INDAGINE SULLE RELAZIONI TRA I CORSI DI LAUREA DELL'UNIVERSITÀ DI PADOVA: UN' APPLICAZIONE ALLE SCUOLE STEM

Corso di Metodi Statistici per i Big Data
Anno accademico 2022/2023
Brunello Irene: matricola 2002694
Gastaldello Matteo: matricola 2012098
Tedesco Gaetano: matricola 2006942

ABSTRACT
L'obiettivo di questo progetto è condurre un'analisi delle connessioni presenti tra i vari corsi di studio STEM offerti dall'Università di Padova, attraverso la creazione di un algoritmo che renda tale analisi replicabile.
In particolare, si vuole indagare come i prerequisiti di ogni insegnamento interconnettono i vari corsi di studio, mediante la realizzazione di una rete e l'applicazione di metodi di Community Detection su questa. Insieme al codice fornito per l'analisi viene allegato il file *school_net_lib* contenente tutte le funzioni utilizzate per manipolare il dataset, creare la rete e analizzarla. Inoltre vengono allegate anche alcune rappresentazioni delle reti per la scuola di scienze e per quella di ingegneria realizzate tramite il software "*Gephi*". I colori dei nodi e degli archi sono attribuiti in base alla scuola di appartenenza, in base al corso di laurea di appartenenza e in base alle comunità trovate nella fase di *community detection*.

INTRODUZIONE DEL DATASET

Per poter disporre di un dataset su cui fare le analisi, è stata svolta un'operazione di web-scraping sul sito internet della didattica dell'Università di Padova, che presenta le informazioni relative a tutti gli insegnamenti nei Corsi di Studio dell'Ateneo di Padova. Nello specifico, si sono prese in considerazione le pagine web relative alle Scuole di Scienze (https://didattica.unipd.it/off/2022/LT/SC) e Ingegneria (https://didattica.unipd.it/off/2022/LT/IN) degli immatricolati nell'anno 2022/2023.

Curr.	Codice	Insegnamento	CFU	Anno	Periodo	Lingua	Responsabile
COMUNE	SCP4067973	SPERIMENTAZIONI DI FISICA 1 Info e programma immatricolati A.A. 2022/23 Attuale A.A. 2022/23 , Prossimo A.A. 2023/24				ITA	ANTONINO MILONE
»	SCP4067974	SPERIMENTAZIONI DI FISICA 1 (MOD. A) Info e programma immatricolati A.A. 2022/23 Attuale A.A. 2022/23 , Prossimo A.A. 2023/24	6	I Anno (2022/23)	Annuale	ITA	ANTONINO MILONE
»	SCP4067975	SPERIMENTAZIONI DI FISICA 1 (MOD. B) Info e programma immatricolati A.A. 2022/23 Attuale A.A. 2022/23 , Prossimo A.A. 2023/24	6	I Anno (2022/23)	Annuale	ITA	PAOLO CASSATA
COMUNE	SC05100190	ANALISI MATEMATICA 1 (Iniziali cognome A-L) Info e programma immatricolati A.A. 2022/23 Attuale A.A. 2022/23 , Prossimo A.A. 2023/24	8	I Anno (2022/23)	Primo semestre	ITA	DAVIDE BARILARI
COMUNE	SC05100190	ANALISI MATEMATICA 1 (Iniziali cognome M-Z) Info e programma immatricolati A.A. 2022/23 Attuale A.A. 2022/23 , Prossimo A.A. 2023/24	8	I Anno (2022/23)	Primo semestre	ITA	PIERPAOLO SORAVIA
COMUNE	SC03101111	CHIMICA Info e programma immatricolati A.A. 2022/23 Attuale A.A. 2022/23 , Prossimo A.A. 2023/24	6	I Anno (2022/23)	Primo semestre	ITA	ALESSANDRO ALIPRANDI

Figura 1. Esempio: Pagina web del corso di laurea di Astronomia

OPERAZIONI DI PREANALISI

ESTRAZIONE DEI DATI

Per l'estrazione dei dati dal sito è stato progettato un *wrapper* che sfrutta al suo interno la funzione `Rcrawler` dell'omonimo pacchetto.
In particolare, la funzione `Rcrawler` svolge il processo di estrazione del dato dal web partendo dall'URL fornito dall'utente. Questa esegue il *fetch* della pagina ed estrae progressivamente i nuovi URL aggiungendoli ad una lista "*di frontiera"*, successivamente suddivide i link in nodi e assegna ogni nodo ad un *thread* del processore della macchina ospite che chiameremo *lavoratore.*

A questo punto ciascun lavoratore inizia un ciclo di *parsing and fetch* dei link relativi alla propria porzione della lista di frontiera salvando l'intero contenuto delle pagine web esplorate e aggiungendo gli *out-link* di queste (link che mandano ad altre pagine) in una propria lista di frontiera.
Il *crawling* della rete viene arrestato quando il lavoratore raggiunge la profondità (dall'URL radice) preimpostata dall'utente e successivamente la funzione applica il processo di estrazione ai documenti html delle pagine, restituendo solo il contenuto targettizzato dall'utente attraverso *XPath*.
All'interno del wrapper i dati vengono manipolati e trasformati in modo tale da restituire un dataset più consono alle successive analisi eliminando campi vuoti o non utili.

PRESENTAZIONE DEL DATASET
Alla fine del processo di estrazione i dati si presentano all'interno di un dataframe che ha come unità statistica il singolo insegnamento e per ognuno di questi le variabili:
-*Codice* = codice dell'insegnamento
-*Corso*= nome dell'insegnamento
-*Laurea*= nome del corso di laurea
-*Settore*= settore disciplinare dell'insegnamento
-*Prerequisiti*= prerequisiti dell'insegnamento
-*Contenuto*= contenuti dell'insegnamento, presi dal syllabus
-*Link*= link url alla pagina web dell'insegnamento
-*Periodo*= periodo di erogazione dell'insegnamento
-*Scuola*= scuola di appartenenza dell'insegnamento

Dagli insegnamenti sono, inoltre, stati rimossi esami come: "idoneità alla lingua inglese", "prova finale" e "tirocinio" comuni a tutti i corsi di laurea e che non rientrano nell'obiettivo delle analisi.

MANIPOLAZIONE DATI GREZZI
Al fine di rendere il più efficace possibile la ricerca dei prerequisiti nei contenuti dei corsi, è stata eseguita una pulizia dei dati. In particolare, è stata effettuata una formattazione dei testi presenti nel dataset per renderli confrontabili.
Le funzioni di manipolazione dei dati grezzi che sono state create sono:
-`normalize_sector` normalizza il settore, prendendo solo le lettere del codice dell'ambito disciplinare, (e.g. il codice FIS/02 viene ricondotto a FIS), mentre gli ambiti non specificati vengono posti NA;
-`format_names` formatta i nomi dei corsi, in modo da permettere la ricerca di questi nei prerequisiti. Il testo viene ridotto in minuscolo, vengono eliminate parentesi, frasi relative alla suddivisione degli insegnamenti in base al numero di matricola, alle iniziali del cognome e/o al canale di appartenenza, specificazioni dei moduli A/B degli esami. Vengono rimosse anche le stop words (preposizioni, articoli, avverbi,..) così come spazi superflui e segni di punteggiatura. Inoltre il nome dell'insegnamento "analisi matematica" viene rimpiazzato con "analisi", per rendere tutte le occorrenze uguali e confrontabili;
-`format_numbers` formatta i numeri romani in numeri decimali;
-`format_requires` formatta il campo dei prerequisiti in modo analogo ai nomi degli insegnamenti, in modo da facilitare il confronto tra il nome dell'insegnamento e il prerequisito;

Vengono poi eliminati i duplicati (dovuti allo stesso esame ripetuto più volte distinto per numero di matricola o iniziale dei cognomi o canale A/B) confrontando i codici degli insegnamenti. Tale scelta è giustificata dal fatto che non si ritengono informativi i record rispettivi, essendo uguali tra di loro.

	Codice	Corso	Laurea	Settore	Prerequisiti
1	SCP4067974	sperimentazioni fisica 1	ASTRONOMIA	FIS	conoscenze base matematica fisica
2	SC05100190	analisi 1	ASTRONOMIA	MAT	funzioni elementari reali potenze modulo esponenzi...
3	SC03101111	chimica	ASTRONOMIA	CHIM	conoscenze base matematica fisica chimica acquisit...

Figura 2. Dataset finale relativo alla scuola di Scienze risultante dalle operazioni di formattazione.

Il risultato finale che si ottiene è una riduzione del dataset da 597 osservazioni a 351 osservazioni per la scuola di Scienze, e una riduzione da 1027 a 276 osservazioni per la scuola di Ingegneria (questo calo drastico è dovuto al fatto che tutti i corsi di laurea in Ingegneria hanno una solida base di corsi comuni, che però, ai fini dell'analisi, viene considerata una sola volta).

Le funzioni per la creazione e la manipolazione del dataset così costruite consentono lo svolgimento delle successive analisi a diversi livelli: si possono esplorare le relazioni tra gli insegnamenti all'interno dello stesso corso di laurea, all'interno della stessa scuola oppure all'interno di più scuole unite.

ANALISI DI RETE

Ottenuto il dataset, si passa alla fase di analisi, dove si considerano solamente le variabili "Corso" e "Prerequisiti" del dataset pulito, mentre "Laurea", "Settore" e "Scuola" fungeranno da covariate dei nodi della rete.
Le funzioni realizzate per la creazione della rete sono:
-`edge` crea archi diretti, restituendo le associazioni prerequisito-corso per ogni singolo nome di corso;
-`make_edges` costruisce gli archi della rete, sfruttando la funzione `edge` sopra descritta. L'arco tra due corsi è definito come segue: "il nome del corso è presente tra i prerequisiti di un altro corso". Se lo stesso insegnamento è presente in due corsi di laurea, viene preso solo una volta, se invece non è prerequisito di alcun corso, viene generato un NA, che viene poi eliminato, insieme ad eventuali self loops. L'unicità di un corso viene identificata attraverso il suo codice che è univoco;
-`make_network`, dalla lista degli archi generata da `make_edges`, costruisce la rete e imposta le covariate di nodo utili alla successiva analisi.

Per quanto riguarda l'analisi di rete, la funzione `net_analytics` produce le statistiche descrittive di rete e di nodo. Viene plottato anche un grafico della rete, con le etichette sugli Hub, trovati ponendo come soglia discriminante il 90esimo percentile della distribuzione del "punteggio Hub", una particolare metrica che assume valori tra 0 e 1. Più il punteggio hub è vicino a 1 più quel nodo è definito come un Hub e viceversa. Essi risultano essere gli insegnamenti di chimica, geometria, matematica, fisica, fisica generale e analisi per la scuola di Scienze, mentre per la scuola di Ingegneria primeggiano analisi 1, elettrotecnica, segnali sistemi e algebra lineare geometria. Questi Hub sono dunque i corsi basilari per le rispettive scuole. Collegano i vari corsi di laurea e fungono da "ponti" che collegano corsi di diverse lauree.

Dato l'interesse nello scovare i prerequisiti predominanti all'interno delle due scuole, tra le statistiche di nodo calcolate si notino quelle definite per le reti dirette. Ad esempio è stato calcolato il grado uscente e anche per l'eccentricità è stata presa in considerazione questa caratteristica della rete .
Dal cammino minimo medio di 2.21 per Scienze e di 1.83 per Ingegneria, dal basso valore del diametro e della densità si può inoltre affermare che entrambe le scuole sono ragionevolmente small world, come ci si aspettava, in quanto l'ateneo suddivide i corsi di laurea per macro aree comuni e gli insegnamenti delineano un percorso durante dei tre anni di università.

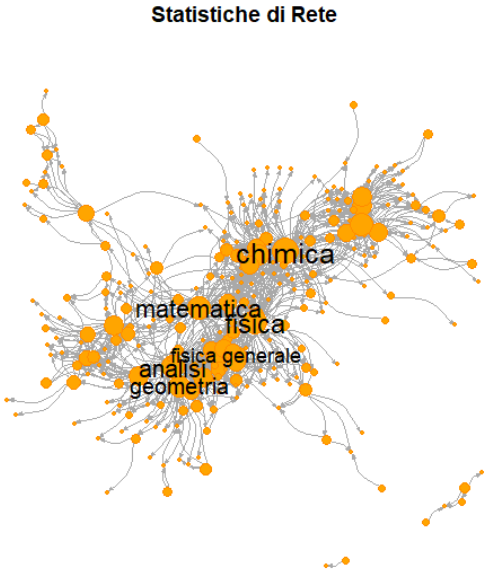

Figura 3. Hub della scuola di Scienze

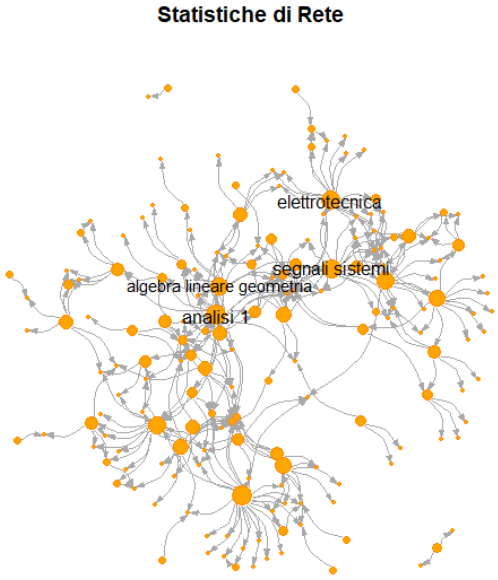

Figura 4. Hub della scuola di Ingegneria

Infine per lo studio delle comunità della rete viene implementato un algoritmo di *community detection,* chiamato Walktrap.

ALGORITMO WALKTRAP

Il walktrap implementa un metodo di clustering gerarchico agglomerativo che sfrutta un "random walk" per trovare strutture di comunità. Questo algoritmo si basa sul fatto che i cammini casuali tendono a restare "intrappolati" in porzioni della rete densamente connesse che corrispondono a comunità. Usando i cammini minimi definiamo una misura di similarità tra i vertici e le comunità. Ad ogni istante di tempo, il cammino minimo che sto considerando si trova su un nodo (nodo *i*) e sceglie a caso e uniformemente dove spostarsi tra i nodi collegati con il nodo *i-esimo*.
Ad ogni step la probabilità di transizione dal nodo *i* al nodo *j* è data dalla probabilità P_{ij} ,definita come segue:

$$P_{ij} = \frac{A_{ij}}{d(i)}$$

Dove A_{ij} vale 1 se il nodo *i* è collegato con il nodo *j*, 0 altrimenti. Invece *d(i)* è il grado del nodo *i*.
Viene definita la probabilità di passare dal nodo *i* al nodo *j* attraverso un random walk di lunghezza *t* come P^t_{ij}.

Inoltre, viene introdotta una misura di distanza tra nodi che cattura la struttura comunitaria del grafo. Questa distanza deve essere grande se due nodi appartengono a comunità diverse e, al contrario, dev'essere piccola se appartengono alla stessa comunità.
La distanza r_{ij} tra due nodi dev'essere sufficientemente elevata da raccogliere informazioni sulla topologia del grafo, ma non dev'essere troppo grande per evitare che la distanza dipenda solamente dal grado del nodo di arrivo (e non dal nodo di partenza).

Si consideri un random walk nella rete di una lunghezza prefissata *t*. Si userà l'informazione data dalle probabilità P^t_{ij} per andare dal nodo *i* al nodo *j* in t passi.
Le informazioni relative alle probabilità del nodo *i-esimo* sono codificate nella matrice P e risiedono nella *i-esima* riga.

Quindi P sarà una matrice che contiene nelle righe tutte le probabilità relative ai collegamenti del nodo *i* con un cammino minimo di distanza massima *t*. Per confrontare due nodi (*i* e *j* per esempio) si noti che:

La distanza tra il nodo *i* e il nodo *j* (r_{ij}) è definita come la norma euclidea tra le due distribuzioni di probabilità P^t_i. e P^t_j (ovvero la riga *i* e la riga *j* della matrice P).

Definita la distanza tra due comunità (r_{C1C2}) come segue: $$r_{C_1C_2} = \sqrt{\sum_{k=1}^{n} \frac{(P^t_{C_1k} - P^t_{C_2k})^2}{d(k)}}$$

l'algoritmo inizia considerando n comunità (ogni nodo ne rappresenta una). Si noti che, al fine di ridurre la complessità, unisce solamente comunità adiacenti (collegate da almeno un arco).
Ad ogni passo *k*, l'algoritmo sceglie le due comunità da unire secondo il metodo di Ward, ovvero minimizzando la media della sommatoria delle distanze al quadrato di ciascun nodo dalla sua comunità, e procede finché non arriva ad avere tutti i nodi cpntenuti in un'unica comunità

Questo procedimento porta ad avere una sequenza $P_{k\ (1 \le k \le n)}$ di partizioni in comunità. Per identificare la partizione che cattura meglio la struttura comunitaria presente nella rete, viene utilizzata la modularità che è la metrica più adatta da utilizzare se si vuole trovare la miglior suddivisione in comunità. La miglior partizione è quella che massimizza la modularità.

COMMENTO AI RISULTATI OTTENUTI
E' stato applicato questo algoritmo alla rete ottenuta sia dalla scuola di Scienze che da quella ottenuta dalla scuola di Ingegneria, con una piccola differenza: per la prima i "passi" dell'algoritmo sono impostati a 4, mentre per la seconda questo parametro viene impostato a 6. Il motivo di questa scelta è la minore densità della rete di Ingegneria che, con un passo pari a 4 avrebbe identificato troppe comunità, molte delle quali sarebbero risultate poco interpretabili. Questa è una chiara conseguenza del fatto che l'aumentare o il diminuire il numero di passi, comporta una modifica alle dimensioni delle comunità trovate. Più passi fa, più sono grandi le comunità trovate e viceversa.

Analizzando la rete dei corsi di Ingegneria si ottengono13 comunità. Più dettagliatamente, la decima è formata da tre corsi e comprende "elementi metallurgia", "leghe metalliche ottimizzazione processi metallurgici", "leghe metalliche processi metallurgici", ed è quindi riconducibile ad un corso che ha una base importante di metallurgia; la sesta invece è composta dai corsi: "scienza costruzioni", "idraulica", "geotecnica", "costruzioni idrauliche", "misure controlli idraulici", "sistemi idropotabili drenaggio urbano", "tecnica costruzioni", "analisi sperimentale tensioni". Sembra quindi essere legata agli insegnamenti che riguardano l'ambito dell'idraulica.

Anche analizzando la rete dei corsi di Scienze si ottengono 13 comunità. Risulta lampante l'associazione tra la sesta comunità e il corso di studi di statistica. Sono inclusi, tra gli altri, i seguenti insegnamenti: "statistica2", "modelli statistici 1", "statistica computazionale", "teoria tecnica indagine statistica campionamento", "analisi dati multidimensionali", "modelli statistici 2", "istituzioni probabilita".
La seconda comunità sembra riconducibile ad un insegnamento di geologia dato che i relativi insegnamenti sono: "rilevamento geologico 2", "cartografia informatizzata", "geologia strutturale", "paleontologia", "geologia italia", "geologia stratigrafica", "geologia applicata". La terza invece è attribuibile ai corsi di chimica e biologia. Molto probabilmente in questo caso l'algoritmo "fonde" i due insegnamenti dato l'elevato numero di collegamenti che ci sono tra i suddetti corsi di laurea. La quarta invece è composta da 83 esami, la stragrande maggioranza dei quali appartiene all'insegnamento di "fisica" ed è appunto riconducibile a questo corso di studi.

Si noti come le comunità risultanti siano più comparabili ai corsi di laurea all'interno della scuola di Scienze. Questa differenza è in parte attribuibile alle descrizioni dei prerequisiti dei corsi di studio presenti sul sito dell'ateneo che risultano essere più "chiare" per le lauree di scienze. Un altro fattore importante che spiega questa differenza di interpretabilità è dato dal diverso numero di archi della rete di scienze e di ingegneria (rispettivamente 1697 contro 432). E' importante anche notare che la scuola di Ingegneria presenta molti corsi di laurea simili per nomi e materia di studio, il che complica la distinzione tra uno e l'altro che cerca di attuare l'algoritmo.

Inoltre si denotano comunità che non rispecchiano pedissequamente né i corsi di laurea, né i settori scientifico disciplinari che sono le due principali covariate di nodo. Questo discende dalla composizione dei corsi all'interno delle lauree nelle due scuole considerate: all'interno dello stesso insegnamento possono essere presenti corsi con settori scientifico disciplinari diversi e, allo stesso modo, in diversi corsi di laurea possono essere presenti esami attribuiti allo stesso settore scientifico disciplinare.

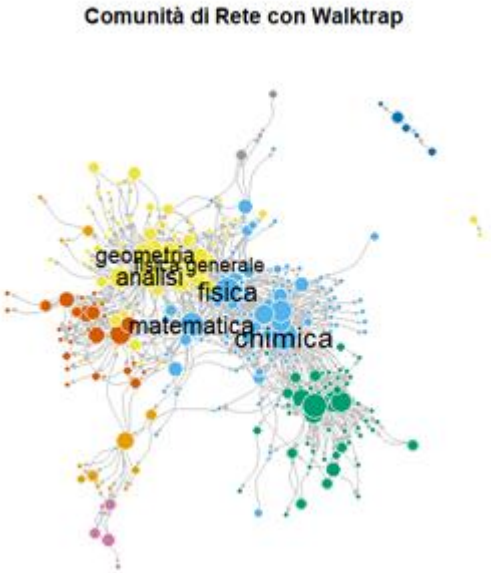

Figura 5. Comunità della scuola di Scienze, con etichette sugli Hub.

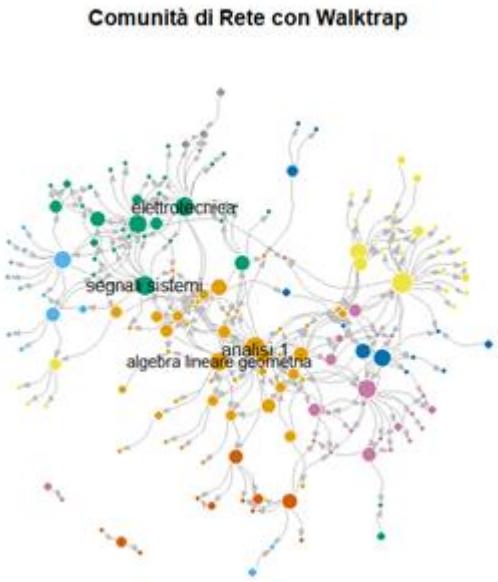

Figura 6. Comunità della scuola di Ingegneria, con etichette sugli Hub.

CONCLUSIONI

Questa esperienza ci ha permesso di mettere in pratica le nostre capacità di problem solving: abbiamo imparato a valutare le diverse opzioni metodologiche e ad adattare le nostre strategie in base alle specifiche esigenze del progetto.

Siamo stati in grado di esplorare le interazioni tra i nodi della rete e di rivelare i pattern emergenti che si nascondono dietro di esse. Questo ci ha permesso di ottenere una comprensione più approfondita delle dinamiche delle reti e dei meccanismi che le governano. Inoltre, il progetto ci ha offerto l'opportunità di sviluppare le nostre competenze nell'utilizzo del software R attraverso l'applicazione pratica di tecniche e algoritmi specifici.

CRITICITA'

-Molti prerequisiti sono scritti in una forma poco consona per il tipo di analisi effettuata, poiché non sono stati compilati utilizzando una nomenclatura standard. Altri invece sono incompleti o completi solo parzialmente. Una delle criticità riscontrate è che lo stesso insegnamento può presentarsi scritto in forme diverse (e.g. "Analisi 1" e "Analisi matematica 1" indicano lo stesso esame). Un altro esempio, è il corso di "sistemi 1" il cui nome completo (e corretto) è "sistemi di elaborazione 1". Un altro problema è il modo in cui sono stati scritti alcuni corsi che prevedono un insegnamento successivo: per esempio c'è un esame che ha come prerequisiti i due corsi "sistemi di elaborazione 1" e "sistemi di elaborazione 2". Il problema sta nel fatto che nei prerequisiti i due corsi sono scritti come segue: "sistemi di elaborazione 1 e 2". In questo modo non è possibile creare un arco in corrispondenza di "sistemi di elaborazione 2" perché il nome del corso non è scritto correttamente nei prerequisiti.

- L'algoritmo Walktrap, per poter funzionare, rende indiretta la rete diretta. Questo approccio, sebbene non sia del tutto corretto poiché ignora la struttura intrinseca della rete degli insegnamenti, è stato considerato sensato in fase di community detection. Ciò è dovuto al fatto che, al fine di valutare la struttura delle comunità all'interno della rete, l'informazione derivante dalla presenza effettiva (o assenza) di un arco è più importante della sua direzione.

BIBLIOGRAFIA

1. "RCrawler: An R package for parallel web crawling and scraping", Salim Khalil and Mohamed Fakir, 2017
2. "Computing communities in large networks using random walks", Pascal Pons and Matthieu Latapy, 2005

SITOGRAFIA

Sito della didattica dell'Università di Padova, https://didattica.unipd.it/off/2022/LT