tutte le seguenti centralità, i valori delle reti seguono lo stesso pattern: le curve RW e BA sono equivalenti, il modello CM differisce da esse ma ha comunque valori comparabili, mentre WS ed ER sono completamente diversi e mostrano un comportamento simile tra loro.

Eigenvector Centrality

I nodi con il valore più alto dell'Eigenvettore sono " IKEDA MASAMI" (0.239), "KASHINO TOSHIO" (0,213), "HIROYUKI" (0,186). Come dichiarato dalla definizione, questi nodi hanno un alto valore perché anche i nodi a cui sono collegati hanno un alto valore. Possiamo presumere che tutti e tre questi nodi siano collegati tra loro.

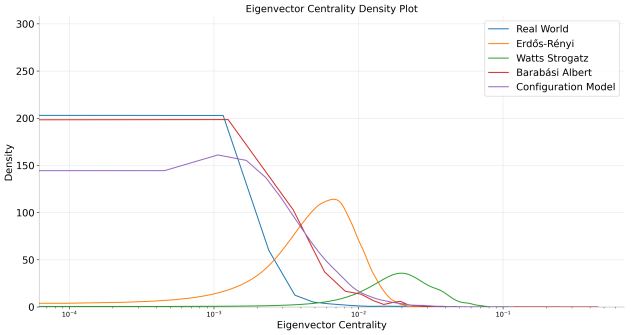


Figure 4: Eigenvector Centrality distribution

Centralità di PageRank

La centralità di PageRank di un nodo dipende da quanti collegamenti riceve, dalla propensione dei nodi

che lo collegano a creare collegamenti e dalla centralità di tali nodi. I nostri tre nodi con il valore di PageRank più alto sono " KONDO TETSUJIRO'" (0,00195), " CHENG KANGGUO" (0,00167) e "SAKO YOICHIRO" (0,00159).In altre parole, questi tre nodi sono considerati particolarmente importanti all'interno della rete in base a quanto sono collegati, a quanto i loro collegamenti sono importanti e a quanto sono importanti i nodi che li collegano. In questo caso troviamo però una somiglianza tra la nostra RW e il modello CM, mentre gli altri modelli si discostano significativamente.

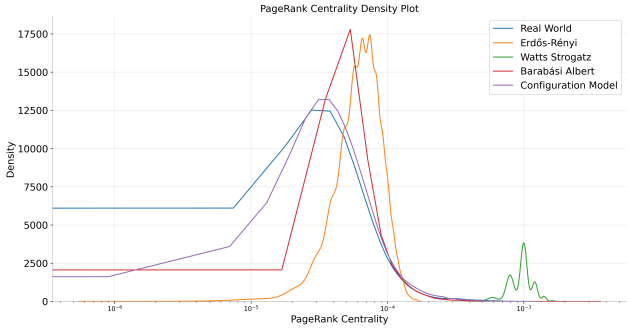


Figure 5: Page Rank Centrality distribution

Centralità di Katz

La Katz Centrality è una misura di centralità dei nodi in una rete, utilizzata per valutare l'importanza o l'influenza di un nodo rispetto agli altri nodi nella stessa rete. Questa metrica è particolarmente utile quando si considerano reti in cui le relazioni tra i nodi possono avere differenti gradi di importanza o in cui alcune connessioni sono più significative di altre.

Per quanto sia ritenuta una misura importante per la nostra analisi purtroppo non é stato possibile riportare questa metrica per insufficienza di potenza di calcolo

2.8 Geometric Centrality

Closeness Centrality

Il nodo con il valore di closeness più alto è il nodo più centrale, ovvero ha la distanza media più bassa con tutti gli altri nodi. Nel nostro caso, i nodi più centrali sono " SAKO YOICHIRO" (0,1555), "TAKANO HIROAKI" (0,1554) e "SAKODA KAZUYUKI" (0,1552). Se il valore fosse 1, significherebbe che questo nodo è direttamente connesso a tutti gli altri nodi, il che è impossibile nella nostra rete, dato che il valore di grado più alto è 133 anziché 13619-1.

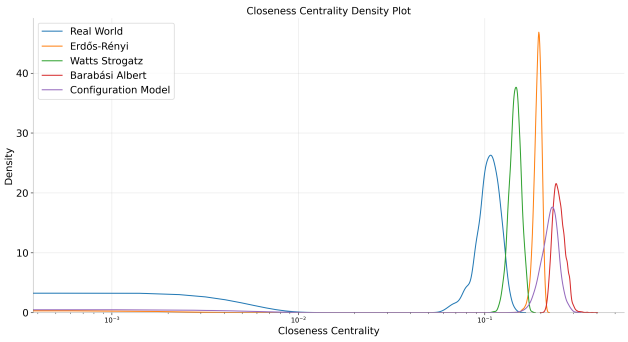


Figure 6: Closeness Centrality distribution

Harmonic Centrality

Il nodo con la più alta centralità armonica ha la media armonica più bassa delle distanze dei percorsi più brevi da un dato nodo a tutti gli altri, ed è quindi il nodo che rappresenta un compromesso tra la sua centralità e il suo grado. Ci aspettiamo che non si discosti dalla parte superiore, e infatti i nodi con i valori più alti sono "SATO HIROSHI" (2590.22), "SAKO YOICHIRO" (2535.05) e "AKAHASHI KENJI" (2519.17).

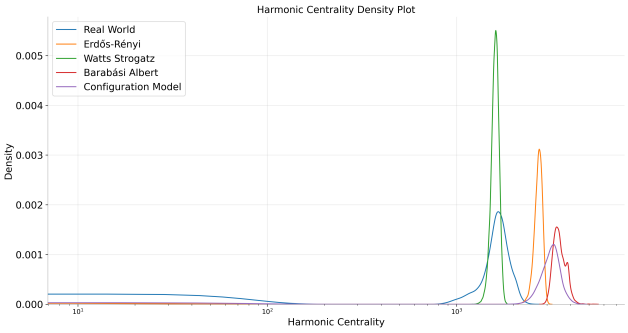


Figure 7: Harmonic Centrality distribution

In questo caso, la Centralità Armonica ha valori più bassi nella nostra rete, le curve assomigliano a distribuzioni normali e la rete WS ha un intervallo molto più stretto rispetto agli altri modelli.

Betweenness Centrality

Il nodo con la più alta Betweenness Centrality è quello attraversato dal maggior numero di percorsi più brevi. Nel nostro caso, "CHEN WEI" ($1.88 \times 10\hat{7}$), "GUO XIN" ($1.80 \times 10\hat{7}$), "TAKANO HIROAKI" ($1.77 \times 10\hat{7}$).

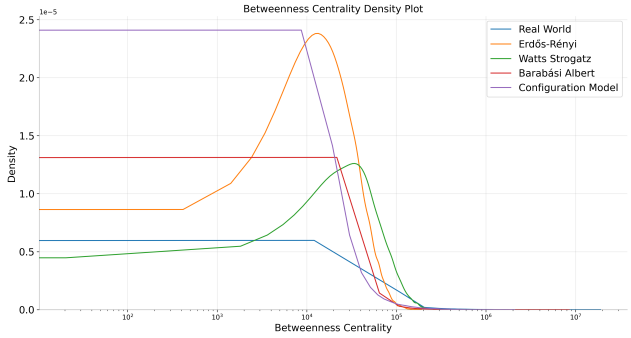


Figure 8: Betweenness Centrality distribution

Considerazioni finali

Confrontando le distribuzioni delle centralità di grado e quelle basate sulla connettività, possiamo notare come esse non siano affatto simili e che le geometric centralities producono risultati migliori sulla nostra rete.

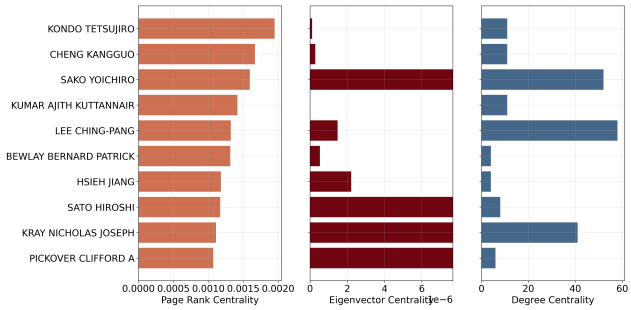


Figure 9: Top Ten Connectivity-Based and Degree Centrality

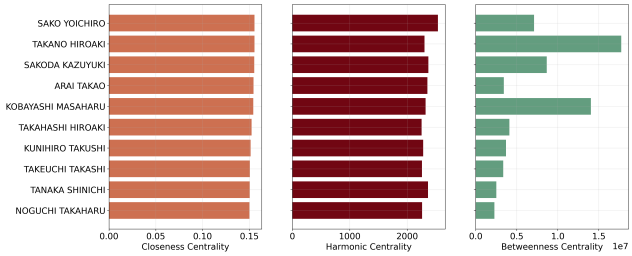


Figure 10: Top Ten Geometric Centrality

Analizzando la nostra rete, le parole più importanti sono quasi sempre le stesse, anche se, particolare importanza va data al nodo" `SAKO YOICHIRO`", che appare tra le prime tre parole di ogni misura, fatta eccezione per la Betweeness Centrality e la Eigenvector Centrality.

—

3 Data manipulation

3.1 Dynamic community discovery

Le comunità dinamiche cercano di riconoscere i rapidi cambiamenti nelle relazioni sociali utilizzando snapshot della rete in momenti diversi. Queste istantanee catturano la topologia della rete in diversi istanti temporali e, analizzando le differenze tra di esse, è possibile ricostruire la possibile evoluzione.

Il problema principale di questo approccio è legato alla inevitabile perdita di informazioni sull'evoluzione degli archi che si formano tra i nodi se l'intervallo di tempo tra ciascuna istantanea è troppo ampio, il che può essere anche descritto, in altre parole, come la mancata identificazione di cambiamenti interessanti - e delle relative intuizioni - nei comportamenti degli utenti in una prospettiva più dettagliata.

Dato che l'approccio dinamico alla scoperta delle comunità richiede una lista di archi arricchita da ulteriori informazioni, ovvero un'indicazione temporale, la divisione iniziale delle collaborazioni estratte in quattro diversi file .csv (uno per decennio) si è rivelata utile: ciò ha permesso di assegnare il rispettivo timestamp a ciascuna collaborazione in modo molto più semplice. In questi casi però il trade-off ricade nel riuscire a scegliere i giusti istanti temporali; ovviamente considerando la mole di dati era impensabile effettuare una grid-search temporale, di conseguenza abbiamo deciso intuitivamente di scegliere un range sufficiente per assicurarci una sufficiente dinamicità tra le collaborazioni dei brevetti.

Considerando la struttura della rete molto dispersiva, dato che presenta molti nodi con poche collaborazioni, ci si aspettava una performance bassa per quanto riguarda gli algoritmi di leiden/louvain i quali lavorano sull'ottimizzazione della modularity. Per verificare ciò infatti sono stati confrontati i risultati di 3 tipologie diverse di algoritmi: louvain/leiden, label propagation, k-clique; inaspettatamente quelli incentrati sulla densità sono risultati essere migliori secondo il nf1 score.

Legenda:

- *Primo timestamp* = 1°
- *Secondo timestamp* = 2°
- *Terzo timestamp* = 3°
- *Quarto timestamp* = 4°
- a = Louvain;
- **b = Leiden**;
- c = k-clique;
- d = Label propagation

Table 5: NF1 score comparison

	1°	2°	3°	4°	Media
a	0.0549	0.0477	0.3224	0.4973	0.2306
b	**0.0484**	**0.5478**	**0.3757**	**0.5055**	**0.3694**
c	0.0449	0.1327	0.2732	0.3465	0.2106
d	0.0353	0.1032	0.2048	0.2691	0.1531

Dai risultati ottenuti possiamo osservare la somiglianza di Louvain (a) con Leiden (b), infatti questi algoritmi non supervisionati ragionano in termini di ottimizzazione della modularità. In termini di stabilità risulta essere migliore Leiden rispetto a Louvain, ma i risultati delle comunità erano quasi equiparabili. La modularity score più alta risulta essere per la comunità di partenza, dopodichè una volta che abbiamo analizzato la loro metamorfosi nel corso dei decenni l'indice di densità è diminuito di circa 0.06-0.08 punti.

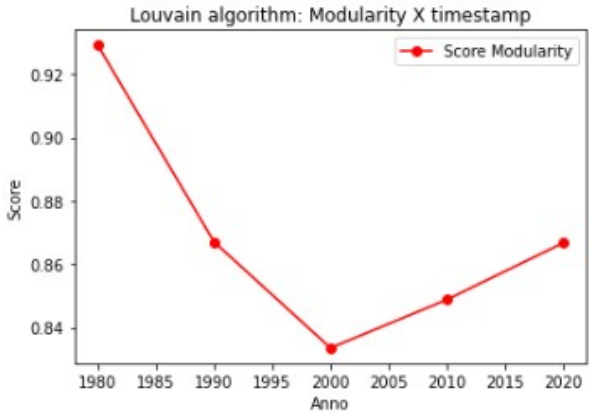


Figure 11: Performance della modularità per timestamp

Per entrare nel dettaglio si è deciso di analizzare le dimensioni delle comunità per ogni timestamp. Riportiamo qui sotto solamente le 20 comunità più grandi per il primo timestamp, ovvero quello con modularity maggiore.

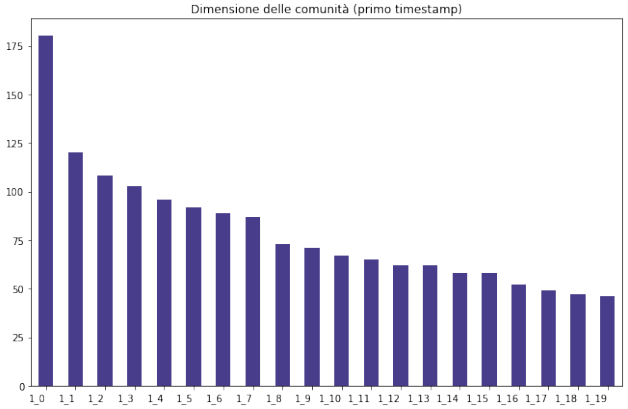


Figure 12: Andamento delle comunità per il primo timestamp

Le comunità sono state poi analizzate osservando il ciclo di vita di una comunità utilizzando il polytree creato da CDlib. I nodi del grafo rappresentano una comunità, contrassegnata con t_id e community_id forniti dall'oggetto di clustering temporale. Il polytree, per scopi di visualizzazione, è stato tracciato escludendo tutti gli altri nodi ad eccezione di quelli coinvolti nell'evoluzione della comunità contrassegnata come 1 nel primo timestamp. I colori delle comunità rispecchiamo i diversi istanti temporali su cui sono state isentificate le comunità. A partire dai nodi gialli contenenti un numero di inventori pari a 182, 58, 12, 62, 4, 13, giungiamo al nodo verde, 5_4, il quale arriva a contenere 3094 inventori.

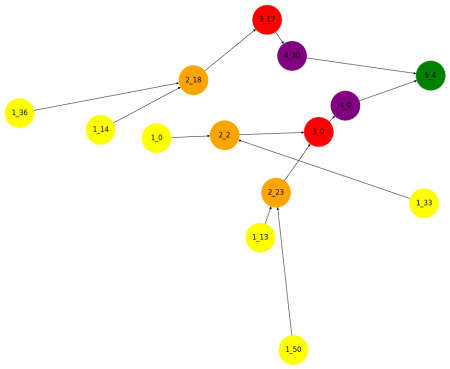


Figure 13: Andamento delle comunità nel tempo

Ai fini della nostra analisi sarebbe interessante sapere gli applicant che hanno mantenuto la collaborazione nel corso degli anni, e dunque le aziende collaboratrici e i settori di riferimento, piuttoso che individuare nuove collaborazioni. In questo modo ai fini di un'analisi settoriale le informazioni ricavate possono avere una confidenza maggiore, considerando che i brevetti non sono documenti che certificano le strate-

gie aziendali nel breve/medio periodo ma delle semplici invenzioni che potenzialmente possono essere impiegate per un determinato prodotto. Riserviamo questo approfondimento nell'ultima parte dell'open question, analizzando più particolari delle comunità dinamiche.

3.2 Graphlets

I graphlet sono piccoli sottografi connessi e non isomorfi all'interno di una grande rete. Lo scopo di questo compito è progettare un algoritmo per stimare il numero di graphlet di dimensione 3 e 4 nella nostra rete.

Definiamo un "fattore di campionamento", che rappresenta la frazione del numero di archi rispetto al numero totale di archi. Inizialmente, campioniamo archi casuali nel grafo in base al fattore di campionamento e, per ciascun arco, consideriamo tutti i nodi vicini collegati all'arco corrente. Successivamente, cerchiamo di contare il numero di graphlet target in tutte le combinazioni di tre e quattro nodi nell'elenco dei vicini, in base ai tipi di graphlet.

Alla fine, dopo aver iterato su tutti gli archi, il valore di conteggio finale viene diviso per il fattore di campionamento in modo da ottenere una stima del conteggio per l'intero grafo. Il fattore di campionamento rappresenta la percentuale del grafo su cui stiamo cercando i graphlet. Ad esempio, 0.5 significa che stiamo analizzando solo il 50 per cento dell'intero grafo. Abbiamo testato il nostro algoritmo su un grafo più piccolo con 3000 archi utilizzando diversi fattori di campionamento.

Nella seguente Tabella 6, presentiamo i risultati di questo test per il graphlet a forma di due stelle. La tabella mostra i risultati del test per il grafo con due graphlet a forma di stella. Come si può osservare, all'aumentare del fattore di campionamento, la conoscenza del grafo aumenta, portando a una stima del conteggio più vicina al numero effettivo. Dalla Tabella 7 si può notare che anche scegliendo un fattore di campionamento molto basso, si ottiene un errore relativo molto basso, il che indica che la nostra stima per l'intero grafo è affidabile.

Sampling Factor	Estimated	Relative Error
0.05	15192.0	0.253024
0.10	17592.0	0.135018
0.25	17878.0	0.120956
0.50	15120.0	0.256564
0.75	20368.0	0.001475
1.00	20338.0	0.000000

Table 6: Estimation of 3000 edges, two star

In base alla Tabella 7, il graphlet più frequente nella nostra rete è il Three star, il che è coerente con la struttura del nostro grafo, nonostante sia calcolato con un valore inferiore per il fattore di campionamento a causa di limitazioni computazionali.

Shape	Sampling Factor	Estimated
Triangle	0.005	0.0
Two Star	0.005	392720.0
Four Clique	0.005	0.0
Chordal Cycle	0.005	0.0
Tailed Triangle	0.005	0.0
Four Cycle	0.005	0.0
Three Star	0.001	5051600.0
Four Path	0.005	0.0

Table 7: Estimation on real network

3.3 Unsupervised link prediction

In questa sezione, ci occupiamo di un'altra sfida nel campo della scienza delle reti, ovvero la predizione dei collegamenti, con l'obiettivo di esplorare le connessioni implicite tra le invenzioni brevettate. La crescente quantità di dati brevettuali disponibili, infatti, ha reso fondamentale l'adozione di approcci avanzati per estrarre valore da queste risorse informatiche preziose, per avere una panoramica esaustiva delle relazioni tra brevetti, individuare tendenze emergenti nel campo tecnologico e identificare potenziali collaborazioni industriali.

Nel nostro caso, il problema della predizione dei collegamenti può essere descritto in questo modo: quali brevetti possono influenzarsi reciprocamente e quali connessioni tra di essi possiamo scoprire?

Il maggior ostacolo nei metodi di previsione dei collegamenti, che possono essere sia di tipo supervisionato che non, è la loro elevata complessità computazionale, che è espressa come $O(—V\hat{2}—)$, dove V rappresenta il numero di nodi presenti nell'intera rete. Un'altra sfida riguarda il fatto che nelle reti del mondo reale, le connessioni spesso sono scarse, il che potrebbe portare gli algoritmi a identificare erroneamente delle relazioni. Per affrontare il primo problema, abbiamo ridotto la complessità creando un grafo più piccolo contenente solo i nodi con un grado uguale o superiore a 20.

Inizialmente, questa analisi è stata condotta utilizzando un semplice approccio non supervisionato, ma successivamente abbiamo cercato anche di affrontare il problema in modo supervisionato, utilizzando diversi algoritmi di Machine Learning. Entrambe le analisi non hanno condotto a risultati ottimali, anzi, il team non si ritiene per nulla soddisfatto, ma avendo impiegato tempo ed energie sull'argomento si è deciso di esporlo ugualmente.

Nei paragrafi seguenti, ci concentreremo sulla previsione dei collegamenti non supervisionata, utilizzando metodi che si basano principalmente sulla struttura della rete, in particolare sulla vicinanza dei nodi.

Questi metodi includono Common Neighbors, Jaccard, Adamic Adar e un approccio basato sul rango dei nodi noto come SimRank.

Un riassunto di tutti i risultati è mostrato nella Tabella 8.

	Common Neighbors	Jaccard
AUC-ROC	0.0222	00242
	Adamic Adar	SimRank
AUC-ROC	0.02336	0.968

Table 8: Unsupervised predictors AUC-ROC

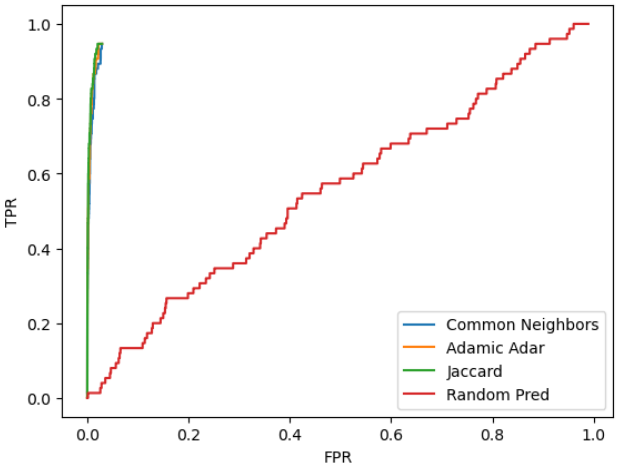


Figure 14: ROC curve

Common Neighbors

L'algoritmo Common Neighbors è basato sull'assunzione che coppie di nodi collegati in una rete, che sono anche connessi a un gran numero di altri nodi collegati, sono probabili a loro volta essere collegati tra di loro. In altre parole, se due nodi sono entrambi vicini a molti altri nodi nella stessa rete, c'è una maggiore probabilità che esista un collegamento diretto tra di loro. Questa idea viene formalizzata tramite la seguente formula: CN $(u,v) =$ —N (u) N (v)—

Dove N(u) e N (v) rappresentano i vicini del nodo u e v, rispettivamente. L'idea di base è che maggiore è il numero di nodi condivisi tra i vicini di u e i vicini di v, maggiore è la probabilità che u e v siano collegati direttamente.

Le prestazioni di questo algoritmo sono ancora oggetto di dibattito e non sempre producono risultati accurati. Questo aspetto è evidenziato dal punteggio AUC-ROC nell'analisi mostrata nella figura 15. La curva ROC misura la capacità dell'algoritmo di classificare correttamente i collegamenti veri e falsi, e l'AUC-ROC riflette l'efficacia complessiva dell'algoritmo. Pertanto, sebbene Common Neighbors possa essere utile in alcuni contesti, è importante considerare anche altri approcci nella predizione dei collegamenti per ottenere risultati più accurati in reti complesse del mondo reale.

Jaccard

L'algoritmo Jaccard è una misura di similarità utilizzata per valutare quanto due insiemi siano simili tra loro, simile all'algoritmo Common Neighbors (CN), che utilizza la similarità di Jaccard per identificare il numero di vicini comuni tra due nodi in una rete. La similarità di Jaccard è formalizzata come segue:

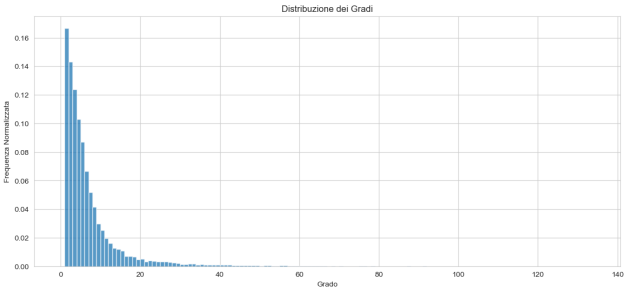


Figure 15: Distribuzione dei gradi

L'obiettivo di questo calcolo è valutare quanto i vicini di due nodi siano simili tra loro. Se due nodi condividono molti vicini, la loro similarità di Jaccard sarà alta, indicando una potenziale connessione tra di loro. Anche questa misura putroppo proprio come Common Neighbors non ha ottenuto risultati ottimali meritevoli di discussione.

Adamic Adar

L'algoritmo Adamic-Adar è incluso nel gruppo di algoritmi basati sulla similarità tra nodi in una rete. La principale differenza rispetto agli algoritmi Common Neighbors e Jaccard è che tiene conto del grado dei nodi: quindi, ciò che accade è che i nodi con un grado più basso hanno un maggiore impatto, mentre i "hub" sono meno importanti per la predizione dei collegamenti, questo perché i nodi con un grado più alto potrebbero condividere molti vicini comuni semplice-

mente a causa del loro elevato grado, ma questo non fornisce molte informazioni uniche sulla loro similarità.

Questo metodo, come gli altri della stessa categoria, mostra molta difficoltà nel trovare nuovi collegamenti possibili.

SimRank

Il SimRank è un approccio basato su un sistema di classificazione alla predizione dei collegamenti. Calcola la probabilità di vedere un nuovo collegamento tra nodi (o brevetti in questo caso) analizzando la similarità dei loro rispettivi vicinati. Questo approccio ha restituito risultati migliori come può essere visto nella Figura 17.

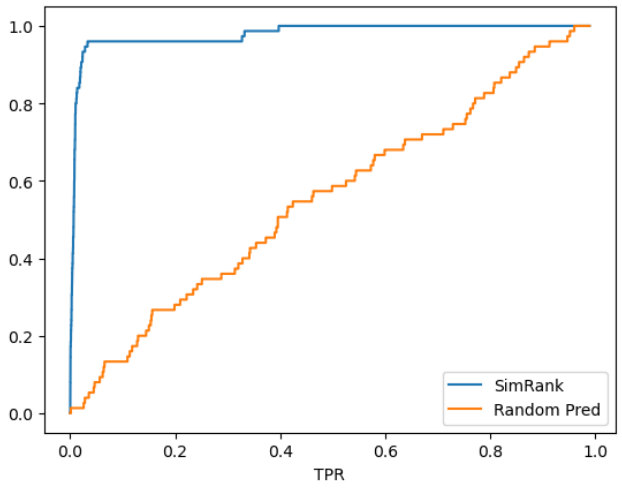


Figure 16: SimRank ROC curve

Supervised Link Prediction

La predizione dei collegamenti supervisionata implica l'uso di un modello di apprendimento automatico per identificare i potenziali collegamenti futuri in una rete. Questo processo è stato eseguito utilizzando la ben nota libreria scikit-learn. La prima fase di questa procedura è stata la preparazione dei dati, che è stata eseguita utilizzando due librerie diverse: StellarGraph13, che ci ha permesso di suddividere i nodi in un set di allenamento e uno di test, e node2vec14, che ha estratto i dati essenziali per applicare gli algoritmi di apprendimento automatico utilizzando un metodo di camminata casuale di secondo ordine.

I nodi sono stati separati in due gruppi distinti: uno per l'allenamento e uno per il test, seguendo le linee guida di StellarGraph15. Questa divisione ha portato a un set di allenamento composto da 6172 nodi, un set di convalida con 2058 nodi e, non da meno, un set di test contenente 9144 nodi. Successivamente, tutti questi nodi sono stati trasformati in vettori che sono stati impiegati per ottimizzare i parametri dei modelli, nonché per l'addestramento e i test successivi.

I risultati, come evidenziati nella Tabella 9, sono notevolmente superiori rispetto a quelli ottenuti mediante gli approcci non supervisionati. Il Random Forest si è dimostrato il modello meno performante, presentando una precisione media e un punteggio AUC-ROC approssimativamente in linea con quelli di un predittore casuale.

Legenda:

- LR = Logistic Regression;
- RF = Random Forest;
- SVM = Singular Value Decomposition;

	LR	RM	SVM
Accuracy	0.51	0.49	0.51
Precision	0.51	0.49	0.51
Recall	0.49	0.48	0.50
F-1 score	0.50	0.48	0.50
AUC-ROC	0.52	0.49	0.49

Table 9: Supervised predictors evaluation

I risultati ottenuti nei test rappresentano il vertice in termini di performance tra gli approcci supervisionati "tradizionali". Come già specificato in principio i risultati della nostra analisi n on risultano essere per niente soddisfacenti, come ben si può notare dalla precision del 51% sul modello più performante.

3.4 Higher-order Network Analysis

L'adozione di un approccio basato su ipergrafi si configura come una strategia avanzata e altamente significativa per sondare la complessità delle relazioni presenti nei dati brevettuali. La struttura flessibile di un ipergrafo consente di catturare non solo le connessioni tra gli inventori dei brevetti, ma anche le relazioni tra concetti e classificazioni correlate. Nel nostro contesto specifico, abbiamo scelto di utilizzare un ipergrafo focalizzandoci sui codici CPC associati ai vari brevetti. Abbiamo successivamente condotto un'analisi dei comportamenti delle connessioni attraverso una task di clustering e una loro valutazione temporale.

Per illustrare in modo più rappresentativo questo approccio, presentiamo un subplot che visualizza gli iper-nodi raffiguranti le classificazioni CPC e gli iperarchi rappresentati dai brevetti.

Questa rappresentazione grafica offre una panoramica visiva delle relazioni tra le classificazioni CPC e i brevetti, fornendo un'anteprima della complessità delle connessioni nel contesto brevettuale.

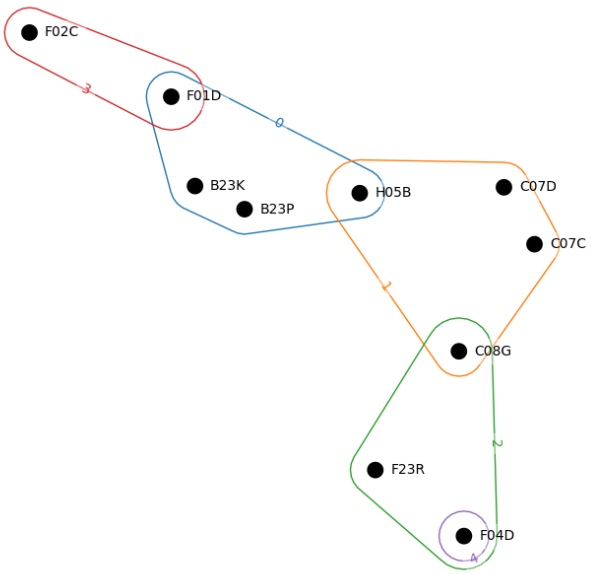


Figure 17: Sotto-ipergrafo dei brevetti e delle CPC classification

Prima di immergerci nelle fasi analitiche, desideriamo avvertire il lettore che la vasta quantità di dati a nostra disposizione ci ha impedito di testare i modelli di clustering sull'intero ipergrafo. Di conseguenza, abbiamo preso la decisione di concentrarci esclusivamente sugli iperarchi con un grado più elevato.

All'interno del nostro dataset, l'unico iper-arco con una centralità di grado superiore a 0.5 risultava essere 'Y10T'.

Pertanto, abbiamo proceduto filtrando i brevetti che contenevano esclusivamente questo codice CPC, riducendo così le righe del dataset da 296639 a 13030. Questa operazione ci ha permesso di evitare possibili errori legati al tempo di esecuzione.

3.4.1 Kumar clustering

Questo approccio fa uso dell'algoritmo di Louvain applicato alla rappresentazione a due sezioni del grafo. In questo contesto la modularità dell'ipergrafo è utilizzata come misura della qualità di una data partizione dei suoi vertici.

La modularità dell'ipergrafo offre una valutazione della bontà di una partizione, dove partizioni casuali tendono a produrre una modularità prossima allo zero (anche potenzialmente negativa). Al contrario, una modularità positiva suggerisce la presenza di comunità dense o moduli all'interno dell'ipergrafo.

È importante notare che ci sono diverse varianti per la definizione della modularità dell'ipergrafo. La principale differenza tra queste varianti risiede nel peso attribuito agli archi. La scelta del modello di pesatura può influenzare significativamente la scoperta di strutture di comunità nell'ipergrafo. Questa variazione consente una maggiore flessibilità nella descrizione delle relazioni e nella rilevazione di comunità significative all'interno dei dati ipergrafici..

I metodi utilizzati per la nostra analisi sono *linear*, *majority*, *strict*. Qui di seguito riportiamo i risultati della modularity in base alla tipologia di peso adottato nel modello.

	Linear	Strict	Majority
Modularity	0.006	0.002	0.003

Table 10: Misure della modularity per una partizione casuale

Questo confronto è stato condotto per esaminare il comportamento del modello rispetto alle tre varianti. Nel nostro caso, le misure risultano essere prossime allo zero, indicando un risultato pressoché indifferente tra le varianti. Tuttavia, va notato che, a causa di problematiche legate alle versioni delle librerie, non è stato possibile calcolare i pesi associati a ciascuna variante.

Per questo motivo, abbiamo eseguito il modello di Kumar, e il metodo basato sulla *Majority* ha restituito una metrica di modularità pari a 0.31. Questo valore fornisce una valutazione della qualità della partizione ottenuta dal modello di Kumar, indicando una tendenza verso la presenza di comunità dense o moduli all'interno dell'ipergrafo. Va sottolineato che, anche se le misure precedentemente menzionate sono prossime allo zero, il risultato specifico ottenuto con il metodo *Majority* suggerisce un qualche livello di struttura e raggruppamento significativo nei dati ipergrafici.

3.4.2 Analisi temporale delle CPC classification

Per completare l'analisi del comportamento degli ipergrafi, abbiamo preso la decisione di esaminare l'evoluzione delle macro classificazioni CPC nel corso dell'ultimo ventennio. A tal fine, il dataset è stato suddiviso in base agli anni, seguendo l'approccio precedentemente adottato per la task di temporal clustering. Per ciascun settore, abbiamo calcolato il grado degli iperarchi.

Questo approccio ci consente di ottenere una visione dettagliata delle variazioni nel numero di connessioni o relazioni tra le diverse macro classificazioni CPC nel corso del tempo. Analizzare il grado degli iperarchi per ogni anno ci permette di identificare trend, cambiamenti significativi e settori particolarmente attivi o rilevanti all'interno dell'ipergrafo.

Questo tipo di analisi temporale aggiunge un elemento dinamico alla nostra comprensione degli ipergrafi, consentendoci di esaminare come le relazioni tra

le classificazioni CPC si siano evolute nel corso degli anni.

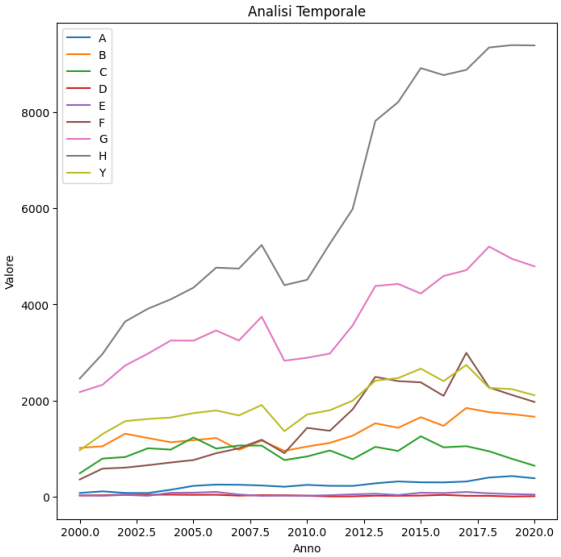


Figure 18: Analisi temporale delle CPC tramite degree

4 Open Question

4.1 Introduzione

Per la nostra task finale, esploriamo l'open question: "Previsioni di investimento delle aziende tramite analisi brevettuale". Il modello principale per le nostre previsioni sarà il modello ARIMA.

4.1.1 Modello ARIMA

Il modello ARIMA, acronimo di AutoRegressive Integrated Moving Average, è una famiglia di modelli statistici ampiamente utilizzata per l'analisi e la previsione dei dati temporali. Questo modello incorpora tre caratteristiche fondamentali:

- **Autoregressività (AR):** Predice i valori futuri sulla base di quelli passati, catturando le relazioni di dipendenza temporale nei dati.
- **Integrazione (I):** Rimuove le tendenze nei dati, rendendoli stazionari e facilitando l'analisi.
- **Media Mobile (MA):** Utilizza una media mobile come metrica per fare previsioni, incorporando informazioni sull'andamento medio dei dati.

4.1.2 Scelta della Metrica tramite Random Forest

La decisione sulla metrica utilizzata per la predizione di ARIMA sarà guidata dalla feature importance di un random forest applicato al grafo di partenza. Questo approccio ci consente di identificare la feature più influente nel nostro grafo, conferendo maggiore capacità di classificare nuovi cpc dal nostro dataset di allenamento.

Il grafo di analisi per questa task è diverso da quello iniziale, dato che tiene conto di un criterio di connessione basato sulle cpc e non sulla collaborazione tra autori. Le caratteristiche della rete sono cambiate, e l'analisi delle metriche mostra un elevato livello di connessione. Per gestire ciò, abbiamo effettuato un campionamento casuale del dataset per garantire tempi di analisi ragionevoli.

Il modello di regressione Random Forest addestrato ha dimostrato ottime performance sia sui dati di addestramento che su quelli di testing. Il punteggio sul training data è di circa 0.9392, indicando un adattamento significativo ai dati di addestramento. Inoltre, il modello mantiene un punteggio elevato anche sui dati di testing (circa 0.9384), suggerendo una buona capacità di generalizzazione.

Questi risultati indicano che il modello Random Forest è in grado di catturare efficacemente le relazioni nei dati e di fare previsioni accurate anche su nuovi dati. Tuttavia, si raccomanda un'ulteriore valutazione delle metriche e una comparazione con altri modelli per garantire la scelta ottimale per il nostro specifico problema di regressione.

Dalla feature importance del random forest emerge che le metriche più rilevanti sono Closeness e Degree Centrality. Tuttavia, nell'attuazione del codice, il calcolo di queste metriche sull'intero grafo risultava essere troppo dispendioso in termini computazionali rispetto alla scelta del Degree. Per questo motivo, è stato preferito quest'ultimo per effettuare le previsioni.

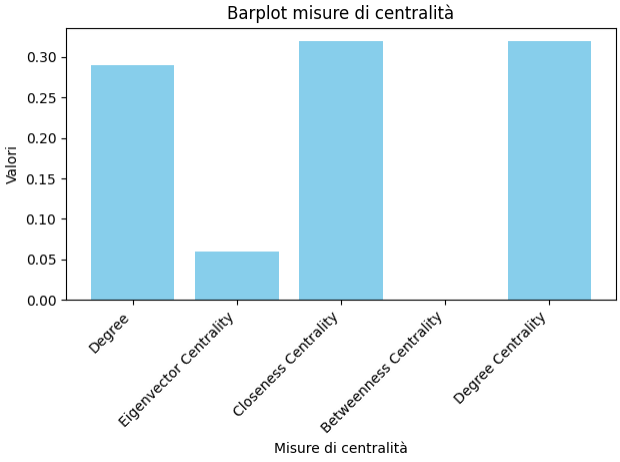


Figure 19: Barplot della feature importance

4.1.3 Spiegazione ed analisi del Modello ARIMA

Successivamente, abbiamo proceduto con l'analisi temporale, dividendo i grafi in snapshot e calcolando le misure di degree per ogni anno e nodo. Le informazioni sulle cpc classifications sono state ottenute dai nodi, corrispondenti ai brevetti stessi. Per maggior chiarezza e per maggior utilità in riferimento alla nostra Open Question abbiamo selezionato i nodi più interessanti analizzando quelli con media delle degree maggiore. Riportiamo qui di seguito i Brevetti e le relative CPC classification:

Brevetti ID	CPC Classification
26	C08G
29	F01D
70	F01D
74	C23C
140	C22C

Table 11: Tracciamento dei settori dai brevetti

I codici che risultano uguali in realtà non lo sono perchè ai fini dell'analisi sono state scelte i primi 4 elementi di una stringa che ne contiene molti di più. Abbiamo creato e addestrato il modello ARIMA per ciascuna serie temporale, incorporando informazioni sull'autocorrelazione e la media mobile. Abbiamo esteso le previsioni per un numero specificato di passi temporali, permettendo di proiettare il comportamento potenziale delle serie nel futuro prossimo. Nel nostro contesto, stiamo considerando un approccio di previsione basato sulla tendenza storica dei dati. Fattori come la stagionalità, gli eventi eccezionali o cambiamenti improvvisi nelle condizioni possono influenzare significativamente la precisione delle previsioni. Gli ordini AR, differenziazione (d), e MA sono parametri la cui scelta può essere ottimizzata in base all'analisi dei grafici ACF (Autocorrelation Function) e PACF (Partial Autocorrelation Function).

La differenziazione delle serie temporali è una tecnica fondamentale nell'analisi delle serie storiche, mirata a rendere stazionaria una serie che presenta tendenze o comportamenti non costanti nel tempo. In un certo senso si normalizzano le tendenze all'interno di un range stazionario. La differenziazione delle serie temporali è una strategia chiave per ottenere serie stazionarie, facilitando l'applicazione di modelli statistici avanzati. Al fine di valutare l'efficacia della differenziazione nel rendere stazionaria la serie temporale, è possibile calcolare il test ADF (Augmented Dickey-Fuller) sulla serie differenziata. Non avendo conoscenza approfondita del tema abbiamo preferito non riportare l'analisi riportata nel codice.

4.1.4 Conclusione: Macro-Settori Emersi

Riportiamo qui di seguito il grafico finale delle previsioni.

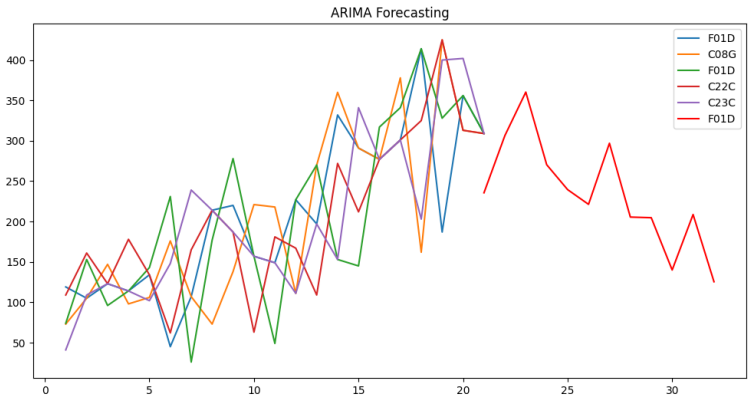


Figure 20: Previsioni temporali dei macro-settori

Le previsioni dei settori di analisi sono state raccolte dentro la linea arancione, la quale rappresenta la media dei valori. Riteniamo utile sottolineare che la scelta dei parametri cambia drasticamente il risultato delle previsioni e di conseguenza la direzione del trend, quindi per un'analisi più efficace è importante effettuare una grid search dei parametri citati in precedenza.

L'analisi brevettuale condotta ha portato all'identificazione di macro-settori significativi, evidenziando tendenze e aree chiave di innovazione. Di seguito sono riportati e descritti i macro-settori risultanti:

1. **Ingegneria Meccanica; Illuminazione; Riscaldamento; Armamenti; Esplosivi:**
   - Questo macro-settore abbraccia un'ampia gamma di tecnologie e invenzioni, spaziando dall'ingegneria meccanica alle soluzioni di illuminazione e sistemi di riscaldamento. La presenza di armamenti ed esplosivi suggerisce un focus su innovazioni nella sicurezza e nelle tecnologie di difesa.
2. **Composti Organici Macromolecolari; Preparazione o Lavorazione Chimica; Composizioni a Base di Questi Composti:**
   - Questo settore si concentra sulla chimica dei composti organici macromolecolari, con un'enfasi particolare sulla preparazione, lavorazione chimica e composizioni derivanti da tali composti. Questa area riflette l'importanza delle innovazioni chimiche e dei materiali organici nel panorama brevettuale.