

UNIVERSITÀ DEGLI STUDI DI PADOVA

CORSO DI LAUREA TRIENNALE IN
INGEGNERIA DELL'INFORMAZIONE

**Analisi di una rete di autori di
pubblicazioni scientifiche**



Laureando:

Pietro Maria NOBILI

Relatore:

Prof.ssa Cinzia PIZZI

Matricola:

1067941

Correlatore:

Dott. Mattia SAMORY

18 Luglio 2018
Anno Accademico 2017/2018

Indice

Abstract	1
Introduzione	1
1 Community detection	3
1.1 Studi precedenti	3
1.2 Metodi usati	3
1.2.1 Girvan–Newman	3
1.2.2 Blockmodel	3
1.2.3 Clauset–Newman–Moore	3
2 Estrazione dati	5
2.1 Struttura database Microsoft	5
2.2 Processo di estrazione	6
2.2.1 Filtro degli autori per affiliation	7
2.2.2 Unione di ID autore in singoli nodi	8
3 Troubleshooting	11
4 Risultati	13
5 Conclusioni	15

Abstract

È stato generato un grafo delle pubblicazioni del DEI.
Sono stati cercati cluster nel grafo.
Sono stati confrontati con la struttura delle comunità del dipartimento.

Introduzione

Breve descrizione del community detection e della sua importanza, in generale e nel caso particolare delle comunità di autori di pubblicazioni scientifiche.

Presentazione della struttura della tesi.

Capitolo 1

Community detection

1.1 Studi precedenti

Descrizione della community detection.

Descrizione della bibliografia attuale sui grafi di coautori.

1.2 Metodi usati

Che metodi vengono usati per generare le comunità.

1.2.1 Girvan–Newman

Descrizione metodo GN.

Nella libreria SNAP di Stanford [1]

1.2.2 Blockmodel

Descrizione metodo blockmodel.

1.2.3 Clauset-Newman-Moore

Descrizione metodo Clauset-Newman-Moore se implementato.

Capitolo 2

Estrazione dati

2.1 Struttura database Microsoft

Il database da cui sono stati estratti i dati è il Microsoft Academic Graph [2], che contiene informazioni relative a pubblicazioni scientifiche, autori, istituzioni accademiche, riviste, conferenze e settori di studio. I record presenti nei file forniscono le relazioni tra queste entità.

Il database è composto da undici file di testo che contengono un record per riga, con i campi separati da tabulazione.

Per il lavoro oggetto di questa tesi sono stati utilizzati in particolare quattro file del database, la cui struttura è illustrata nella tabella 2.1.

L'ultimo aggiornamento del database disponibile risale all'agosto 2015.

Tabella 2.1: Struttura del database

Nome file (Numero record)	Campi
Authors.txt (123.017.489)	Author ID Author Name
PaperAuthorAffiliations.txt (325.498.063)	Paper ID Author ID Affiliation ID Original affiliation name Normalized affiliation name Author sequence number
Affiliations.txt (2.719.436)	Affiliation ID Affiliation name
Papers.txt (122.695.085)	Paper ID Original paper title Normalized paper title Paper publish year

	Paper publish date
	Paper Document Object Identifier (DOI)
	Original venue name
	Normalized venue name
	Journal ID mapped to venue name
	Conference series ID mapped to venue name
	Paper rank

2.2 Processo di estrazione

Dal sito di dipartimento (<http://www.dei.unipd.it/lista-docenti>) sono stati estratti i nomi degli attuali afferenti DEI, includendo Docenti, Assegnisti di ricerca, Collaboratori di ricerca e Dottorandi.

La lista ottenuta comprende 379 autori. Dove presente, è stato estratto il Settore Scientifico Disciplinare dell'afferente, che ha fornito la partizione in classi utilizzata come riferimento alla fine dell'elaborazione dei dati.

I nomi propri degli autori sono stati abbreviati in tutte le possibili combinazioni per rispecchiare la struttura del database Microsoft, ed è stato creato un file con la struttura seguente:

Tabella 2.3: PersoneComunitaDEI.txt

```
a a pietracaprina          INF/01 - INFORMATICA
a alberto pietracaprina    INF/01 - INFORMATICA
andrea a pietracaprina     INF/01 - INFORMATICA
andrea alberto pietracaprina INF/01 - INFORMATICA
...
```

Utilizzando questa lista di nomi ed abbreviazioni, sono stati estratti dal file *Authors.txt* le coppie (ID autore, nome autore), che risultano essere 8.135, con una media di 21,5 ID per nome.

Con il set di ID autori ottenuto, sono stati estratti dal file *PaperAuthorAffiliations.txt* i record relativi ai paper, nella forma (ID paper, ID autore, ID affiliation), per un totale di 62.291 paper.

Dalla lista così ottenuta, non necessariamente ordinata, sono stati creati gli edge fra ID autore, che sono poi stati aggregati creando una edge list pesata, come indicato in 2.4

Tabella 2.4: Creazione edge

IDpaper1	IDautore1				
IDpaper1	IDautore2		IDautore1	IDautore2	1
IDpaper1	IDautore3	→	IDautore1	IDautore3	1
IDpaper2	IDautore2		IDautore2	IDautore3	2
IDpaper2	IDautore3				

A partire da questi edge è stato generato il primo grafo che rappresenta la rete degli autori. Questo metodo di creazione del grafo, che considera le collaborazioni tra autori, riduce a 778 il set di ID autori per un totale di 287 nomi univoci. Il grafo è formato da 1.830 edge con 7.803 di peso totale.

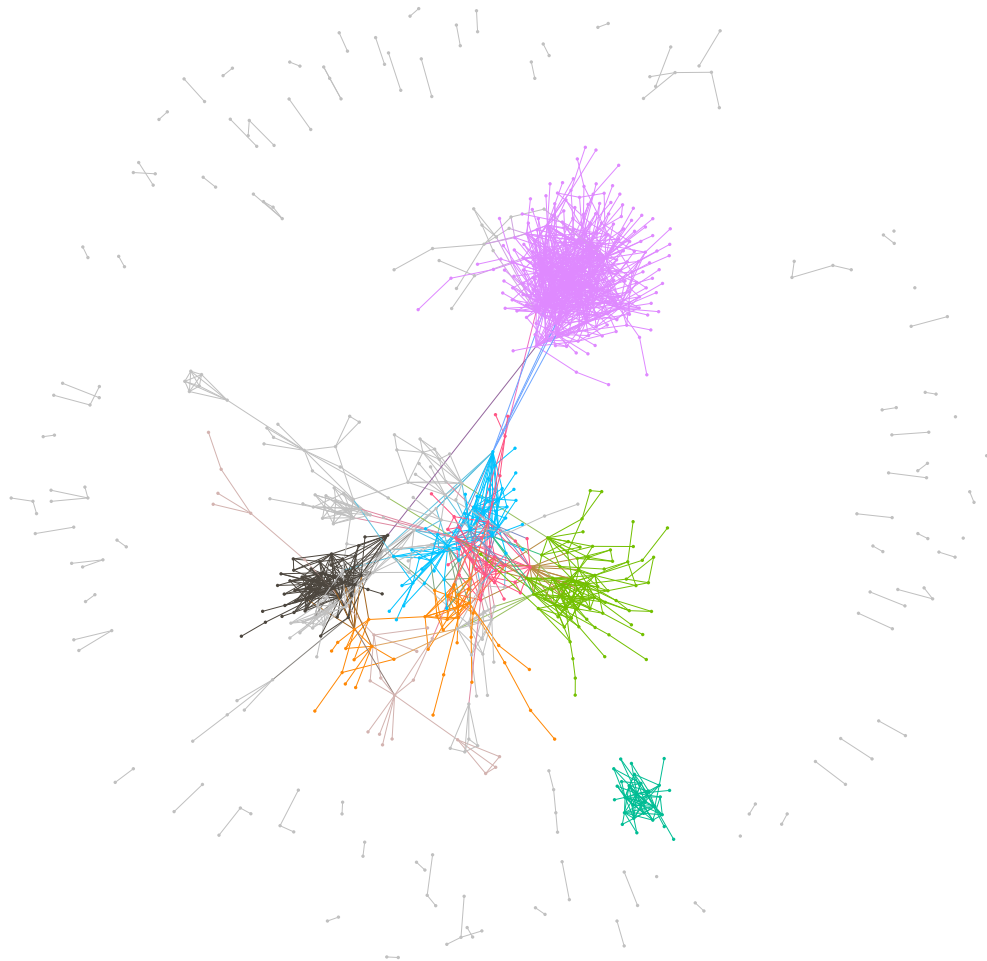


Figura 2.1: Grafo collaborazioni

Analisi della prima estrazione

I due problemi principali di questo processo emergono al momento dell'estrazione degli ID autore dal file *Authors.txt*:

- Vengono selezionati anche gli ID riferiti ad autori omonimi, non affiliati al DEI.
- Ad un singolo autore DEI sono associati più ID autore.

2.2.1 Filtro degli autori per affiliation

Il primo problema è già in parte risolto dal metodo di creazione del grafo, che considera solo gli ID autori che hanno almeno una collaborazione con un altro ID nel set. In questo modo più del 90% degli ID viene filtrato.

È stato sviluppato un secondo metodo di estrazione dei dati che esclude gli ID autore se non hanno mai pubblicato un paper a Padova, considerando le informazioni sulle affiliation dei paper, come viene illustrato di seguito.

- Dal file *Affiliations.txt* si estrae la lista delle affiliation in cui risulta nel nome un match all'espressione regolare “*pad(ov|u)a*”.
- Dalla lista di terne (ID paper, ID autore, ID affiliation) si mantengono solo quelle in cui l'ID affiliation compare nel set di affiliation padovane appena estratte.
- Dalle terne selezionate, si estrae un set di ID autori che hanno pubblicato almeno un paper con affiliation padovana.
- Si estraggono i paper scritti da questi ID autore e si procede alla generazione degli edge pesati.

Applicando questo metodo, gli ID autore si riducono a 306, relativi a 201 nomi univoci. Il grafo generato include 850 edge con un peso totale di 5.195.



Figura 2.2: Grafo Padovani

2.2.2 Unione di ID autore in singoli nodi

Per risolvere il secondo problema insorto nell'estrazione dei dati, sono state seguite due vie alternative fra loro. Per risolvere il secondo problema, si propongono due modi:

Per nome

I nodi con nomi uguali o abbreviazioni l'uno dell'altro sono considerati un unico nodo.

Per distanza

Il metodo per nomi introduce un errore potenzialmente molto grave, nomi come Michele Zorzi e Mattia Zorzi vengono confusi e considerati un unico nodo. Nella lista di nomi considerata questo succede solo nel caso citato, ma in dataset più ampi i falsi positivi aumentano considerevolmente. Si tenta di risolvere questo problema considerando unici due nodi in base alla distanza minima che hanno nel grafo.

I nodi con nomi uguali o abbreviazioni l'uno dell'altro sono considerati un unico nodo solo se sono anche vicini nel grafo: si calcola il cammino minimo tra i nodi e li si unisce se è di lunghezza minore di x

TODO Il valore ottimo di x sarebbe bello scoprirlo sperimentalmente

Capitolo 3

Troubleshooting

Il grafo generato presenta ancora delle carenze.

La lista di partenza include gli afferenti DEI attuali, mentre il database è del 2015. Questo comporta la mancanza di professori, non più a Padova, che avevano ruolo di aggregatore di una comunità.

TODO inserire il nome di Apostolico nella lista di partenza e valutare i cambiamenti

Un modo proposto per includere i nomi mancanti è: estrarre gli ID autore; estrarre i paper-aut-aff; da questa lista di paper estrarre tutti gli ID autore; potenzialmente ridurre i paper (e gli autori) estratti in base alle affiliation (anche solo del DEI e non di tutta Padova); iterare il processo con la nuova lista di autori.

In questo modo si estraggono anche troppe comunità, una singola collaborazione con un dipartimento esterno comporta alle iterazioni successive l'inclusione di molti autori di quel dipartimento. Un modo per risolvere il problema è, alla fine delle iterazioni e della creazione dei cluster, considerare solo quelli che contengono almeno un nome che era nella lista originale: in questo modo dipartimenti esterni, anche se connessi al grafo, non vengono inclusi.

Capitolo 4

Risultati

Descrizione della v-measure

Presentazione dei valori di v-measure per i vari grafi e commenti.

Capitolo 5

Conclusioni

Partendo da una lista di nomi di un dipartimento si può estrarre un grafo che lo rappresenti? Le comunità generate rispecchiano quelle reali?

Bibliografia

- [1] Jure Leskovec and Andrej Krevl. SNAP Datasets: Stanford large network dataset collection. <http://snap.stanford.edu/data>, June 2014.
- [2] Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June (Paul) Hsu, and Kuansan Wang. An overview of microsoft academic service (mas) and applications. In *Proceedings of the 24th International Conference on World Wide Web, WWW '15 Companion*, pages 243–246, New York, NY, USA, 2015. ACM. URL: <http://doi.acm.org/10.1145/2740908.2742839>.