Analisi delle Reti Semantiche nel Panorama Mediatico Italiano

Giordana Mazzone
g.mazzone7@studenti.unipi.it
Student ID: 692262

a.a. 2024-2025

Sommario

Questo studio utilizza la Social Network Analysis (SNA) per mappare il panorama mediatico italiano e verificare l'esistenza di echo chamber ideologiche. È stata costruita una rete basata sulla similarità semantica (TF-IDF) di migliaia di articoli di testate giornalistiche bilanciate. L'analisi rivela una rete scale-free e altamente frammentata in migliaia di "isole" tematiche. L'analisi delle community identifica tre archetipi di aggregazione: 1) cluster di convergenza fattuale, 2) arene di dibattito polarizzato, 3) echo chamber ideologiche. Si conclude che, oltre a cluster guidati da pratiche editoriali condivise, l'analisi fornisce prova quantitativa della formazione di echo chamber, specialmente nella politica interna. La segregazione ideologica si manifesta sia nel framing delle notizie che nella selezione degli argomenti da amplificare.
Keywords: Social Network Analysis, Echo Chamber, Panorama Mediatico, Similarità Semantica, Analisi delle Comunità, Polarizzazione, Giornalismo Italiano.

1 Introduzione

Nell'odierno ecosistema dell'informazione, caratterizzato da un'enorme abbondanza di fonti e da un'accelerata polarizzazione del dibattito pubblico, i concetti di "filter bubble" e "echo chamber" sono diventati di fondamentale importanza. Questi fenomeni descrivono ambienti informativi isolati in cui gli individui sono esposti prevalentemente a opinioni e linguaggi che confermano le proprie convinzioni preesistenti, limitando il confronto con prospettive divergenti. L'obiettivo di questo progetto è quello di indagare questo fenomeno cercando di costruire una rete che modella il panorama mediatico italiano tramite gli strumenti della Social Network Analysis. Cercherò quindi di rispondere ad una domanda precisa:

La somiglianza nel linguaggio utilizzato negli articoli di testate giornalistiche rivela la formazione di cluster ideologici o echo-chamber?

2 Data Collection

Durante la fase di collezione di dati ho utilizzato la versione gratuita del dataset NewsAPI che mi permetteva di effettuare fino a cento richieste al giorno. Le richieste, inoltre, possono riguardare solo gli articoli degli ultimi 30 giorni. All'interno del codice ho selezionato le seguenti fonti: Ansa, Corriere, La Repubblica, La Stampa, Il Sole 24 Ore, ADN Kronos, Il Fatto Quotidiano, Il Messaggero, Il Giornale, Avvenire, Internazionale, Il Mattino, Il Gazzettino, Unione Sarda, Il Secolo XIX, GDS (Giornale Di Sicilia), Quotidiano Nazionale, Libero Quotidiano, HuffPost e FanPage.

La ricerca tramite API ha restituito un totale di 32900 articoli. L'obiettivo è ora quello di costruire una rete in cui gli articoli rappresentano i nodi e gli archi vengono invece rappresentati dalla similarità semantica, cioè la similarità del linguaggio, tra

due articoli. Per evitare una rete molto densa, però, prendo due precauzioni:

1. Non utilizzo tutti gli articoli trovati durante la ricerca ma li filtro per quattro categorie principali (Politica, Economia, Cronaca e Guerra) attraverso l'uso di parole chiave: in questo modo ottengo che:

   - "Politica" contiene 10510 articoli
   - "Economia" contiene 9188 articoli
   - "Cronaca" contiene 3710 articoli
   - "Guerra" contiene 2885 articoli

   Mi concentro quindi solo sugli articoli che appartengono alla categoria "Politica" in maniera da poter fare un'analisi più precisa dei dati.

2. Stabilisco una soglia di similarità semantica che garantisce che due articoli vengano connessi tra loro solo se sono effettivamente molto simili, evitando così di avere una rete super densa e che mi porterebbe a trovare cluster che non esistono.

3 Network Analysis

Alla fine della fase di "pulizia dei dati", si ottiene una rete con 10510 nodi e 4321 archi. Come detto nella prima parte, il numero degli archi è determinato dalla similarità semantica tra due articoli che viene determinata tramite il confronto di parole chiave. L'algoritmo usato per individuare queste parole chiave è il TF-IDF [1] che sta per term frequency-inverse document frequency. Si tratta di un algoritmo che misura l'importanza di una keyword all'interno di una pagina: calcola un punteggio che tiene conto della frequenza con cui una parola compare in un documento e tiene conto di quanto quella parola sia rara in tutto il corpus. Inoltre, per assicurare una certa similarità semantica tra gli articoli, è stata specificata una soglia di similarità pari al 0,6 che fa in modo che due articoli siano collegati tra loro solo se raggiungono questa soglia. Per migliorare l'accuratezza dell'analisi di similarità semantica avremmo potuto usare BERT o semantic BERT (sBERT). È stato usato TF-IDF per motivi di risorse computazionali e di semplicità di processo.

3.1 Degree Distribution Analysis

Analizziamo la distribuzione del grado per verificare quale grado hanno gli articoli all'interno della rete creata. Osservando le immagini si nota che la rete è composta da un picco enorme di articoli con grado molto basso. Nel contesto della rete mediatica, questo fenomeno potrebbe essere letto come il fatto che la stragrande maggioranza di articoli all'interno del dataset sono abbastanza "standard", cioè sono semanticamente simili solo ad un piccolo gruppi di articoli. Potremmo interpretare questo dato come l'indizio che la maggior parte degli articoli di politica non sono così influenti o unici da essere "citati" semanticamente da molti altri.

La coda lunga nella destra del grafico indica che ci sono pochissimi articoli con un grado alto. Questi rappresentano gli hub della nostra rete: sono gli articoli più importanti.

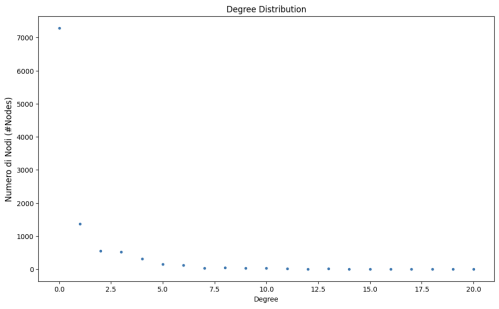

Figura 1: Distribuzione del grado (scala normale)

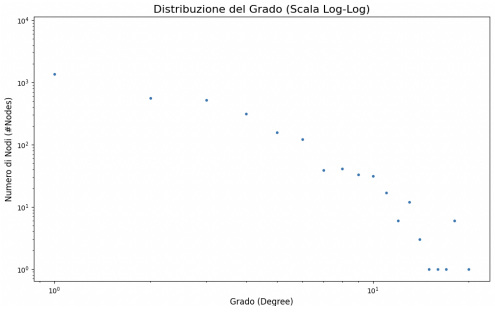

Figura 2: Distribuzione del grado (scala logaritmica)

3.2 Connected component analysis

Su un totale di 10510 nodi, la rete si frammenta in 8426 componenti connesse distinte. Questo conferma i risultati ottenuti dall'analisi della distribuzione del grado: la maggior parte degli articoli costituisce una componente a sè stante. Quindi, nonostante il grande numero di componenti, la rete si presenta come una rete sparpagliata con molte "isole semantiche" differenti. Infatti, l'analisi delle 5 componenti più grandi si presenta così:

- La componente gigante è formata da 24 nodi che costituiscono lo 0.23% dei nodi totali della rete
- La seconda componente in ordine di grandezza è formata da 20 nodi, cioè lo 0.19% totale dei nodi
- La terza e la quarta sono formate da 18 nodi, cioè lo 0.17% dei nodi totali
- La quinta componente è formata da 17 nodi, cioè lo 0.16% dei nodi totali

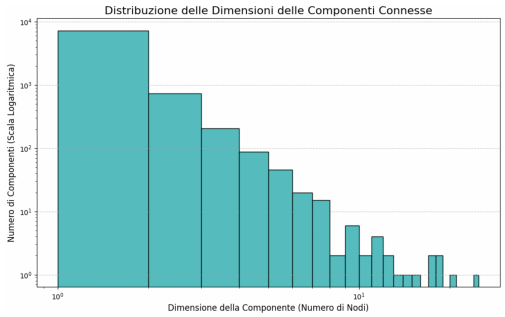

Figura 3: Le prime 10 componenti per dimensione

La sparsità della rete ha senso perché ogni componente si riferisce tendenzialmente ad articoli che affrontano uno stesso argomento. Si deduce quindi che ci sono articoli più "importanti" che, proprio per la loro rilevanza, vengono affrontati da diverse testate. La componente gigante, quindi, corrisponde alla singola notizia più grande e con la maggiore risonanza mediatica nel dataset.

3.3 Analisi dei cammini

Considerando la sparsità della rete dimostrata anche dall'analisi delle componenti, sarebbe poco rilevante fare un'analisi dei cammini sull'intera rete perché ci sono troppe isole semantiche scollegate dal resto della rete. Analizziamo quindi i cammini della componente gigante:

- Il cammino medio minimo è uguale a 1.96. Questo significa che, in media, per collegare due articoli scelti casualmente, non servono nemmeno 2 salti
- Il diametro della rete è pari a 5. Questo vuol dire che la massima distanza tra i due articoli semanticamente più diversi nel cluster sarà pari a 5

Questi risultati rivelano una struttura di rete estremamente compatta, quasi con proprietà delle reti "ultra-small-world". Questo implica che gli articoli all'interno di questo cluster sono semanticamente quasi sovrapponibili e formano un nucleo tematico incredibilmente denso.

3.4 Densità

La rete ha una densità di $7.8e^{-5}$. Si tratta di un numero bassissimo che indica che, di tutti i collegamenti possibili nella rete, solo una frazione infinitesimale esiste realmente. Questo risultato è un'ulteriore prova della sparsità della rete, confermando che non esiste un discorso mediatico unificato. Al contrario, il panorama semantico è costituito da piccoli gruppi di articoli densamente connessi al loro interno ma quasi completamente isolati l'uno dall'altro.

3.5 Coefficiente di clustering

Il coefficiente di clustering è una misura che ci dice "quanti triangoli si chiudono nella rete". Cioè, si tratta di una misura che calcola quanti vicini all'interno di una rete sono connessi tra loro. Il valore di clustering che otteniamo dalla rete è di 0.143: un valore relativamente "alto" rispetto alla densità bassissima. Questa combinazione di bassa densità e alto clustering è tipica delle reti complesse che modellano

il mondo reale e delle strutture "small-world". Nel nostro contesto si tratta di un buon risultato perché dimostra che, sebbene gli algoritmi siano semanticamente isolati l'uno dall'altro (bassa densità), all'interno di ogni argomento gli articoli fanno cluster coesi e auto-referenziali.

3.6 Centrality Analysis

Per l'analisi della centralità ho utilizzato tre metriche:

1. Degree Centrality: misura il numero di connessioni dirette di un nodo
2. Betweenness Centrality: misura la frequenza con cui un nodo si trova sui percorsi più brevi tra altre coppie di nodi
3. PageRank: analizza come i nodi sono collegati tra loro. L'idea è che se un nodo N è connesso ad un nodo X più "importante", allora anche il nodo N sarà reso più importante da questo collegamento

Dall'analisi è risultato che tutte e tre le classifiche sono dominate da comunicati stampa proveniente da un'unica testata: Il Giornale. Questo indica che in questa rete la centralità non è determinata dal dibattito politico ma dalla standardizzazione del testo.

- **Degree centrality:** gli articoli con più alta Degree Centrality sono quelli che rappresentano il nucleo di una notizia specifica. Dai dati raccolti, in particolare, sembra che si tratti spesso di dichiarazioni politiche riprese da più fonti. I nodi centrali, infatti, corrispondono a diverse testate che riportano la stessa dichiarazione del ministro Tajani. In questo contesto, l'importanza di un articolo è legata alla ripetizione e alla sincronizzazione del sistema mediatico nel riportare un'informazione specifica
- **Betweenness Centrality:** questa metrica identifica come "ponti" gli articoli relativi a eventi di cronaca con forti implicazioni politiche, come lo sgombero del Leoncavallo. Questi articoli fungono da ponti perché sono semanticamente simili sia agli articoli di cronaca fattuale sia a quelli di commento ideologico, collegando discorsi altrimenti distanti. La loro importanza, quindi, risiede nella loro capacità di unire diverse narrazioni
- **PageRank:** gli articoli con PageRank più alto sono quindi articoli che si trovano al centro dei dibattiti. Questi nodi non si limitano a riportare una notizia, ma ne definiscono l'inquadramento e sono semanticamente connessi a molti altri articoli importanti all'interno dello stesso cluster tematico-ideologico. Temi come la gestione dei migranti, le dichiarazioni politiche e gli eventi culturali di rilievo generano i nodi con il PageRank più alto, indicando che la loro influenza deriva dall'essere al centro di una conversazione attiva

3.7 Confronto con Modelli Teorici

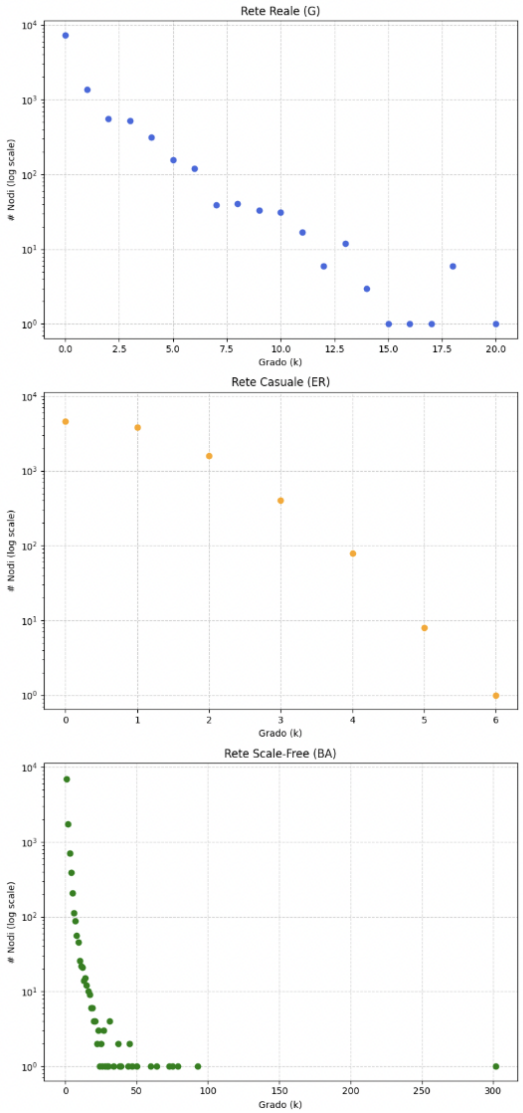

Figura 4: Confronto tra la mia rete (rete reale), rete random e rete scale-free

Dalla figura 4 si evince che:

- La nostra rete (a sinsitra) mostra una distribuzione a coda lunga, come detto precedentemente. Questo è il segno di una rete non casuale dove alcuni elementi (gli hub) sono più "importanti" di altri
- La rete random (al centro) mostra una distribuzione molto diversa da quella della rete reale: la distribuzione ha la forma di una campana stretta e bassa, cioè si tratta di una distribuzione Poissoniana concentrata attorno a un grado medio molto basso. In questa rete non ci sono hub
- La distribuzione della rete scale-free (a destra) è molto simile a quella della rete reale. Anche qui si osserva infatti una distribuzione a coda lunga che conferma il meccanismo di preferential attachment per cui gli articoli importanti attirano più connessioni

Si può quindi affermare che il confronto visivo delle distribuzioni del grado nella figura 4 evidenzia la natura non casuale della rete semantica. A differenza del modello ER, con distribuzione Poissoniana, la rete reale mostra una distribuzione a legge di potenza, quasi identica a quella generata dal modello BA. Questo conferma la presenza di una struttura scale-free, dove pochi articoli hub agiscono come centri semantici per la maggior parte degli altri articoli.

Rete	# Nodi	# Archi	# Componenti	Dim. Comp. Gigante
Reale (G)	10510	4321	8426	24
Casuale (ER)	10510	4290	6221	64
Scale-Free (BA)	10510	10509	1	10510

Rete	Coeff. Clustering Medio(LCC)	Cammino Medio (LCC)
Reale (G)	0.6874	1.9565
Casuale (ER)	0.0000	10.1419
Scale-Free (BA)	0.0000	NaN

Rete	Diametro (LCC)
Reale (G)	5.0
Casuale (ER)	26.0
Scale-Free (BA)	NaN

Figura 5: Confronto con le metriche dei modelli teorici

La differenza principale tra i modelli emerge dall'analisi delle componenti. Come si può osservare dalla Figura 5, mentre il modello BA crea un'unica rete

interconnessa, con una componente gigante che corrisponde al numero di nodi totali della rete, la mia rete è altamente frammentata. Questa frammentazione appare come una caratteristica fondamentale del dominio che analizzo: le conversazioni mediatiche, se analizzate con un'alta soglia di similarità, non formano un unico discorso, ma si dividono in migliaia di "isole tematiche" distinte, ciascuna rappresentante un singolo evento o notizia. Il modello BA, pur catturando la dinamica degli hub, non riesce a rappresentare questo fenomeno di isolamento tematico.

Un'altra differenza viene riscontrata nel coefficiente di clustering (Figura 5). Il coefficiente medio della rete reale (coefficiente di clustering della componente gigante) è molto alto, indicando che gli articoli all'interno di un singolo evento mediatico sono iperconnessi e formano una bolla quasi completa. Al contrario, i modelli ER e BA hanno un clustering praticamente nullo. Questo dimostra che la rete reale costruita ha una struttura locale molto più forte di quanto i modelli teorici possano prevedere.

4 Open question: struttura semantica del panorama mediatico italiano

In base a quanto detto nell'introduzione e ai risultati delle analisi precedenti, mi sono chiesta se questo fenomeno delle echo-chamber o di cluster ideologici può effettivamente essere individuato tramite un'analisi delle comunità interne alla rete. L'analisi delle community ha rivelato una frammentazione della rete in oltre 8mila piccoli cluster molto coesi. Per avere un'idea dei processi interni alla rete ho individuato le prime dieci comunità per dimensione.

L'analisi qualitativa delle 10 community più grandi rivela che il panorama mediatico non si divide secondo semplici linee ideologiche, ma si aggrega secondo tre archetipi principali, ognuno con una diversa dinamica di formazione.

4.1 Echo-chamber (non ideologica) del "Copia-Incolla"

Questo archetipo include i cluster che si formano attorno a eventi di cronaca o politica internazionale dove i fatti sono predominanti. La composizione di queste community è tipicamente eterogenea e spesso dominata dalle agenzie di stampa, indicando una convergenza del linguaggio giornalistico piuttosto che una divisione ideologica. Ne fanno parte tre comunità:

- la numero 7654 formata da 13 nodi che corripsondono ad articoli provenienti prevalentemente dall'agenzia di stampa ANSA.it (30.77%), seguita da testate generaliste locali e nazionali. Questo dimostra come, su notizie fattuali, le agenzie fungano da fonte primaria, creando un'alta omogeneità semantica tra i vari articoli che ne riprendono le informazioni
- La numero 1532 composta da 13 nodi. Trattandosi di una notizia di politica economica internazionale (Dazi USA), mostra una copertura trasversale che include testate come Fanpage, Il Fatto Quotidiano e il Giornale.it, dimostrando un interessa condiviso per la notizia
- La numero 979, che tratta degli scontri politci tra Italia e Francia ed è composta da 17 nodi. Anche in questo caso, la notizia viene coperta prevalentemente da agenzie di stampa come Adnkronos, QuotidianoNet, il Giornale.it e ANSA.it indicando la formazione di un cluster che si basa sulla notizia stessa

Quindi, in questo archetipo troviamo delle bolle informative perfette ma non di natura ideologica. Invece, questi articoli sono connessi da una forte dipendenza da fonti condivise e dalla necessità di riportare gli stessi fatti.

4.2 La convergenza mediatica

Questo archetipo descrive community relative a notizie altamente divisive. A differenza di una echo chamber, qui testate di schieramenti opposti entrano nello stesso cluster semantico, creando una sorta

di "arena" dove si dibbatte sullo stesso campo di battaglia lessicale, anche se con framing diversi. Le tre comunità principali sono:

- La comunità 6488 relativa al Caso Almasri e formata da 18 nodi. Qui, la presenza fianco a fianco di testate come Il Fatto Quotidiano e Libero dimostra che, pur partendo dagli stessi fatti e condividendo l'attenzione centrale (il video), il dibattito si è sviluppato lungo linee ideologiche che i dati raccolgono dentro la stessa arena semantica

- La comunità 5991, formata da 15 nodi, dove un evento diplomatico ad alta tensione (scontro Salvini-Macron) viene coperto da un mix di testate, indivando che il tema è centrale nel dibattito pubblico e unendo gli attori mediatici attorno ad uno stesso campo lessicale di partenza

- La comunità 5691 formata da 20 nodi. Anche qui, si tratta di un evento di politica interna raccontato da un ampio numero di testate che si concentrano tutte su un'unica affermazione ("Forza idee vince sempre su violenza delle parole"

Questi cluster dimostrano che la polarizzazione non sempre porta a parlare di cose diverse, ma spesso a parlare della stessa cosa in modo conflittuale all'interno dello stesso cluster

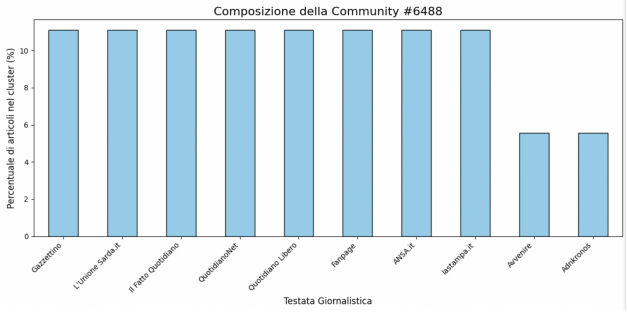

Figura 6: Distribuzione delle testate nella comunità 6488

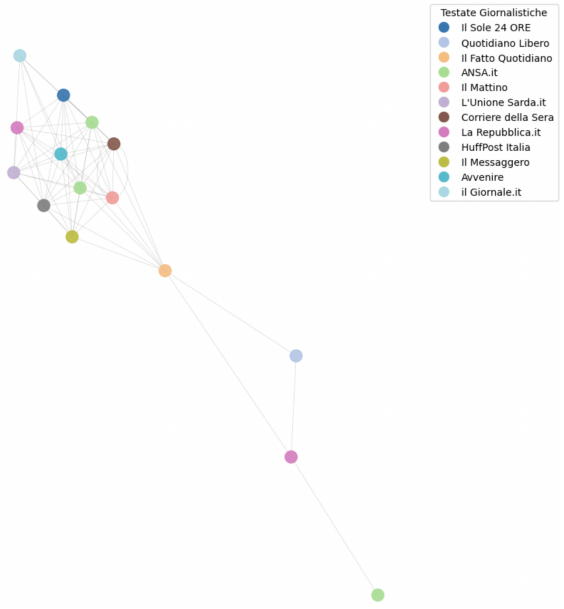

Figura 7: Visualizzazione grafica della community 5991

Quindi, la somiglianza che forma il cluster non deriva da un'opinione condivisa ma dalla necessità di usare lo stesso vocabolario di base (nomi dei protagonisti, luoghi, fatti scatenanti) per partecipare al dibattito.

4.3 Echo Chamber Ideologica

È in questo terzo archetipo che l'analisi identifica le prove più forti della formazione di "bolle" informative. Si tratta di community dominate da un gruppo ristretto e ideologicamente omogeneo di testate, che trattano temi cari alla loro linea editoriale. Gli esempi più lampanti di questo fenomeno sono due:

- La comunità 3735, formata da 17 nodi e composta quasi esclusivamente da articoli prodotti da testate di centro-destra e regionali allineate. I titoli si concentrano su temi programmatici del governo (come la giustizia, i migranti e la politica estera) con un framing positivo. La quasi totale assenza di voci di opposizione indica la creazione

di una bolla mediatica in cui un discorso specifico viene trattato da una sola ideologia politica

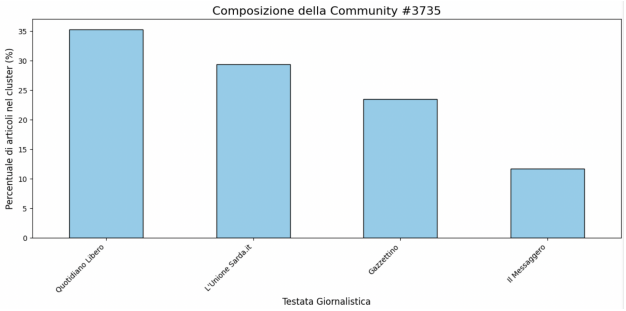

Figura 8: Visualizzazione grafica della community 3735

- La comunità 6833, formata da 18 nodi. Questa comunità, pur avendo una base di cronaca, mostra una forte inclinazione a destra. La maggior parte dei titoli con una forte connotazione ideologica ("Meloni fa impazzire la sinistra", "Salvini ...'Afuera!'") provengono da testate come Libero, che dominano la narrazione ideologica all'interno del cluster, spingendolo verso questo archetipo.

5 Conclusioni

L'obiettivo di questa analisi era determinare se fosse possibile mappare la formazione di cluster ideologici ed echo chamber nel panorama mediatico italiano attraverso una rete basata sulla similarità semantica degli articoli. I risultati confermano questa ipotesi, rivelando una struttura complessa.

L'analisi delle community ha dimostrato che il discorso mediatico non si organizza in un unico flusso, ma si frammenta in migliaia di isole tematiche. L'analisi qualitativa di queste isole ha permesso di identificare tre archetipi principali di aggregazione:

- **Echo Chamber del "Copia-Incolla":** su eventi di cronaca o politica internazionale non divisivi, le testate convergono attorno a un linguaggio comune, spesso dettato dalle agenzie di stampa. In questo caso, la similarità semantica indica una pratica editoriale condivisa, non un'affinità ideologica
- **Convergenza mediatica:** su temi ad alto impatto e controversi, testate di orientamento opposto entrano nella stessa "arena" semantica. Questo conferma che la polarizzazione non sempre porta a parlare di argomenti diversi, ma spesso a dibattere sullo stesso tema, usando un vocabolario di base condiviso
- **Echo Chamber Ideologica:** rappresentano la scoperta più significativa perché consistono in cluster tematici dominati da un gruppo omogeneo di testate. La community 3735, focalizzata su dichiarazioni programmatiche del governo e composta quasi esclusivamente da fonti di centro-destra, rappresenta un chiaro esempio di echo chamber. In questo caso, la similarità semantica non deriva solo dal tema, ma da una selezione e un framing condivisi che escludono voci alternative

In conclusione, questa analisi ha dimostrato che la similarità lessicale è uno strumento efficace per mappare le fratture del discorso mediatico. Inoltre, possiamo affermare con più sicurezza che l'esistenza di echo chamber si manifesta non solo nel modo in cui un argomento viene presentato ma anche nella scelta stessa di quali argomenti e quali dichiarazioni amplificare. Con l'approccio usato è stato possibile identificare cluster guidati sia da pratiche editoriali neutre sia da affinità ideologiche.

Per approfondire ulteriormente l'analisi si potrebbero usare tecniche di Sentiment Analysis all'interno dei cluster di convergenza mediatica, così da quanatificare numericamente le differenze di tono tra testate opposte che affrontano uno stesso argomento.

Riferimenti bibliografici

[1] G. Salton e C. Buckley. "Term Weighting Approaches in Automatic Text Retrieval". In: *Information Processing & Management* 24.5 (1988), pp. 323–328. DOI: 10.1016/0306-4573(88)90021-0.