Verifica delle proprietà dei grafi di Kronecker

Irene Dini

2 luglio 2018

1 Introduzione

Questo progetto illustra alcuni esperimenti effettuati su grafi di Kronecker: lo scopo è quello di verificare se essi rispettano alcune proprietà che caratterizzano la rete sociale a partire dalla quale sono stati generati. Ci siamo concentrati in particolare sulla distribuzione dei gradi e sul coefficiente di clustering. Dato che quest'ultima quantità è definita solo per nodi di grafi non diretti, abbiamo preso in considerazione due reti sociali non dirette.

1.1 Strumenti utilizzati

Gli strumenti utilizzati per questo progetto sono python per il pre-processing, l'elaborazione e la visualizzazione dei dati, e gli algoritmi KRONFIT e KRONGEN disponibili su Snap Stanford.

1.1.1 KRONFIT

KRONFIT è l'algoritmo che, presa in input la lista di archi del grafo $\mathcal{G}$ da "simulare" e la dimensione desiderata N_1, stima l'*initiator matrix* $\mathbf{K_1}$ (di dimensione $N_1 \times N_1$) da cui generare il grafo di Kronecker, utilizzando il metodo della discesa del gradiente e l'algoritmo di Metropolis. Lo pseudocodice è riportato nel seguente listato:

```
input: size of parameter matrix N_1, graph G, and learning rate λ
output: MLE parameters Θ̂(N_1 × N_1 probability matrix)

initialize Θ̂
while not converged do
  evaluate gradient: δ/δΘ̂_t l(Θ̂_t)
  update parameter estimates: Θ̂_{t+1} = Θ̂_t + λ δ/δΘ̂_t l(Θ̂_t)
end
return Θ̂ = Θ̂_t
```

Per come è strutturato l'algoritmo, se si lavora con grafi non diretti è necessario un leggero pre-processing: sia

$$EL = \{(v_0, v_1), (v_2, v_3), \ldots, (v_{n-1}, v_n)\}$$

la lista degli archi del grafo G, per ogni arco $(v_i, v_j) \in EL$, bisogna inserire in EL anche (v_j, v_i).

1.1.2 KRONGEN

KRONGEN è invece l'algoritmo che, data l'*initiator matrix* $\hat{\mathbf{\Theta}}$ di dimensione $N_1 \times N_1$ ed il numero di iterazioni da eseguire k, genera un grafo con N_1^k nodi, la cui matrice di adiacenza è appunto la matrice di Kronecker $\mathbf{K_k} = \underbrace{\mathbf{K_1} \otimes \mathbf{K_1} \otimes \ldots \otimes \mathbf{K_1}}_{\text{k volte}}$. Il numero di iterazioni viene calcolato secondo la regola: $N_1^{k-1} < N(G) \leq N_1^k$. Indichiamo con $\hat{\mathbf{\Theta}} = \{\theta_{ij}\}$ e $E_1 = \sum_{ij} \theta_{ij}$

```
input: Initiator matrix Θ̂, number of iterations k
output: Kronecker Matrix K_k with N_1^k nodes

repeat E_1^k times
  sample edge: simulate the Kronecker product k times, each time choosing
      an edge from Θ̂ with probability θ_ij / E_1
  add edge to K_k
return K_k
```

La matrice restituita avrà un numero di nodi che tende a E_1^k. In questo caso è però necessario del post-processing: anche se la matrice $\hat{\mathbf{\Theta}}$ è simmetrica, data la natura probabilistica del processo, è molto poco probabile che il grafo di Kronecker ottenuto risulterà essere non diretto. Per renderlo tale, basta semplicemente considerare solo gli archi (v_i, v_j) con $i \leq j$. Questa operazione equivale a considerare solo la parte triangolare inferiore della matrice di adiacenza del grafo.
NB: In realtà si potrebbe anche cosiderare la parte triangolare superiore: da un'unica matrice di Kronecker "diretta", si possono generare due grafi non diretti.

2 Esperimenti

Le reti sociali su cui abbiamo scelto di lavorare sono non dirette e non pesate.

2.1 Grafo di Facebook

La prima rete sociale presa in considerazione è una parte del grafo di Facebook disponibile su Snap Stanford. Le sue caratteristiche sono le seguenti:

- Nodi: 4039
- Archi: 88234
- Diametro: 8

- Nodi GCC: 4039
- Grado medio: 43.6910

Dopo il pre-processing abbiamo applicato KRONFIT e KRONGEN per ottenere 4 diversi grafi di Kronecker e confrontare le loro caratteristiche con quelle del grafo di partenza.

2.1.1 Kronfit

Per prima cosa abbiamo applicato KRONFIT al grafo di Facebook, per ottenere una initiator matrix di dimensione 2×2 e una di dimensione 3×3:

$$\hat{\Theta}_2 = \begin{pmatrix} 0.9999 & 0.691526 \\ 0.691459 & 0.349524 \end{pmatrix}$$

$$\hat{\Theta}_3 = \begin{pmatrix} 0.4372 & 0.6576 & 0.08469 \\ 0.6544 & 0.9999 & 0.3151 \\ 0.084 & 0.3151 & 0.9999 \end{pmatrix}$$

Come ci si poteva aspettare, dato che il grafo in ingresso è non diretto, entrambe le matrici risultano praticamente simmetriche.

2.1.2 Krongen

A partire dalle *initiator matrix*, dopo aver calcolato il numero di iterazioni da fare, abbiamo applicato KRONGEN per ottenere dei grafi di Kronecker. Studiamo i due casi separatamente.
Per quanto riguarda $\hat{\Theta}_2$ si ha che:

$$k = 12 \text{ infatti: } 2^{11} = 2048 < 4039 \leq 4096 = 2^{12}$$

Il grafo $\mathcal{G}_2$ ottenuto, dopo il post-processing, ha le seguenti caratteristiche:

- Nodi: 4096
- Archi: 86365 (173200 prima del post-porocessing)
- Diametro: 5
- Nodi GCC: 4096
- Grado medio: 42.1704

Escludendo il diametro, che risulta essere piuttosto più piccolo, tutte le quantità stimate si avvicinano molto a quelle del grafo di partenza.
Per quanto riguarda $\hat{\Theta}_3$ invece:

$$k = 8 \text{ infatti: } 3^7 = 2187 < 4039 \leq 6561 = 3^8$$

Le caratteristiche del grafo G_3 sono:

- Nodi: 6561
- Archi: 91346 (183012 prima del post-porocessing)
- Diametro: 5
- Nodi GCC: 6561
- Grado medio: 27.8451

In questo caso notiamo subito che le caratteristiche del grafo di Kronecker sono abbastanza lontane da quelle del grafo di partenza: il numero di nodi ed il numero di archi risultano essere più alti, mentre il diametro ed il grado medio sono molto più bassi.

2.2 Studio delle proprietà

Abbiamo fatto uno studio grafico della distribuzione dei gradi e del coefficiente di clustering, per vedere quanto i grafi $\mathcal{G}_2$ e $\mathcal{G}_3$ si avvicinassero al grafo originario $\mathcal{G}$. Il grafico 1 rappresenta la distribuzione dei gradi dei 3 grafi: ad ogni grado è associato il numero di nodi che hanno quel grado.

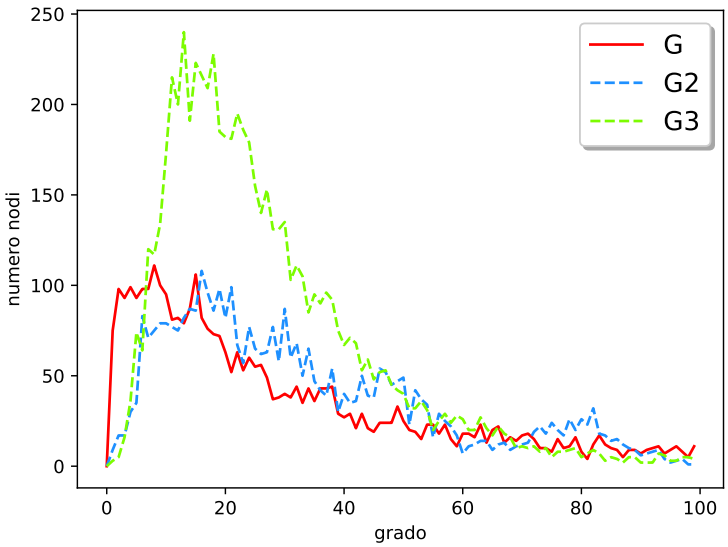

Figura 1: Distribuzione dei gradi

Come si può notare, la distribuzione dei gradi di $\mathcal{G}_2$ non si discosta molto da quella di $\mathcal{G}$, si ha una differenza sostanziale solo per i gradi che vanno da 1 a 5. La distribuzione dei gradi di $\mathcal{G}_3$ è invece molto diversa.
Vediamo adesso se queste distribuzioni seguono una power-law, trasferendo il grafico in scala logaritmica (Figura 2).

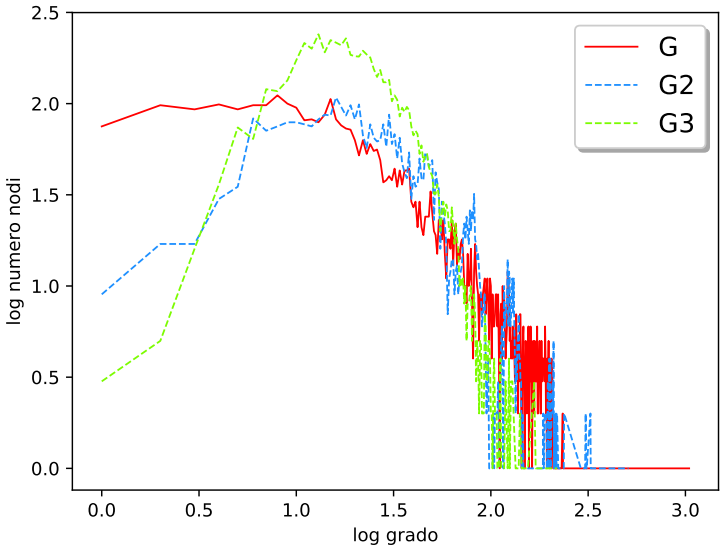

Figura 2: Degree Power-Law

L'andamento dei 3 grafici è molto simile: da 1.3 circa in poi, anche se molto oscillatorio, ricorda una retta. Non si può comunque concludere che i grafi di Kronecker seguano una power law ma, dato che non la seguiva nemmeno il grafo di partenza, era prevedibile.

Come ultima cosa, abbiamo confrontato i coefficienti di clustering dei nodi. Nel grafico in figura 3 ad ogni nodo è associato il suo coefficiente di clustering (in ordine decrescente per questo valore).

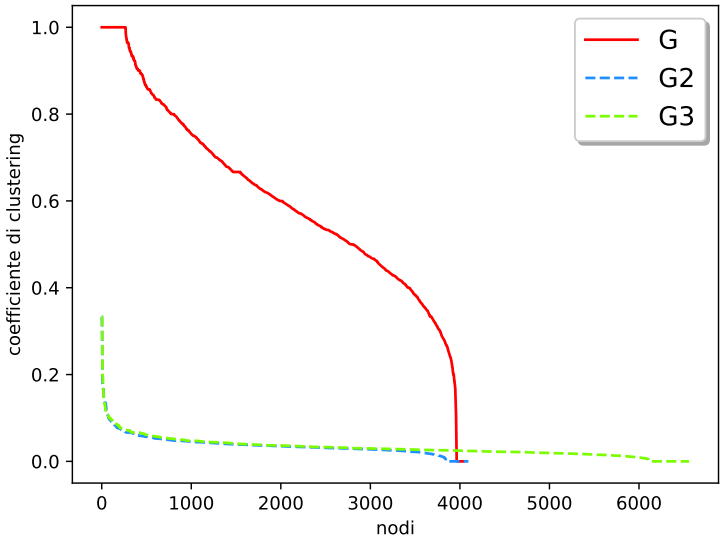

Figura 3: Clustering coefficient

Mentre negli altri due casi $\mathcal{G}_2$ approssimava almeno qualitativamente la distribuzione dei gradi del grafo $\mathcal{G}$, in questo caso si ottiene una distribuzione totalmente differente.

Riportiamo di seguito altri esempi, per vedere se, partendo da sampling diversi, si potessero ottenere risultati migliori. Nei grafici riportati nelle figure 4 e 5 si confronta quindi il grafo $\mathcal{G}_2$ (che portava una migliore approssimazione di $\mathcal{G}$ rispetto a $\mathcal{G}_3$), con altri due grafi generati sempre a partire da una *initiator matrix* di dimensione 2×2. $\mathcal{G}_{2.1}$ è semplicemente un altro sampling di $\hat{\mathbf{\Theta}}_\mathbf{2}$ e le sue caratteristiche sono:

- Nodi: 4093
- Archi: 91346 (173200 prima del post-porocessing)
- Diametro: 5
- Nodi GCC: 4093
- Grado medio: 42.5121

Per $\mathcal{G}_{2.2}$ è stato invcee nuovamente eseguito KRONFIT ottenendo la matrice

$$\hat{\mathbf{\Theta}}_{\mathbf{2.2}} = \begin{pmatrix} 0.9999 & 0.6914 \\ 0.6914 & 0.3501 \end{pmatrix}$$

Da essa, applicando KRONGEN si è ottenuto un grafo con le seguenti caratteristiche:

- Nodi: 4093

- Archi: 86763 (173421 prima del post-porocessing)
- Diametro: 5
- Nodi GCC: 4093
- Grado medio: 42.3958

In entrambi i casi, le caratteristiche elencate sono molto vicine a quelle di $\mathcal{G}_2$ e quindi di $\mathcal{G}$.
Le figure 4 e 5 mostrano come l'andamento dei gradi e del coefficiente di clustering di questi due grafi siano molto simili a quelli di $\mathcal{G}_2$. Questo porta a pensare che non potremo ottenere approssimazioni di $\mathcal{G}$ molto migliori utilizzando i grafi di Kronecker.

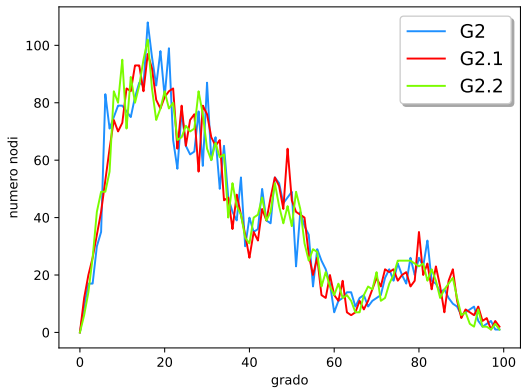

Figura 4: Distribuzione dei gradi

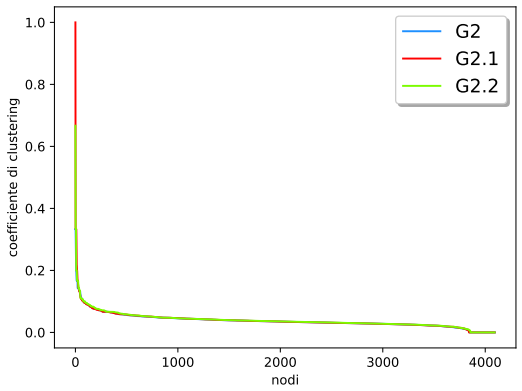

Figura 5: Coefficiente di clustering

2.3 Grafo citazioni

La seconda rete considerata è Arxiv GR-QC (General Relativity and Quantum Cosmology), una rete di citazioni disponibile su Snap Stanford. Rappresenta le collaborazioni tra gli autori che hanno scritto articoli per GR-QC: se un autore i, ha scritto un articolo insieme all'autore j, il grafo contiene un arco non diretto tra i e j. Le caratteristiche di questa rete sono:

- Nodi: 5242
- Archi: 14496
- Diametro: 17
- Nodi GCC: 4158
- Grado medio: 5.5307

2.3.1 Kronfit

Dopo il pre-processing, abbiamo applicato l'algoritmo di KRONFIT per ottenere una *initiator matrix* di dimensione 2×2 e una di dimensione 3×3, cioè:

$$\hat{\boldsymbol{\Theta}}_{\mathbf{2}} = \begin{pmatrix} 0.9999 & 0.3624 \\ 0.3624 & 0.4497 \end{pmatrix}$$

$$\hat{\boldsymbol{\Theta}}_{\mathbf{3}} = \begin{pmatrix} 0.3631 & 0.5434 & 0.01157 \\ 0.5434 & 0.6927 & 0.1968 \\ 0.01157 & 0.197 & 0.9999 \end{pmatrix}$$

Come ci si doveva aspettare, anche in questo caso le matrici ottenute sono simmetriche, in quanto il grafo di partenza è non diretto.

2.3.2 Krongen

Per prima cosa è necessario calcolare il numero di iterazioni k, basandosi sul numero di nodi del grafo di partenza. Per quanto riguarda $\hat{\boldsymbol{\Theta}}_{\mathbf{2}}$ si ha che:

$$k = 13 \text{ infatti: } 2^{12} = 4096 < 5242 \leq 8192 = 2^{13}$$

Applicando KRONGEN abbiamo ottenuto il grafo G_2 con le seguenti caratteristiche:

- Nodi: 6451 (8192 prima del post-processing)
- Archi: 12181 (24289 prima del post-porocessing)
- Diametro: 16
- Nodi GCC: 6119
- Grado medio: 3.7765

Il numero di nodi risulta essere piuttosto maggiore di quello di $\mathcal{G}$ (dovevamo aspettarcelo dato che non esiste una potenza di 2 vicina a 5242), e notiamo come sarebbe stato ancora maggiore se avessimo avuto a che fare con un grafo diretto. Il grado medio ed il numero di archi risultano invece più piccoli.
Per $\hat{\boldsymbol{\Theta}}_{\mathbf{3}}$ si ha invece che:

$$k = 8 \text{ infatti: } 3^{12} = 2187 < 5242 \leq 6561 = 3^8$$

Applicando KRONGEN abbiamo ottenuto il grafo $\mathcal{G}_3$ con le seguenti caratteristiche:

- Nodi: 6158 (6561 prima del post-processing)
- Archi: 12738 (24766 prima del post-porocessing)
- Diametro: 15

- Nodi GCC: 6130
- Grado medio: 4.1371

Le caratteristiche di questo grafo sono simili a quelle del precedente.

Il grafo $\mathcal{G}_2$ prima del post-processing aveva un numero di nodi molto maggiore rispetto a $\mathbf{G}$, ma rendendolo non diretto, dato che non vengono considerati i nodi di grado 0, produce un'approssimazione più precisa.

2.4 Studio delle proprietà

Come per il grafo di Facebook abbiamo per prima cosa studiato l'andamento dei gradi, mettendo a confronto quelle di $\mathcal{G}$, $\mathcal{G}_2$ e $\mathcal{G}_3$, in figura 6

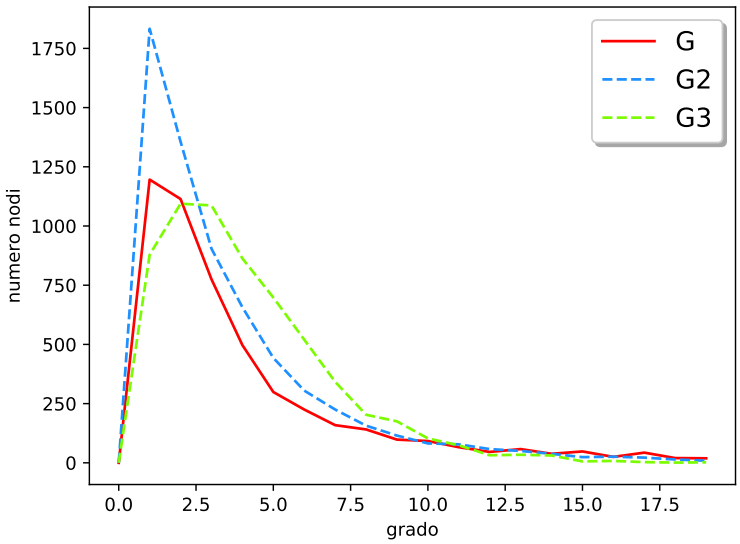

Figura 6: Distribuzione dei gradi

Sono riportati solo i primi 20, in quanto l'andamento dei successivi è praticamente identico per le 3 reti. Notiamo come $\mathcal{G}_2$ abbia un numero di nodi di grado 1 molto più alto di $\mathcal{G}$ ma, per il resto, la sua distribuzione risulti essere una buona approssimazione. La distribuzione di $\mathcal{G}_3$ risulta invece essere abbastanza spostata verso destra, mostrando che il grafo ha, in generale, molti più nodi di grado compreso tra 3 e 5 rispetto a $\mathcal{G}$. Verifichiamo adesso se queste distribuzioni seguono una power-law: in figura 7 è riportato il grafico precedente in scala logaritmica.

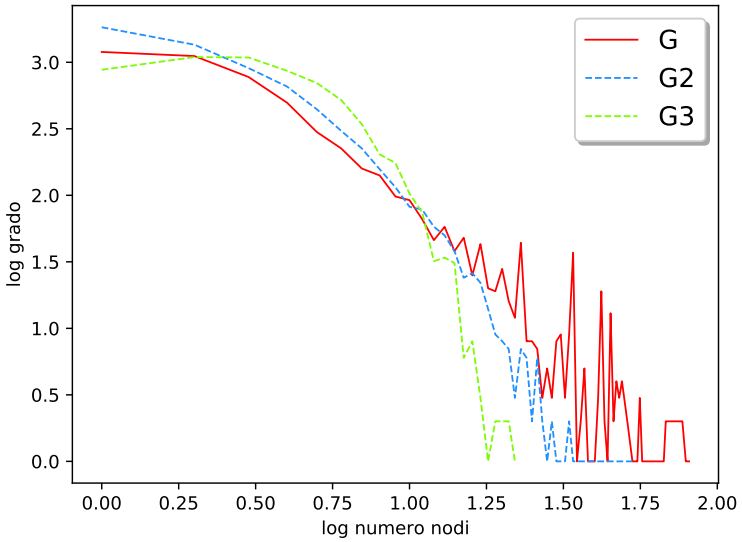

Figura 7: Power-law

La figura mostra come i grafi di Kronecker seguano una degree power-law in modo più fedele rispetto al grafo originale.
Come ultima cosa verifichiamo la distribuzione dei coefficienti di clustering. La figura 8, come per il grafo di Facebook, associa ad ogni nodo il suo coeffciente di clustering, ordinandoli in modo decrescente. Anche questa volta risulta essere totalmente diversa da quella del grafo di partenza.

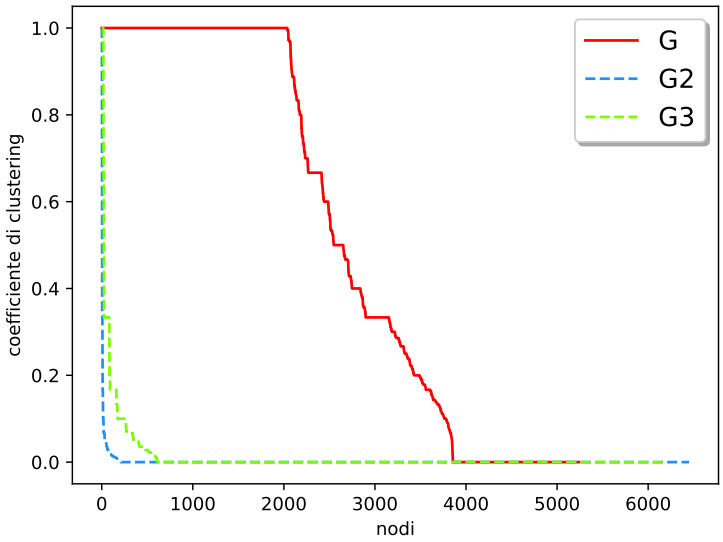

Figura 8: Coefficiente di clustering

3 Conclusioni

Dagli esperimenti condotti si può dedurre che, in realtà, almeno per reti di dimensioni piuttosto contenute, i grafi di Kronecker non producono approssimazioni soddifacenti. Anche semplicemente confrontando caratteristiche "semplici" come il grado medio e il diametro, non è detto che si ottengano valori uguali a quelli del grafo originale. Inoltre, mentre per la distribuzione dei gradi, i grafi di Kronecker ricordano almeno qualitativamente l'andamento del grafo di origine, per quanto riguarda il coefficiente di clustering, si ottengono distribuzioni totalmente diverse dalla sua, indipendentemente dai parametri utilizzati per stimare il modello di generazione.