Cora Citation Network

Network Analysis di una rete di citazioni accademiche in ambito Machine Learning

Progetto finale Advanced Data Scince

Edoardo Diana

Indice

Domande di ricerca

Network quick view

1 : Qual è il numero di paper in ogni categoria ?

2 : Qual è la distribuzione del numero di citazioni fatte dai paper ?

3 : Per ogni catgoria, quante sono le citazioni verso la stessa categoria ?

4.1 : Quali sono le top coppie di paper con più paper citati in comune ?

4.2 : Quali sono le top coppie di paper con più paper citanti in comune ?

Centrality

5 : Analizza i top paper per PageRank e scopri se tale valore è principalmente causato

6 : Analizza Authority ed Hub score per capire quali paper risultano maggiormente influenti.

7 : Identifica categorie potenti che non sono centrali e categorie centrali che non sono potenti.

8 : Visualizzare la rete tra categorie con la dimensione dei nodi proporzionale alla power e alla centralità, colore archi proporzionale al flusso di citazioni.

Similarità ed Eterogeneità

9 : I paper con alta centralità tendono ad essere interdisciplinari ?

10 : I paper interdisciplinari (alto Rao da out-degree) sono più influenti (alto in-degree/PageRank) ?

11 : Quali categorie producono i paper più interdisciplinari ? E quali beneficiano di più dai ponti interdisciplinari?

Group Analysis

12 : Identifica le comunità di paper e verifica se esse coincidono con le categorie scientifiche (ground truth).

13 : Quali categorie sono più pure ? (una sola community o meno possibili)

Global Analysis

14 : Studia i K-connected components e verifica se i gruppi con connettività più alta sono composti da soli paper di una stessa catgoria.

15 : Analizza gli SCC per verificare l'aciclicità della rete.

16 : Verifica lo Small-World Effect sul grafo non diretto.

17 : Verifica che la Assortativity by category sia molto alta (dato che ho tante citazioni intra-categoria)

18 : Verifica se l’elite delle pubblicazioni crea un “club” chiuso

Network quick view

Domanda 1 : Qual è il numero di paper in ogni categoria ?

La rete è un grafo Diretto Non Pesato composto da :

- 2708 paper (nodi)
- 5429 citazioni (archi)

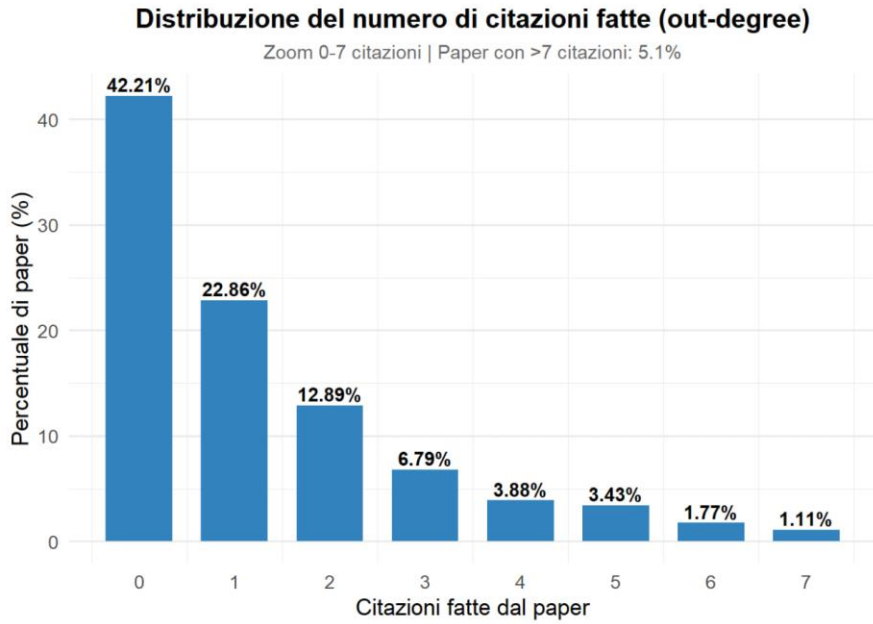
I paper appartengono ad 1 categoria (di 7) nell'ambito Machine Learning.

	category	numero_paper
	<chr>	<int>
1	Neural_Networks	818
2	Probabilistic_Methods	426
3	Genetic_Algorithms	418
4	Theory	351
5	Case_Based	298
6	Reinforcement_Learning	217
7	Rule_Learning	180

Domanda 2 : Qual è la distribuzione del numero di citazioni fatte dai paper ?

Nel grafico viene mostrata la percentuale dei nodi che hanno un dato valore per out-degree centrality.

Vediamo come quasi la metà dei paper in questa rete non citi nessun altro paper (magari perché tra i più datati tra quelli presenti).



Domanda 3 : Per ogni categoria, quante sono le citazioni verso la stessa categoria ?

	from_cat	citazioni_intra	citazioni_totali	perc_intra
	<chr>	<int>	<int>	<dbl>
1	Genetic_Algorithms	848	935	90.7
2	Neural_Networks	1220	1463	83.4
3	Probabilistic_Methods	696	840	82.9
4	Case_Based	427	521	82.0
5	Rule_Learning	255	313	81.5
6	Reinforcement_Learning	414	535	77.4
7	Theory	558	822	67.9

Numero di volte che un paper, di una certa categoria, cita un paper che sta nella stessa categoria.

Citazioni intra-disciplinari: 4418
Citazioni totali: 5429
Percentuale citazioni intra-disciplinari: 81.38%

Genetic_Algo ha la percentuale maggiore, Theory la minore.

Domanda 4.1 : Quali sono le top coppie di paper con più paper citati in comune ? (common successors)

Per farlo, uso le proiezioni della matrice di adiacenza.

	paper_i	paper_j	shared
1	114	6213	20
2	35	82920	15
3	4584	6213	13
4	2658	2665	12
5	19621	1365	12

`A * t(A)`

Domanda 4.2 : Quali sono le top coppie di paper con più paper citanti in comune ? (common predecessors)

	paper_i	paper_j	shared
1	63832	1104999	5
2	1154123	1154124	5
3	31349	686532	4
4	193742	6155	4
5	124064	6155	4

`t(A) * A`

Centrality

Calcolo dapprima queste misure di centralità : indegree, outdegree e pagerank.

Domanda 5 : Analizza i top paper per PageRank e scopri da cosa tale valore è principalmente causato

La PagRank Centrality è ottenuta sulla base di

- numero di link che il nodo riceve (in-degree)
- centralità dei nodi da cui riceve i link
- quanto i nodi da cui si ricevono i link sono propensi a linkar altri nodi

name <chr>	category <chr>	pagerank <dbl>	indegree <dbl>	outdegree <dbl>	avg_pr_citing <dbl>	max_pr_citing <dbl>	avg_outdegree_citing <dbl>
683355	Probabilistic_Methods	0.004771088	5	1	0.0012547852	0.004582902	2.60
683404	Probabilistic_Methods	0.004582902	4	1	0.0013824136	0.004771088	2.50
39210	Neural_Networks	0.003490741	4	2	0.0012214163	0.001831824	2.50
578347	Neural_Networks	0.003442287	5	1	0.0009574864	0.003310904	10.00
578309	Neural_Networks	0.003310904	5	1	0.0009286370	0.003442287	10.40
32698	Probabilistic_Methods	0.003267067	4	1	0.0009889463	0.002974560	5.75
289085	Neural_Networks	0.003228009	4	1	0.0010460292	0.003204155	2.25
689152	Neural_Networks	0.003204155	3	1	0.0012292476	0.003228009	2.00
9513	Theory	0.003116922	5	1	0.0007550796	0.002901496	5.00
95719	Probabilistic_Methods	0.002974560	2	1	0.0017294972	0.003267067	15.00

Numero di link che riceve :

- i top 10 papers hanno mediamente un alto indice

Centralità dei nodi da cui riceve i link :

- spesso vi è la presenza di un nodo molto centrale che sposta in alto la media delle citazioni ricevute da quel paper

Quanto i nodi da cui si ricevono i link sono propensi a linkar altri nodi :

- Non tutti, ma solo alcuni paper tra i top per pagerank ricevono citazioni da paper molto propensi a citare.
- Esempio ne sono i paper 578347 e 578309 di Neural_Networks con 10 e 10.4, o 95719 di Probabilistic_Methods con 15 (alto PageRank anche i paper citanti citano molto)

Domanda 6 : Analizza Authority ed Hub score per capire quali paper risultano maggiormente influenti.

	name	category	authority	indegree	citato_da_hub
	<chr>	<chr>	<dbl>	<dbl>	<lgl>
1	1154459	Genetic_Algorithms	1	4	TRUE
2	1152421	Genetic_Algorithms	1	4	TRUE
3	1153280	Genetic_Algorithms	1	4	TRUE
4	1153943	Genetic_Algorithms	0.983	5	TRUE
5	1119708	Genetic_Algorithms	0.960	5	TRUE
6	84021	Genetic_Algorithms	0.958	5	TRUE
7	273152	Genetic_Algorithms	0.949	4	TRUE

	name	category	hub	outdegree	num_cat_diverse_citate	perc_verso_stessa_cat
	<chr>	<chr>	<dbl>	<dbl>	<int>	<dbl>
1	35	Genetic…	1	166	5	92.2
2	82920	Genetic…	0.107	23	2	95.6
3	85352	Genetic…	0.0818	16	1	100
4	1688	Genetic…	0.0653	15	1	100
5	287787	Genetic…	0.0614	10	1	100
6	14062	Genetic…	0.0488	11	1	100
7	210871	Genetic…	0.0469	13	2	92.3
8	41714	Genetic…	0.0380	11	2	90.9
9	12576	Genetic…	0.0348	19	3	79.0
10	103515	Genetic…	0.0315	9	1	100

Authority Score

Nella tabella emerge il fatto che i paper con l'authority maggiore appartengono tutti alla categoria Genetic_Algorithms.

Nella domanda 3 abbiamo visto come tale categoria sia quella percentuale maggiore di citazioni intra-categoria (90.7%).

Questo alto livello di omofilia causa l'amplificazione reciproca degli score hub-authority all'interno della categoria.

Bias di HITS (favorisce le comunità densamente connesse e omofile).

Hub Score

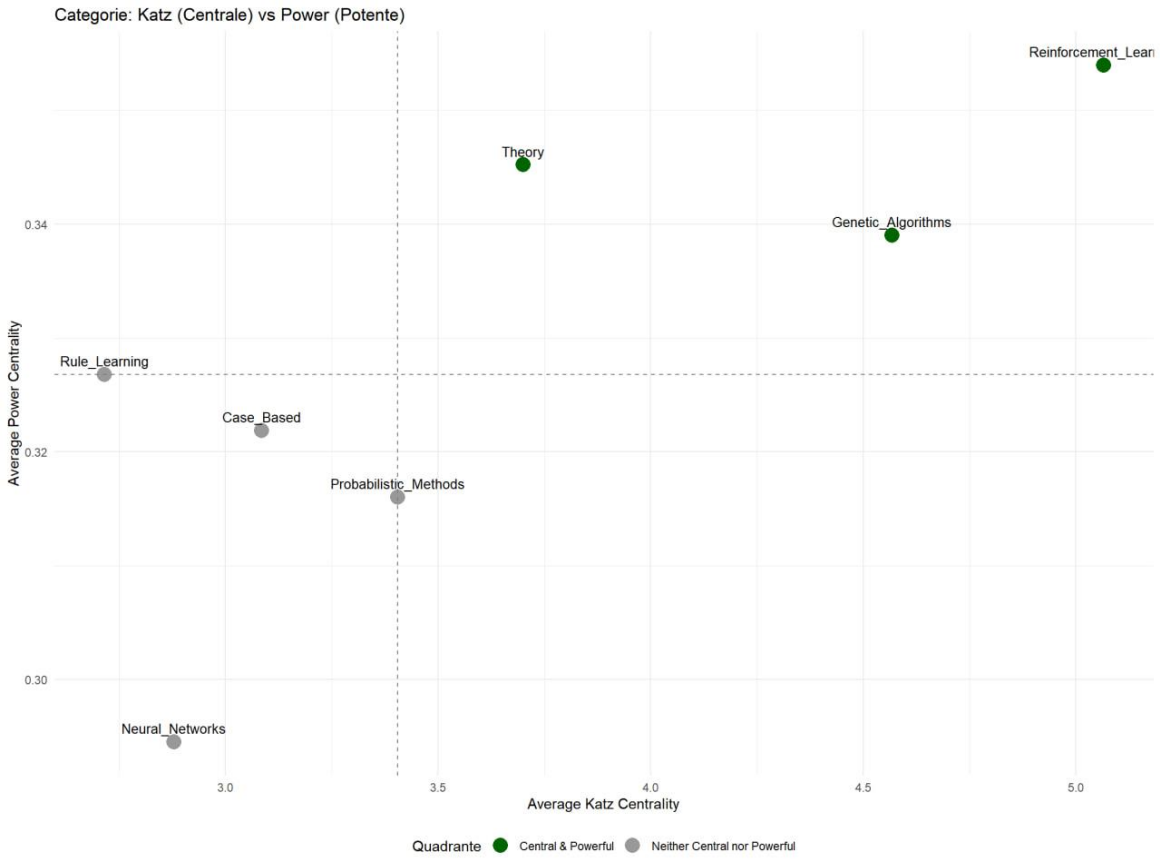
Nota che da analisi precedenti sappiamo che in questa rete ogni paper riceve al massimo 5 citazioni in totale.

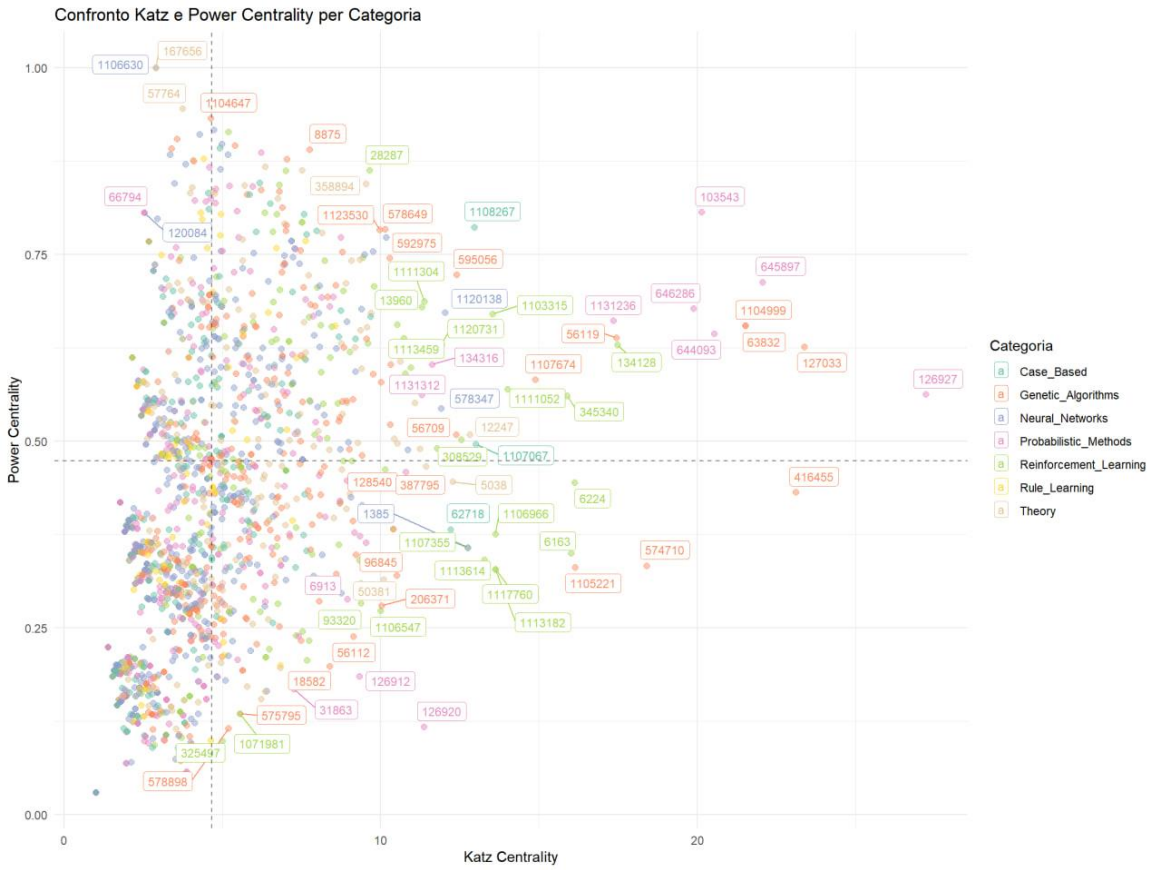
Gli hub linkano autorità nella stessa comunità, creando un circolo chiuso di rinforzo reciproco.

Per un’analisi più accurata si potrebbe :

- Calcolare HITS su sottografi per categoria separatamente
- Filtrare i link intra-categoria prima di calcolare HITS

Domanda 7 : Identifica categorie potenti che non sono centrali e categorie centrali che non sono potenti.



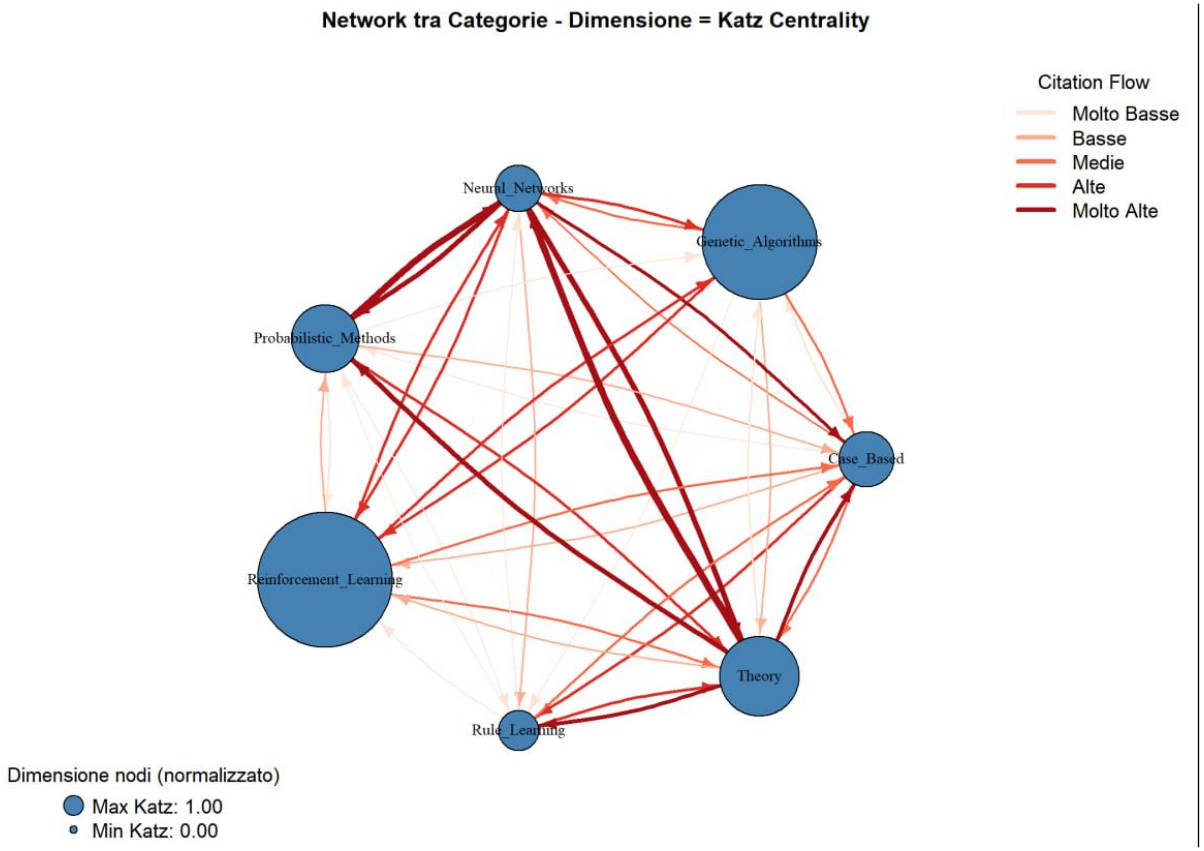


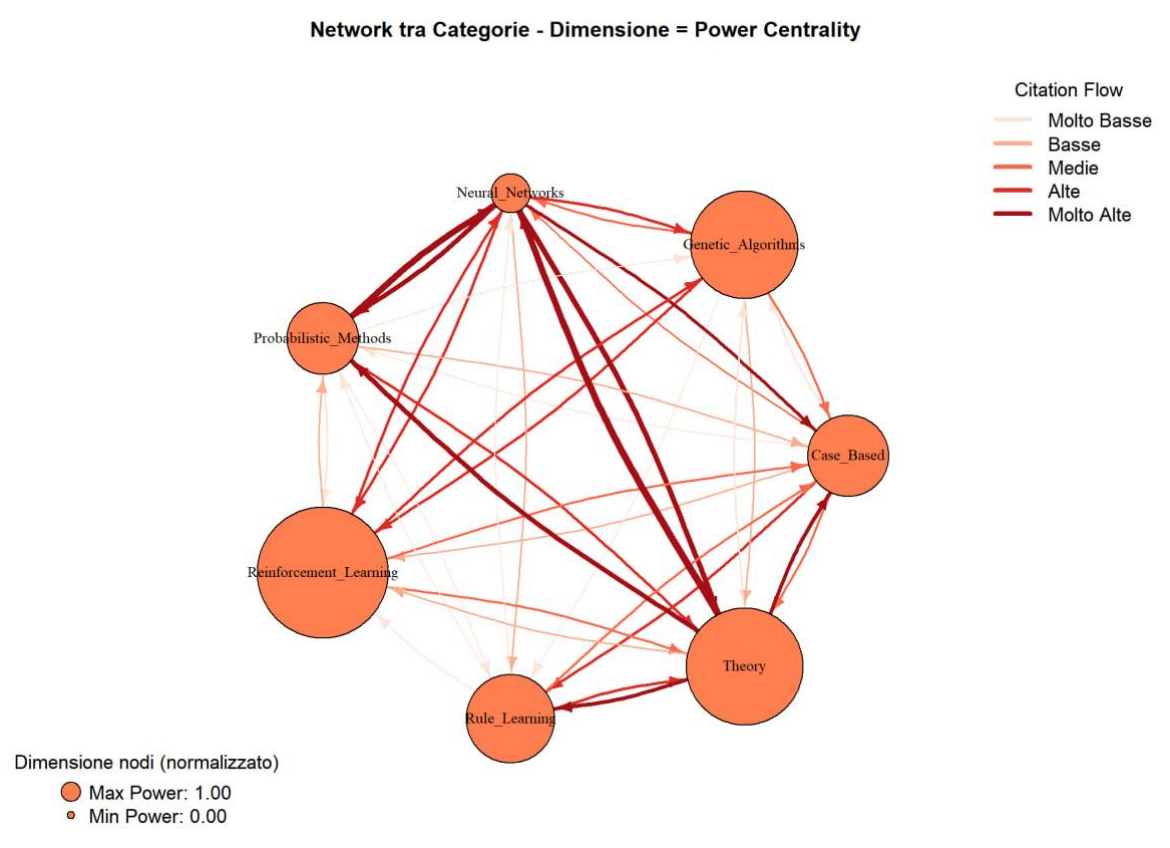
Nessuna delle categorie della rete presenta il comportamento ricercato.

Una categoria “mediamente” potente e per nulla centrale è Rule_Learning. Invece la categoria più centrale possibile rimanendo “mediamente” a bassa power è Probabilistic_Methods.

Spiccano alcuni paper della categoria Probabilistic_Methods come maggiormente centrali e potenti.

Domanda 8 : Visualizzare la rete tra categorie con la dimensione dei nodi proporzionale alla power e alla centralità, colore archi proporzionale al flusso di citazioni.





Ho che Genetic_Alghoritms e Reinforcement Learning son quelli sia più potenti che centrali.

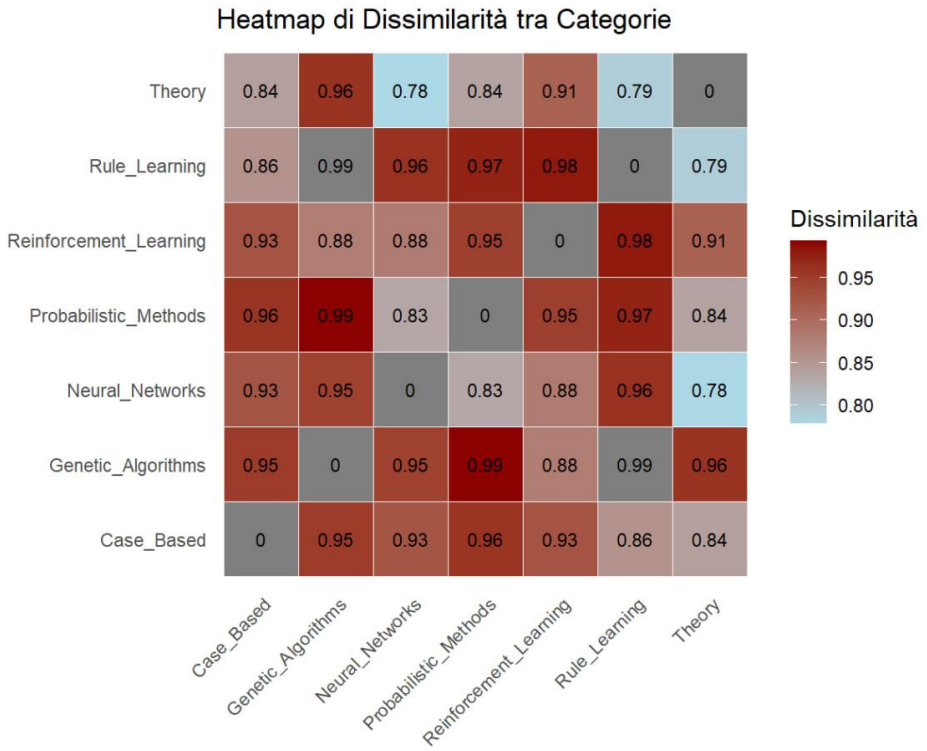
Poi possiamo notare come Neural_Network, pur essendo poco potente e centrale, riceve e invia molte più citazioni rispetto a Genetic_Alghoritms che invece è molto più centrale e potente.

Similarità ed Eterogeneità

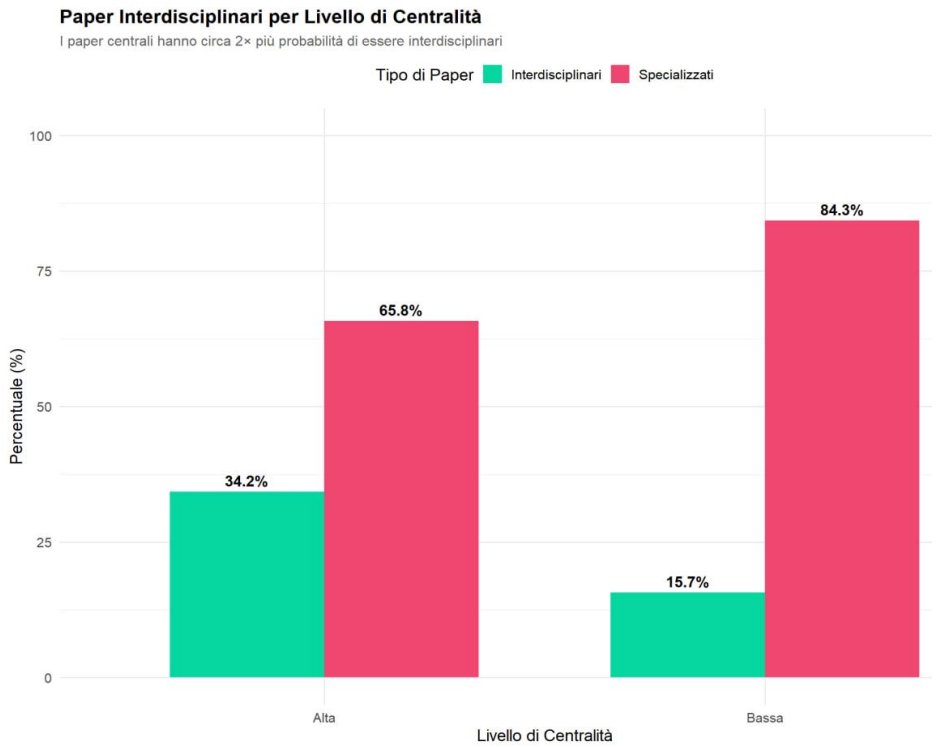
STEP 1 : Creare una matrice di flusso F
STEP 2 : Cosine Similarity tra categorie
STEP 3 : Dissimilarità D tra categorie
STEP 5 : Calcolo RAO per ogni paper
STEP 7 : Calcolo della Betweenness Centrality

L'idea è di misurare quanto è interdisciplinare il pattern di citazioni in uscita.
Rao dunque misura quanto sono diverse tra loro le categorie che il paper i cita.

Domanda 9 : I paper con alta centralità (PageRank o Betweenness) tendono ad essere interdisciplinari ?



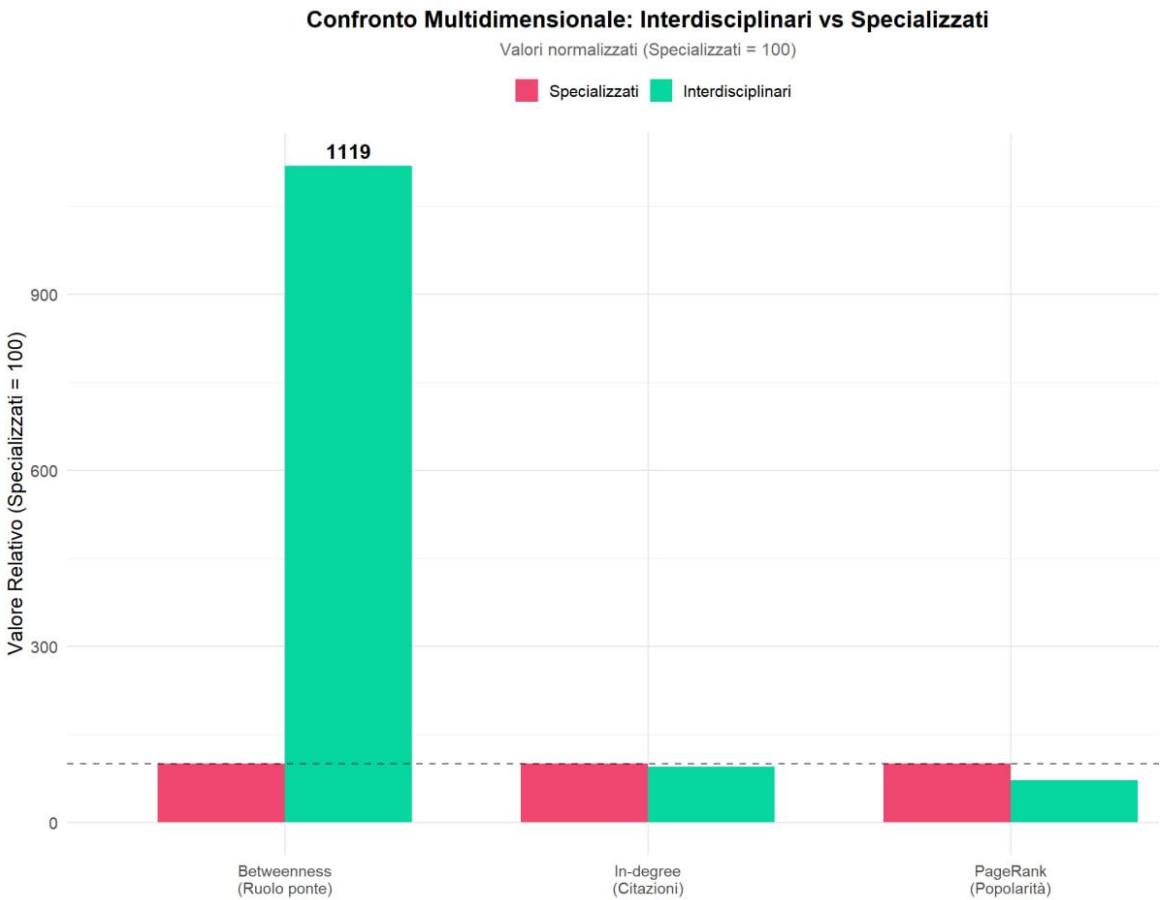
- Alto valore di Dissimilarità media : 0.911
- Le categorie sono mediamente molto dissimili tra loro.
- Ottimo per distinguere bene l'interdisciplinarità dato che le categorie son ben separate (mediante la la Rao Entropy).

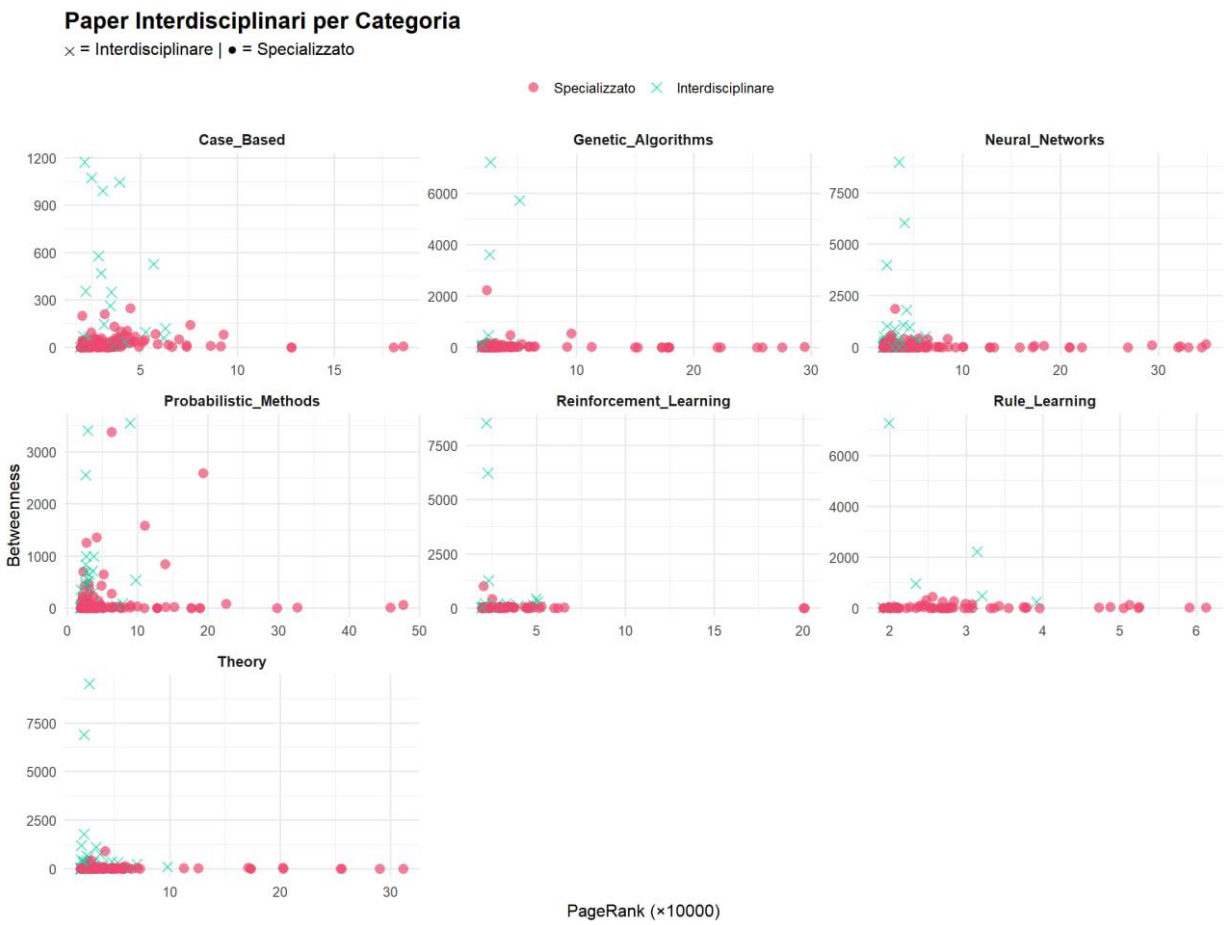


I Paper interdisciplinari (citano categorie diverse) son 300, l'11.08 % dei paper totali.

I Paper ad alta centralità analizzati sono il top 10% per PR o Betw. sono 295.
Di essi, ho che la % con Rao > 0 è di 34.24 %

Domanda 10 : I paper interdisciplinari (alto Rao da out-degree) sono più influenti (alto in-degree/PageRank) ?





I paper interdisciplinari (Rao > 0) hanno una Betweenness Centrality molto più alta.

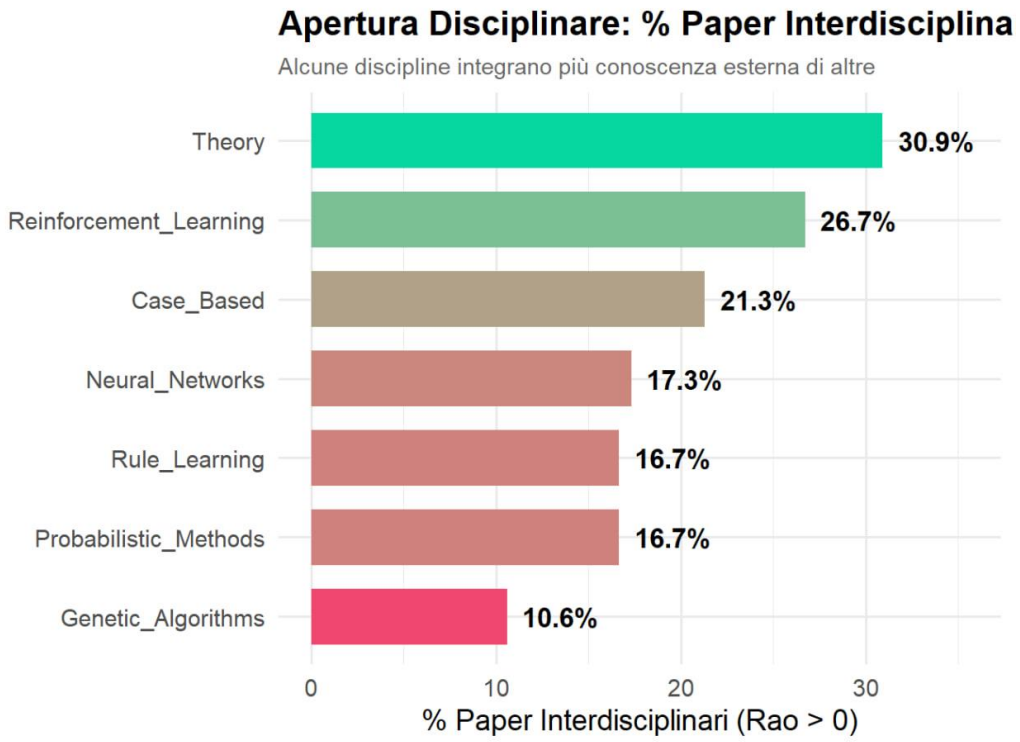
Il che sembra sensato in quanto si troveranno a connettere parti distinte della rete molto più dei paper specializzati.

Questa maggiore influenza dei paper Interdisciplinari è evidenziata solo nel caso della Betweenness Centrality.

Domanda 11 : Quali categorie producono i paper più interdisciplinari ? E quali beneficiano di più dai ponti interdisciplinari?

Considero qui un paper interdisciplinare se cita al di fuori della propria categoria.

Considero invece altri paper che beneficiano da ponti come citati da paper di categoria diversa.



Totale citazioni cross-disciplinari: 874

Ho che la categoria che produce i paper più interdisciplinari è Theory.

Le categorie che beneficiano di più da ponti interdisciplinari sono Theory e Reinforcement_Learning.

category	citazioni_out	citazioni_in	balance
Theory	242	170	72
Reinforcement_Learning	90	81	9
Probabilistic_Methods	123	117	6
Genetic_Algorithms	75	70	5
Neural_Networks	204	214	-10
Rule_Learning	52	91	-39
Case_Based	88	131	-43

Group Analysis

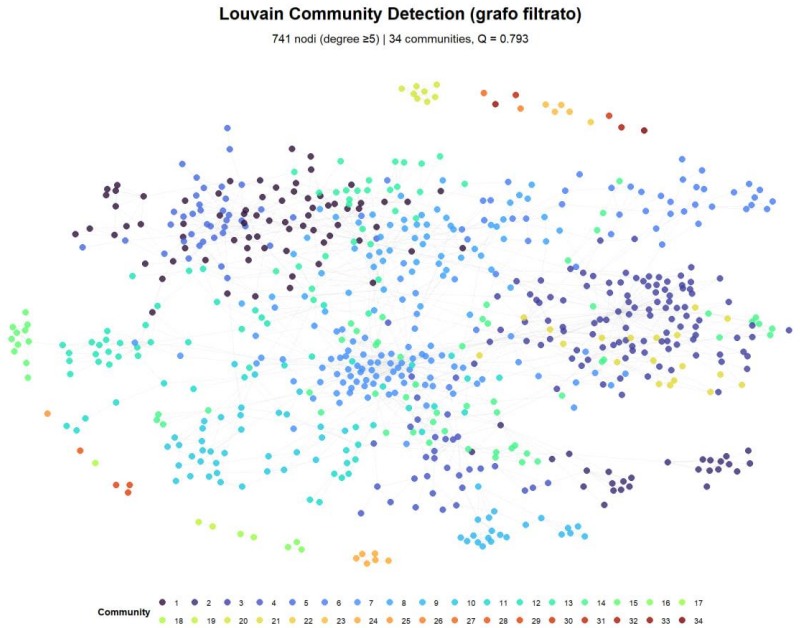
Domanda 12 : Identifica le comunità di paper e verifica se esse coincidono con le categorie (ground truth).
Se no, quale divisione in gruppi risulta esser più adatta alla rete? Quella iniziale (categorie) o quella trovata (communities) ?

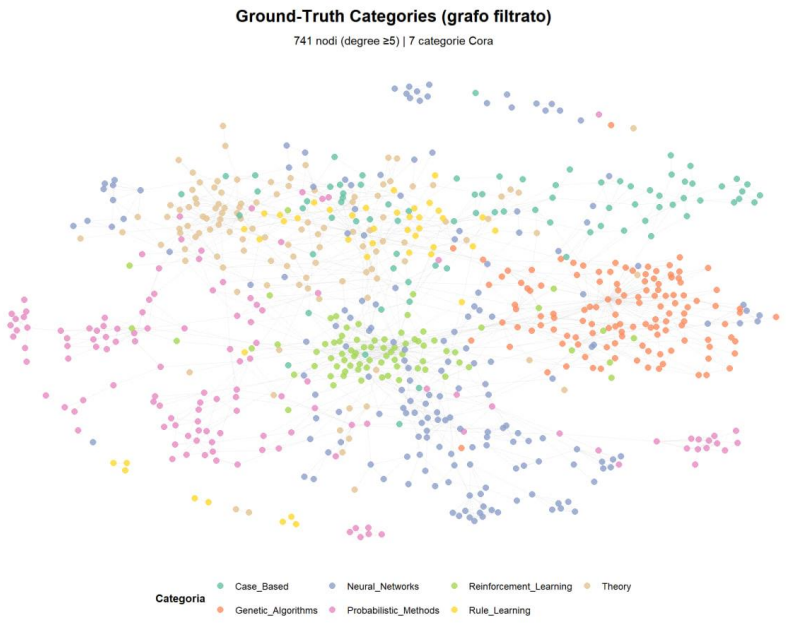
Method	Communities	Modularity	Time_seconds
Louvain	34	0.7926	0.004031
Edge Betweenness	36	0.7904	8.048498
Fast Greedy	35	0.7864	0.004056
Walktrap	57	0.7761	0.021425
Label Propagation	65	0.7584	0.002831
Infomap	90	0.7393	0.363834
Leading Eigenvector	49	0.6627	0.068629

Filtro i nodi per avere una rete meno sparsa e setto una soglia minima di 5 come grado totale (in degree + out degree).

Nodi mantenuti: 741 (27.4%), Archi : 1685

Converto il grafo in non diretto ed applico gli Ehuristic Algorithms, il migliore tra questi è Louvain.





A sinistra ho la rete filtrata in cui sono mostrate le 34 communities trovate da Louvain.

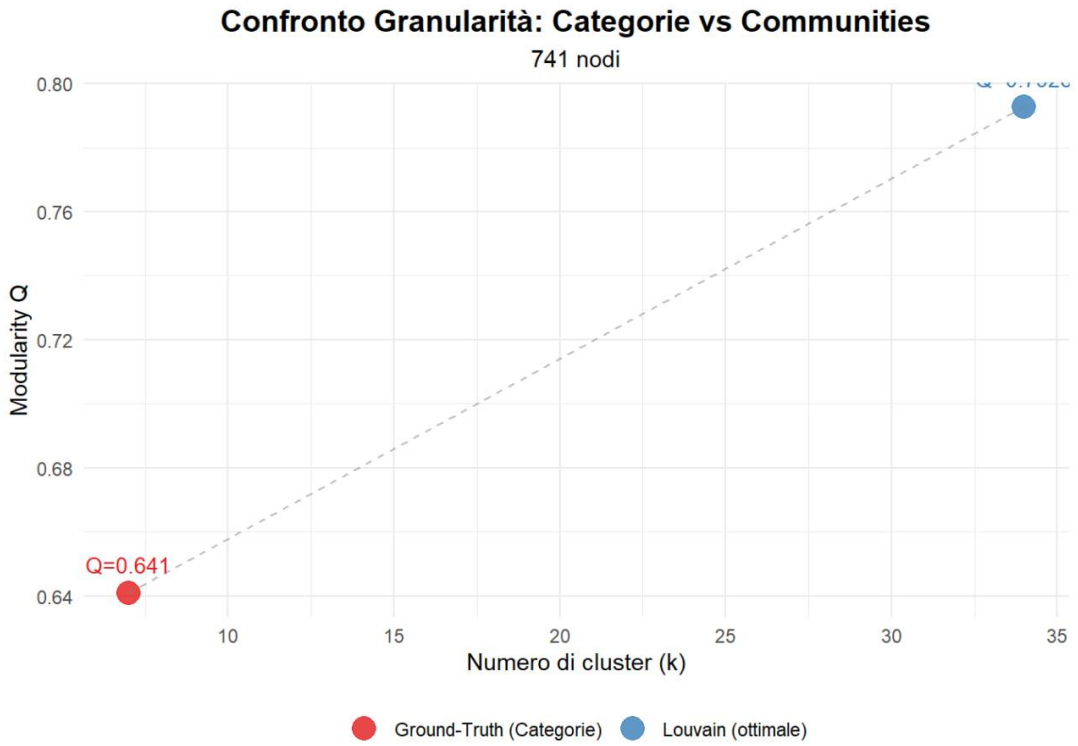
A destra ho la rete filtrata in cui sono mostrate le 7 categorie della Cora Network.

A pagina seguente valuto la modularità.

K	Source	Modularity	Avg_size	Min_size	Max_size
7	Ground-Truth (Categorie)	0.6410	105.9	42	175
34	Louvain (ottimale)	0.7926	21.8	1	107

Louvain trova una struttura più modulare che cattura meglio le sotto comunità.

La Modularity è migliorata del 23.7 %



Le 7 categorie sono macro-raggruppamenti grossolani.

La struttura di citazione riflette specializzazioni più fini.

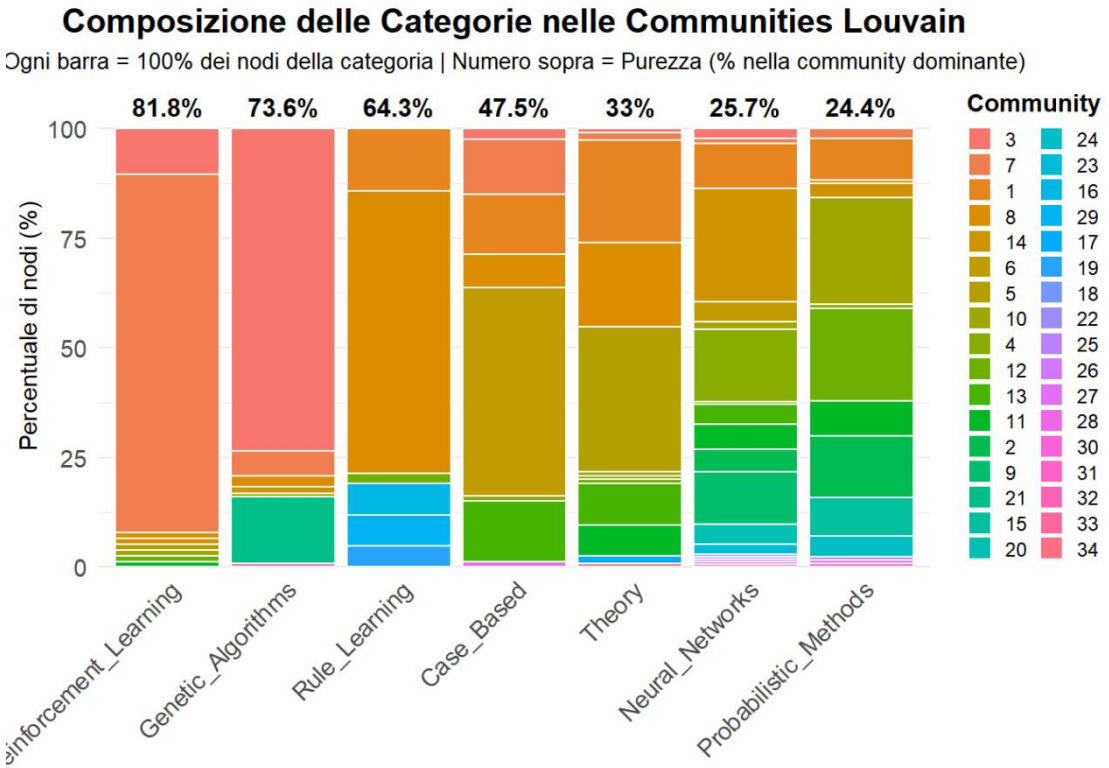
L'idea è che ciò indichi l'esistenza di sottocategorie tematiche (es. CNN o RNN all'interno di Neural Networks).

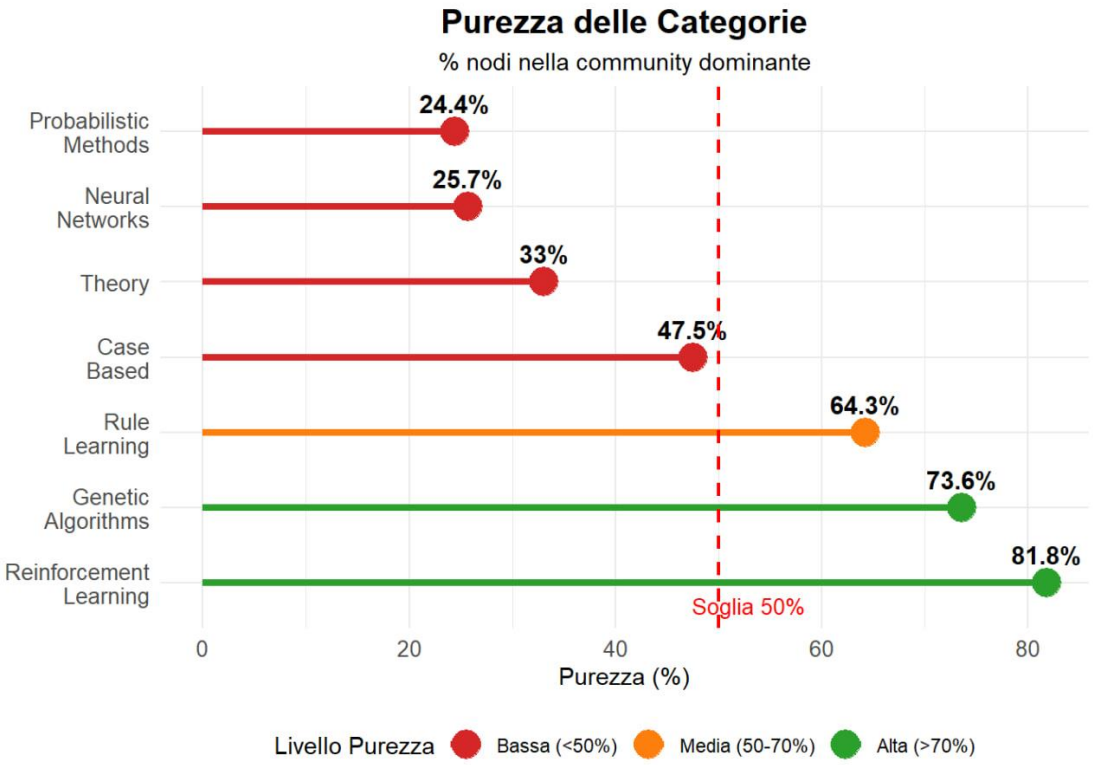
Domanda 13 : Quali categorie sono più pure ? (una sola community o meno possibili)

Qui l'idea è di incrociare la partizione fornita da Louvain con la classificazione originale delle categorie.

Si noti come tutte le categorie siano frammentate. Prendiamo ad esempio :

- Neural Networks (la 2^a più frammentata) : probabile esistenza di molte sottocategorie tematiche.
- Reinforcement Learning (la meno frammentata): 81.8% dei nodi in una sola community (la 7), ma comunque divisa in 8 communities.



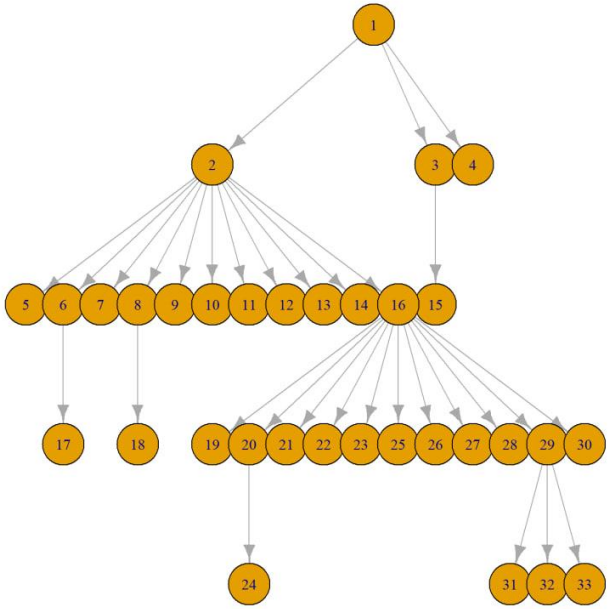


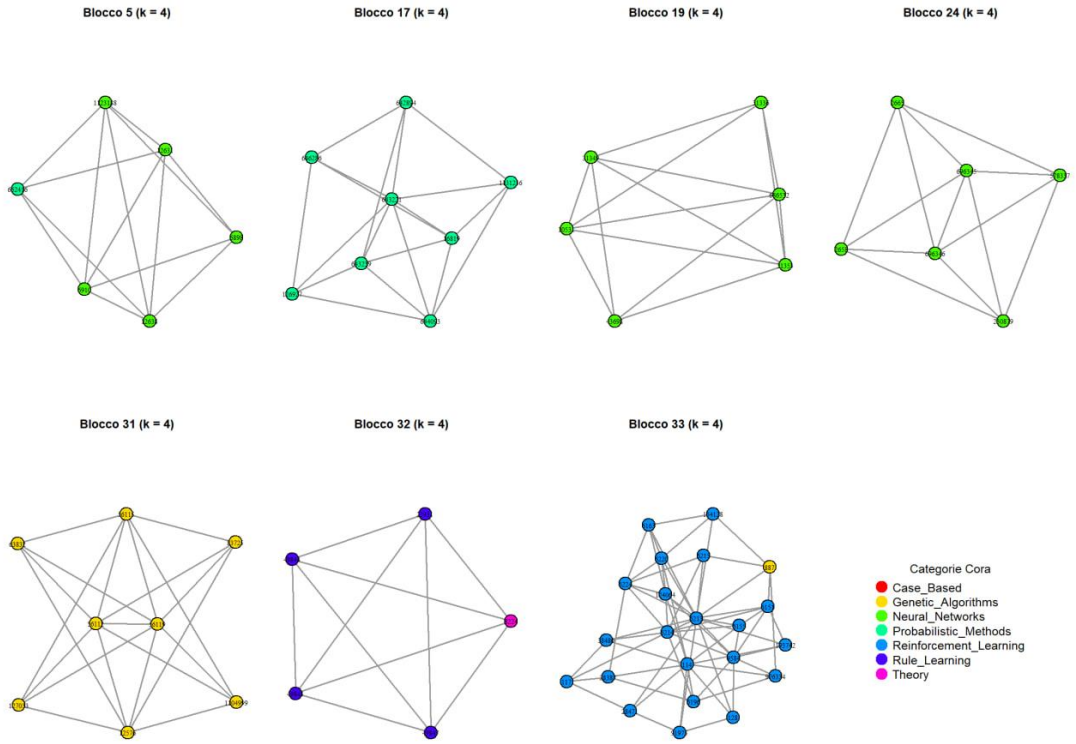
Global Analysis

Domanda 14 : Studia i K-connected components e verifica se i gruppi con connettività più alta sono composti da soli paper di una stessa catgoria.

Uso come prima il grafo non diretto e filtrato (indegree + outdegree >= 5).

La struttura dei cohesive blocks è fatta da 33 blocchi, con alcuni che hanno una connettività pari a massimo 4.





L'analisi dei K-connected components con k max (4) si è dimostrata interessante nei blocchi 5, 32 e 33.

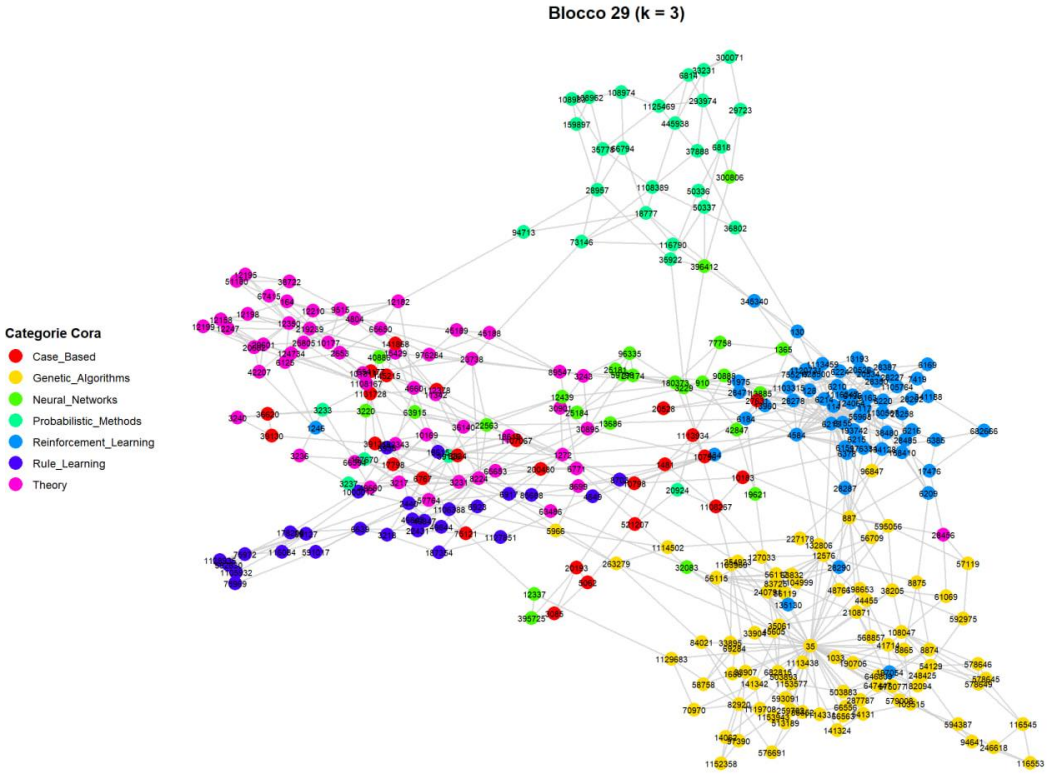
Infatti al loro interno troviamo dei paper outsider, ossia con una categoria diversa da quella di tutti gli altri.

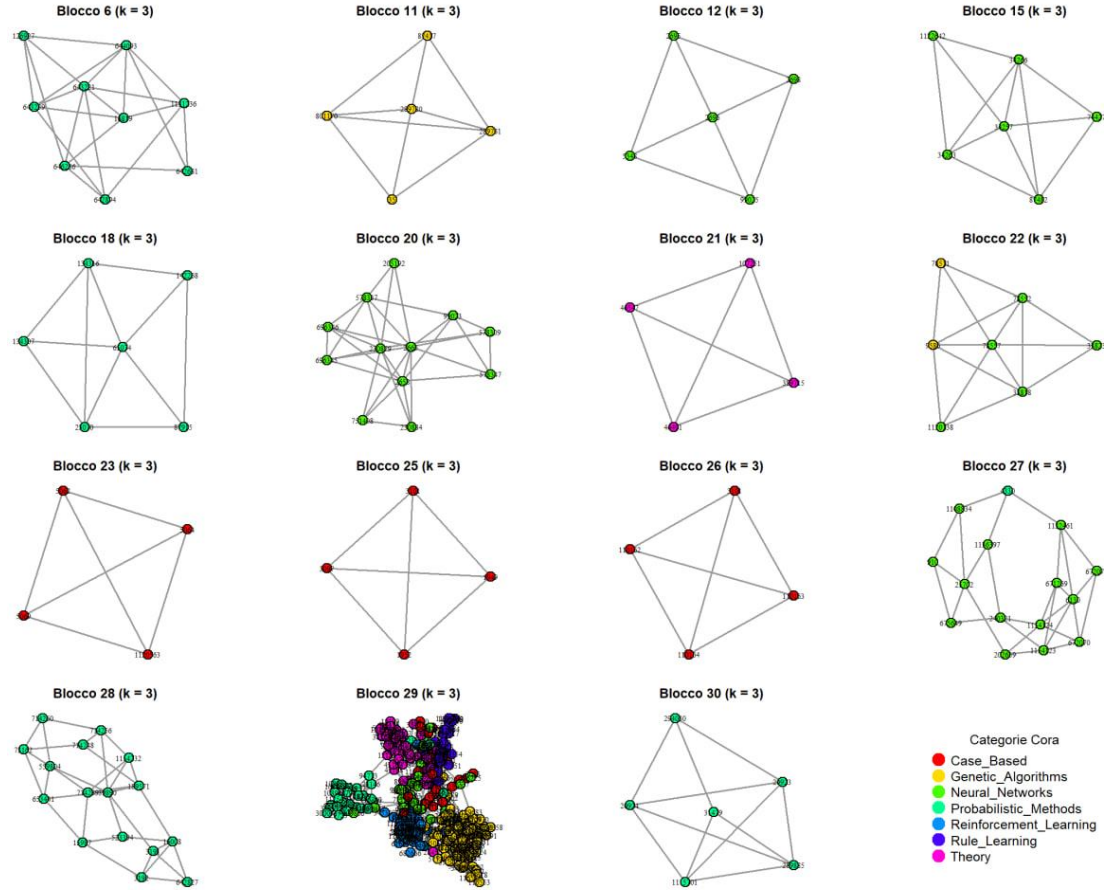
Proseguo con questa analisi, vediamo ora i blocchi con coesione max – 1 (3).

Abbiamo che i blocchi sono quasi tutti omogenei, tranne :

- il blocco 22 e 27, che vedono entrambi la presenza di 1 paper eterogeneo
- il blocco 29 che risulta interessante in quanto molto eterogeneo.

Analizziamo il blocco 29 più in dettaglio.





La categoria Neural_Networks risulta esser più dispersa rispetto alle altre categorie. Tante connessioni verso le altre categorie.

Inoltre ho che Reinforcement_Learning risulta esser denso ma allo stesso tempo importante per varie altre categorie.

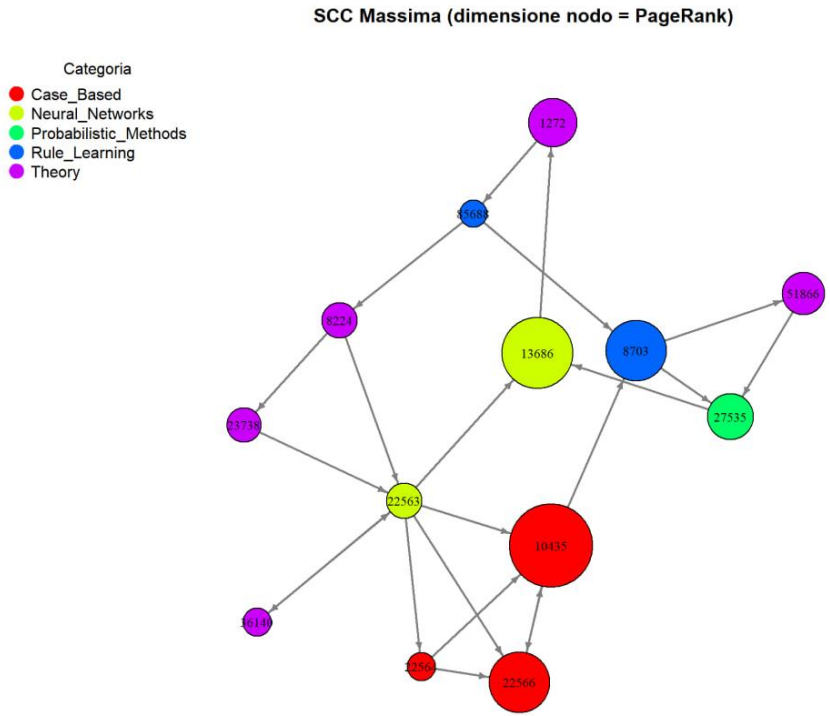
Domanda 15 : Analizza gli SCC per verificare l'aciclicità della rete.

Torniamo ora a lavorare con il grafo diretto.

2 nodi sono nello stesso Strongly Connected Components se sono a vicenda raggiungibili mediante un path diretto.

In questo caso stiamo lavorando con una citation network, la quale è una quasi aciclica.
Questo è dovuto al fatto che quando un paper ne cita un altro, questo citato sarà stato scritto in passato.
Come prvisto, non si trovano SCC di grandi dimensioni.
A destra vien mostrata la distribuzione del numero di componenti degli SCC (sono 2526 in totale).

Dimensione	Numero_SCC
13	1
7	1
6	1
5	4
4	5
3	18
2	92
1	2404



Lavoro ora con il sottografo della SCC massima.

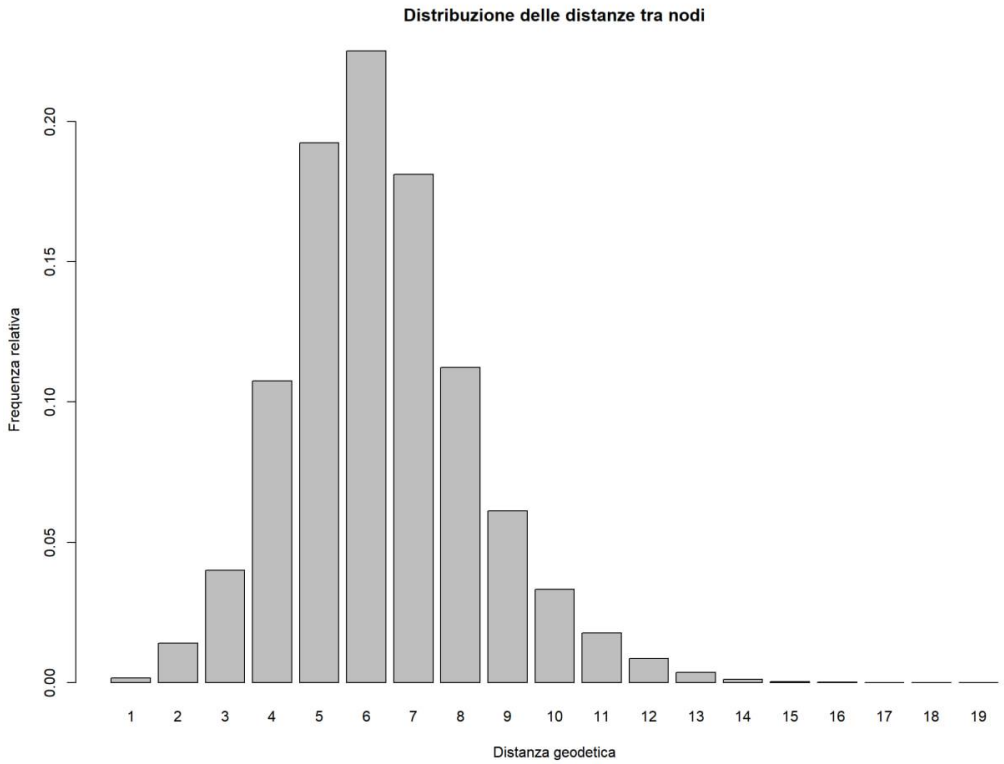
Notare come in questo SCC ci siano dei paper che si citano a vicenda, il che risulta controintuitivo per una citation network.

Domanda 16 : Verifica lo Small-World Effect sul grafo non diretto.

Lo Shortest Path tra 2 nodi (Geodesic Path) in un grafo è il path con il numero minore di edges.

La lunghezza di tale path è detta Shortest Distance (Geodesic Distance).

Lo Small-World effect mi dice che la Shortest Distance è tipicamente sorprendentemente corta.



Uso il grafo non diretto.

La Distanza media è pari a : 6.31

In questo caso lo Small-World Effect mostra che questa rete rispetta la teoria dei 6 gradi di separazione.

Domanda 17 : Verifica che la «Assortativity by category» sia molto alta (dato che ho tante citazioni intra-categoria).

La rete è Assortative se una frazione significante delle edges corre tra vertici dello stesso tipo. Misura di ciò è la Modularità.

Posso normalizzare la Modularità andando a dividere per il massimo valore che essa può assumere.
Una misura dell' Assortative Mixing in base ad una caratteristica scalare è la Covarianza.

L'Assortativity Coefficient è la misura della Assortativity normalizzata (Coefficiente di correlazione di Pearson).
In questo contesto misura quanto i paper della stessa categoria tendono a collegarsi tra loro.

Assortativity by category : 0.7710854 (forte omofilia di categoria)

Domanda 18 : Verifica se l'elite delle pubblicazioni crea un "club" chiuso.

Si chiede se i paper ad alta centralità (PageRank) tendono a citare/ricevere citazioni da altri paper ugualmente centrali.

Se l'assortatività è positiva, allora l'elite dei paper centrali cita prevalentemente altri paper influenti, creando una sorta di "elite club" della conoscenza scientifica nellarete.

Assortatività rispetto al PageRank : -0.07957071

Non si forma un "elite club".

I paper centrali non creano una sottorete chiusa e autoreferenziale, ma sono piuttosto ponte verso molti altri paper meno centrali.
Il valore è vicino a zero, quindi si tratta di una tendenza debole.