Dispersion centrality: applicazione della dispersione in casi di studio reali

Amedeo Leo | Alessio Petrozziello | Simone Romano
amedeo.leo92@gmail.com | alessio92p@gmail.com | s.romano1992@gmail.com
Università degli Studi di Salerno

Abstract

In questo lavoro si vogliono mostrare alcuni casi di studio reali in cui la dispersione[1](una nuova misura per Social Network Analysis) è utilizzata per evidenziare nodi all'interno di una rete con particolari caratteristiche. In particolare verrà mostrato come la ***Dispersion centrality****(nuova misura che introdurremo) insieme alla Betweenness centrality rappresentano un metodo efficace per isolare nodi che si differenziano dal resto della rete.*

1. Introduzione

L'articolo [1] mostra la dispersione, un'unità di misura applicabile agli archi di un grafo. Lo studio proposto parte da un'analisi di tale misura per confrontarla con alcune delle misure già presenti e diffuse in letteratura(degree centrality, closeness centrality, betweenness centrality) e punta a cercare una relazione con queste. In particolare scopo dello studio è utilizzare la dispersione in unione ad una delle altre misure per ricavare informazioni aggiuntive inerenti la struttura della rete analizzata. Ciò che semanticamente rappresenta la dispersione è **quanto amici in comune di due nodi non sono ben collegati**. Innanzitutto quello che è stato fatto è portare una misura applicata ad archi(quale la dispersione) ad un singolo nodo, introducendo la **dispersion centrality** definita, a partire da un nodo, come la somma delle dispersioni da tutti i suoi vicini. Per ricavare dei risultati valutabili è stata implementata una libreria scritta in python utilizzando la libreria [2]. Tale libreria consente, dato un grafo in input in formato *edge*, di calcolare per tutti i nodi del grafo alcune delle misure principali utilizzate in social network analysis aggiungendo dunque la dispersion centrality e di stampare i risultati in formato *.cvs* analizzabile poi con altri tool descritti in seguito. I paragrafi successivi formalizzeranno in maniera più precisa il concetto di dispersione così come proprosto in [1]; a questo punto verranno dettagliate le metodologie utilizzate, i casi di studio e i risultati. Verranno dunque proposte delle conclusioni.

2. Dispersione

Nella sociologia matematica, i legami interpersonali sono definiti come connessioni che "inglobano" delle informazioni tra le persone. La forza di un legame costituisce una dimensione importante in cui si possono caratterizzare i collegamenti di una persona ai suoi "vicini di rete". Informalmente, si riferisce alla vicinanza di un collegamento: definisce un insieme che varia da legami forti a deboli; i primi sono spesso incorporati nella rete, circondati da un vasto numero di vicini, mentre gli ultimi coinvolgono spesso pochi vicini e sono utilizzati come "ponti" per diverse zone della rete, fornendo accesso a importanti informazioni. Il punto critico è dunque identificare gli individui più importanti in una rete sociale. La caratteristica fondamentale in queste analisi è l'embeddedness ("annidamento" o "radicamento"): è una quantità che tipicamente aumenta con la forza dei legami, poiché rappresenta il numero di persone che due nodi vicini hanno in comune. Tuttavia, è un mezzo debole per poter identificare le relazioni di una rete. In questo lavoro proponiamo una misurazione alternativa, chiamata dispersione, che è significativamente più efficace. Tale misura non prende in considerazione solo il numero di individui comuni a due persone, ma anche alla struttura della rete determinata da questo nu-

mero; approssimativamente, una connessione tra due persone ha un'alta dispersione quando i loro vicini non sono connessi tra loro. La dispersione riflette la teoria del "social foci": molte persone hanno grandi gruppi (o "cluster") di amici o conoscenti corrispondenti a interazioni ben definite nella loro vita, come, nel nostro caso di studio, individui che frequentano un dojo e relativi istruttori oppure una rete criminale con i corrispondenti boss. Poiché molte persone all'interno di tali cluster si conoscono, questi ultimi contengono dei link con alta embeddedness, anche se non corrispondono necessariamente a legami particolarmente forti. Al contrario, i collegamenti agli amici di una persona possono avere radicamento minore, ma coinvolgono i vicini comuni di investimento da numerosi e diversi foci. Quindi, invece dell'embeddedness, ipotizziamo che i collegamenti tra un nodo u e un suo vicino v riflettano una struttura "dispersa": in tal modo, i vicini comuni di u e v non sono ben connessi l'un l'altro, e quindi u e v agiscono congiuntamente come gli unici intermediari tra queste diverse parti della rete.

Per poter definire la dispersione in termini matematici, formuliamo una serie di definizioni: per iniziare, consideriamo G_u come il sottografo indotto su u e sui suoi vicini, e per ogni nodo v definiamo C_{uv} come l'insieme dei vicini comuni di u e v. Per esprimere l'idea che le coppie di nodi in C_{uv} devono essere distanti in G_u quando non consideriamo i path in due step attraverso u e v stessi, definiamo la dispersione assoluta della collegamento u-v, disp(u,v), come la somma di tutte le distanze a coppie tra i nodi in C_{uv}, misurata in G_u - u, v; ovvero,

$$disp(u,v) = \sum_{s,t \in C_{uv}} d_v(s,t)$$

dove d_v è la distanza tra due nodi in C_{uv}; tale distanza d_v (s,t) equivale a 1 se s e t non sono direttamente connessi e non hanno nodi in comuni in G_u, oltre a u e v stessi, e equivale a 0 altrimenti. Nei nostri casi di studio intervengono anche altre misurazioni: la dispersione normalizzata, norm(u,v), definita come il rapporto tra la dispersione e l'embeddness; da questa sono state create altre due metriche per aumentare le prestazioni: la prima è la dispersione parametrizzata, definita da:

$$\frac{(disp(u,v)+b)^{\alpha}}{emb(u,v)+c}$$

La seconda è l'applicazione della dispersione in maniera ricorsiva: individuare i nodi v il cui collegamento u-v raggiunge una elevata dispersione normalizzata sulla base di una serie di vicini comuni C_{uv}, che, a loro volta, hanno alta dispersione normalizzata nei loro legami con u. Per realizzare questa idea, assegniamo valori ai nodi che riflettono la dispersione nei loro legami con u, e quindi aggiorniamo questi valori con quelli di dispersione associati ad altri nodi. In particolare, definiamo x_v= 1 per tutti i vicini v di u, e aggiorniamo iterativamente ogni x_v:

$$x_v = \frac{\sum_{w \in C_{uv}} x_w^2 + 2\sum_{s,t \in C_{uv}} d_v(s,t) x_s x_t}{emb(u,v)+c}$$

Tuttavia, tali misurazioni, data una coppia di nodi definita da un arco, determinano un insieme di valori: per poter confrontare i risultati ottenuti, occorrerebbero nuove funzioni che restituiscono una sola misura. Per questo obiettivo, abbiamo definito diverse nuove metriche: centralità di dispersione, dispersione massima, dispersione minima, dispersione media. Tali funzioni non si basano su una coppia di nodi, ma su un singolo nodo: determinano quanto questo nodo venga coinvolto nel calcolo delle dispersioni dei suoi vicini. Sommando le dispersioni di tali vicini, ne abbiamo determinato l'effettiva misurazione, la dispersione massima, la minima e la media. Inoltre, abbiamo effettuato i test sui grafi anche su altre funzioni ben note: la betweenness centrality, la degree centrality e la closeness centrality. La prima, per un nodo v, è determinata dall'espressione

$$g(v) = \sum_{s \neq v \neq t} \frac{\sigma_{s,t}(v)}{\sigma_{s,t}}$$

Dove σ_{st} è il numero totale di cammini minimi da s a t e σ_{st} è il numero di tali cammini che attraversano v. La degree centrality è stata una delle prime metriche sviluppate: è definita come il numero di archi incidenti su un nodo,

ovvero la quantità di legami che ha un nodo. Nei nostri casi di studio, gli archi non sono direzionati, quindi useremo solo una misurazione (e non indegree o outdegree). Il significato dato da tale metrica è rappresentato dal fatto che gli individui più importanti sono coloro con più legami nella rete. La closeness centrality, invece, misura quanti passaggi devono essere fatti, partendo da un nodo v, per raggiungere il maggior numero possibile di nodi. In tal caso, la persona più importante può raggiungere facilmente le altre nella rete.

3. Metodologie

Durante lo sviluppo di questo progetto ci siamo avvalsi del supporto di diverse tecnologie per la manipolazione e visualizzazione dei grafi tra cui:

- Python
- Xml
- Csv
- Snap
- Gephi
- Cran-R
- Weka

3.1. Python

Python è un linguaggio di programmazione dinamico orientato agli oggetti utilizzabile per molti tipi di sviluppo software. Offre un forte supporto all'integrazione con altri linguaggi e programmi, è fornito di una estesa libreria standard e può essere imparato in pochi giorni. Molti programmatori Python possono confermare un sostanziale aumento di produttività e ritengono che il linguaggio incoraggi allo sviluppo di codice di qualità e manutenibilità superiori. Python gira su Windows, Linux/Unix, Mac OS X, OS/2, Amiga, palmari Palm e cellulari Nokia; è stato anche portato sulle macchine virtuali Java e .NET.

3.2. Snap

Stanford Network Analysis Platform (SNAP) è una libreria per network analysis e graph mining. E' scritta in C++ e scala facilmente con grafi massivi con centinaia di milioni di nodi e miliardi di archi. E' capace di manipolare grafi molto grandi, calcolare proprietà strutturali, generare graphi random e regolari e supporta gli attributi su nodi e archi. Snap.py è una interfaccia Python per SNAP. Essa provvede le performance di SNAP, uniti alla flessibilità di Python. Molte delle funzioni di SNAP in C++ sono disponibili nella libreria Snap.py

3.3. Gephi

Gephi è un tool di visualizzazione ed esplorazione interattiva di qualsiasi tipo di rete, sistemi complessi, grafi dinimaci e gerarchici. Gephi gira su Windwos, Linux e OS X. E' opensorce e free

3.4. Implementazione

Varie funzioni necessarie e di supporto sono state sviluppate tra cui:

- fromGexfToEdge: permette la trasformazione da un file di gephi ad un file di tipo EDGE, importabile in SNAP
- commonNeighbors: permette di trovare il vicinato comune tra due nodi
- dispersion: calcola la dispersione tra due nodi, così come da formula
- embeddedness: corrisponde al vicinato comune tra due nodi
- norm: calcola la dispersione normalizzata

$$\frac{dispersione}{embeddedness}$$

- performance: calcola la norma aggiungendo dei parametri costanti di ottimizzazione
- recDisp: calcola la dispersione ricorsiva da un nodo a tutti gli altri come da formula
- printNodeInformations_XML: crea un file XML con tutti le informazioni su ogni nodo del grafo (degree centrality, betweenness centrality, closeness centrality, dispersion centrality)
- dispersionCentrality: calcola la dispersione su ogni nodo vicino e ne effettua la somma

- printNodesInformations_CSV: crea un file CSV con tutte le informazioni sui nodi, standard facilmente importabile in CRAN-R o Weka, tool di data analisys

4. Casi di studio

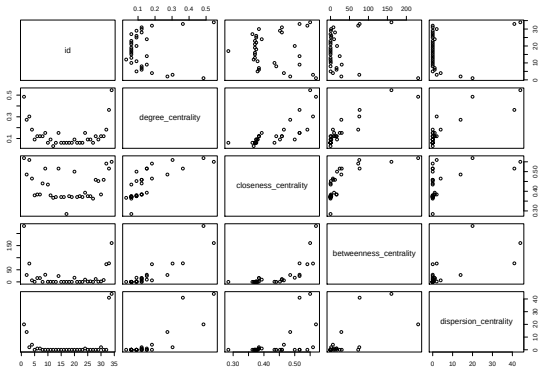
Al fine di valutare quanto ipotizzato in fase di analisi del problema sono stati studiati alcuni casi reali. In particolare il primo caso studiato è stato quello rappresentatne un club di karate(presentato in [3]). Lo studio [3] risale al 1977. Gli autori hanno osservato il comportamento del club di karate dal 1970 al 1972, registrando ciò che accadeva all'interno del club. In particolare il club era formato da due istruttori i quali, a seguito di litigi dovuti a disaccordi sul prezzo delle lezioni, litigarono. La rete si divise in due parti, ciascuna con a capo uno degli istruttori e i relativi allievi. Infine, ulitmo caso di studio è il grafo di una rete criminale. Ciò che si vuole mostrare è che la dispersion centrality consente di evidenziare nodi che rappresentano nodi particolari(dal punto di vista semantico) all'interno del grafo.

5. Risultati

Di seguito sono mostrati e commentati i risultati dei test effettuati sui vari casi di studio. Le metriche utilizzate per analizzare i risultati della dispersion centrality sono degree centrality, closeness centrality e betweenness centrality.

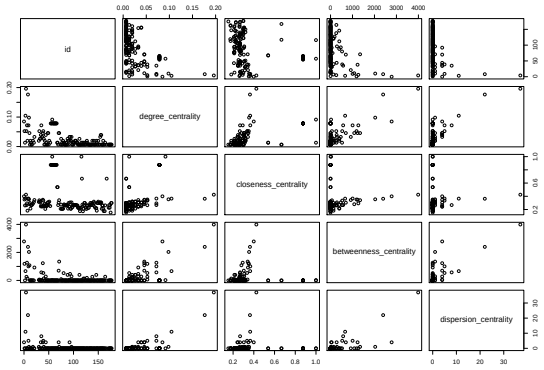
5.1. Karate

Nella figura seguente si evince che, nella rete del Karate analizzata, esiste una netta correlazione tra betweenness centrality e dispersion centrality, come si può notare nel relativo quadrante. In particolare gli outlier rappresentano i nodi 1, 34, 33 relativi ai 3 nodi più importanti della rete ovvero il capo istruttore e i 2 maestri.



5.2. Rete criminale

La figura seguente mostra i risultati ottenuti applicando le stesse metriche ad una rete criminale. Ciò che si evince è, ancora una volta, che gli outlier evidenziati dalla betweenness centrality sono gli stessi evidenziati dalla dispersion centrality.



6. Conclusioni

Partendo dall'analisi della dispersione[1](unità applicabile ad un arco di una rete) abbiamo considerato, dato un nodo, la somma delle dispersioni tra questo nodo ed i suoi vicini(dispersion centrality) in modo da ottenere una misura confrontabile con degree centrality, closeness centrality e betweenness centrality. Ciò che è emerso dai risultati sui dataset utilizzati è

quanto segue: la dispersion centrality fornisce le stesse informazioni della betweenness centrality. La differenza sostanziale è che il calcolo della dispersion centrality è **parallelizzabile**.

References

[1] Romantic Partnerships and the Dispersion of Social Ties: A Network Analysis of Relationship Status on Facebook - http://arxiv.org/pdf/1310.6753.pdf

[2] Snap http://snap.stanford.edu/

[3] Journal of Anthropological Research, Vol. 33, No. 4 (Winter, 1977), pp. 452-473 http://www1.ind.ku.dk/complexLearning/zachary1977.pdf