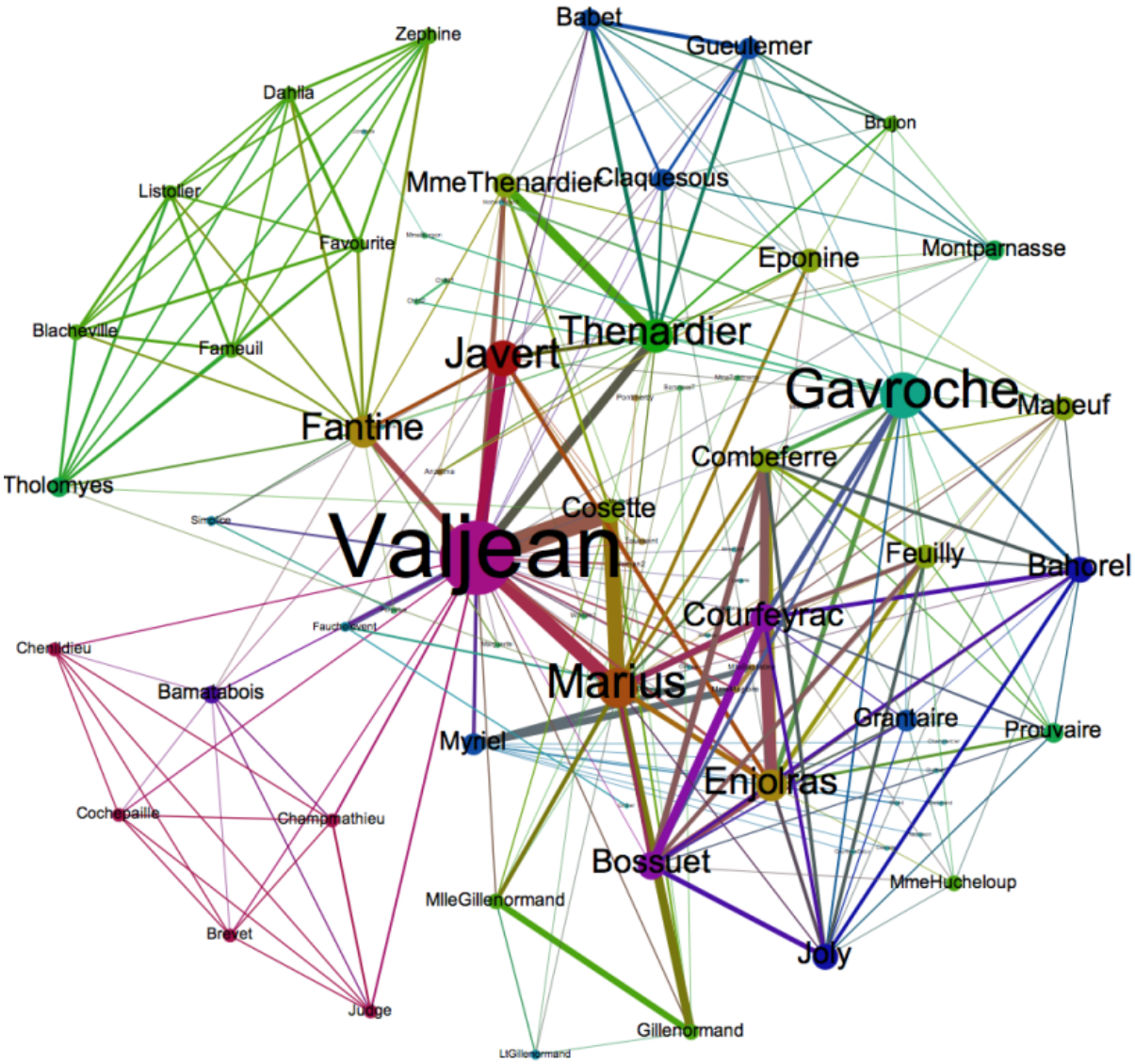
Network Analysis of Les Miserables Graph



Fabio Nan - matricola 2735065.

Graph Analytics a.a 2017/18

INDICE

*Una rete sociale è tecnicamente un **grafo** che rappresenta relazioni tra entità indipendenti. Un grafo è un insieme di nodi (detti anche vertici) V, collegati tra loro da un insieme di archi (spigoli, collegamenti o archi non orientati)*

Perché è importante studiare i Grafi?

Lo studio dei grafi ci aiuta a capire come si formano, come si diffondono le informazioni, come si propagano le epidemie o i malware, la resistenza ai guasti (nel caso di infrastrutture critiche), come evolve il web e a predire l'emergenza di nuovi fenomeni.

Setup

Come nella vecchia esercitazione, anche in questa ho riproposto la combo: Python + NetworkX.

Python è un linguaggio di programmazione ad alto livello interpretato, tirato dinamicamente la cui filosofia di design sottolinea la leggibilità del codice,

Che cos'è NetworkX?

"E' una libreria per Python per la creazione, la manipolazione e lo studio della struttura, delle dinamiche e delle funzioni di reti complesse".

- Strutture dati per rappresentare molti tipi di reti o grafici
- I nodi possono essere qualsiasi oggetto Python (dizionari) e gli archi possono contenere dati arbitrari
- Offre flessibilità per rappresentare reti trovate in molti campi diversi
- Primo rilascio pubblico nell'aprile 2005

```
# read the graph (gml format)
G = nx.read_gml('lesmiserables.gml',relabel=True)

# drawing the full network
figure(1)
nx.draw_spring(G,node_size=0,edge_color='b',alpha=.2,font_size=10)
show()
```

Per Plottare il grafico mi sono appoggiato a ***Pylab*** una libreria che aggiunge un sacco di funzionalità scientifiche.

Siccome nell' esercitazione svolta con i miei compagni Mattia e Leonardo, avevamo raccolto poche misurazioni (grafo troppo grande) ho pensato di provare a misurare su grafo più piccolo.

Il dataset scelto é tratto dal famossissimo libro “les Miserables” e precisamente il dataset catturla la coapparizione dei vari personaggi. (Lo potete trovare https://github.com/gephi/gephi/wiki/Datasets)

In questa rete ogni nodo rappresenta un personaggio e la connessione tra due personaggi é rappresentata dalla contemporanea apparizione nello stesso capitolo.

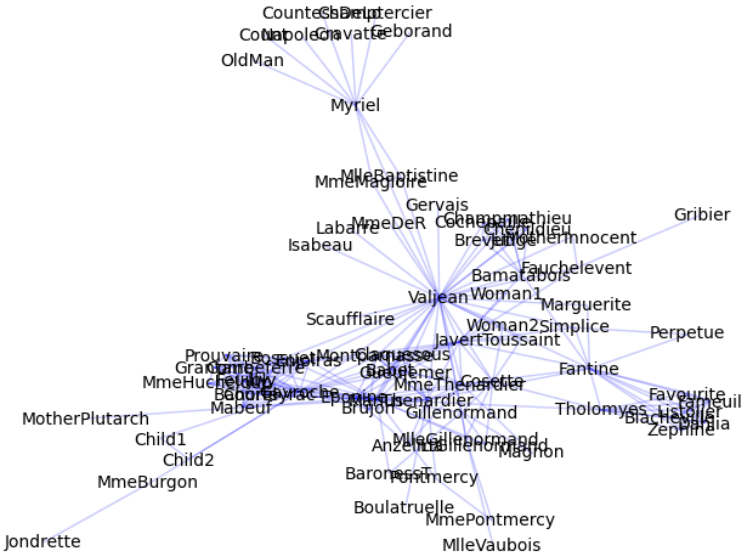
Trama

Les Miserables è uno dei più grandi romanzi storici francesi, scritto da Victor Hugo e pubblicato per la prima volta nel 1862.
Ambientato nella Francia dei primi del XIX secolo, è la storia dell'ex detenuto Jean Valjean (e della sua ricerca di redenzione), che viene inesorabilmente rintracciato da un ispettore di polizia chiamato Javert. Lungo la strada, Valjean (e una lunga lista di personaggi) sono trascinati nel periodo della rivoluzione Francese, dove un gruppo di giovani idealisti fa la sua ultima resistenza in una barricata di strada. Molti personaggi appaiono e scompaiono attraverso le oltre mille pagine del libro, proprio per questo é interessante esempio di rete sociale da studiare.
La storia si svolge nell'era post-napoleonica dopo la rivoluzione francese e parla di un detenuto, Jean Valjean, che è appena stato rilasciato dalla prigione dopo aver scontato 19 anni per aver rubato una pagnotta di pane. Convinto dal vescovo a ricominciare la sua vita, Valjean ricrea la sua identità con un nuovo nome e si trasferisce in una nuova città per diventare un cittadino rispettato e guadagna ricchezza grazie alla sua innovazione nel settore manifatturiero e lo porterà a diventare sindaco. Il resto della storia è ambientato a Parigi, dove Valjean cambia continuamente identità e domicili per evitare il carcere. Durante il suo viaggio, salva e protegge la giovane Cosette dalla terribile famiglia Thénardier come promessa alla sua madre morente, Fantine. Dopo aver trascorso molti anni in un convento, Valjean e Cosette si separano, permettendo a Cosette di vivere una vita più normale c. Cosette si innamora di un giovane avvocato, Marius, che nel seguito si unisce a un gruppo di rivoluzionari in una barricata. Con sorpresa di Marius, Valjean è anchesso sulla barricata e salva il ferito Marius e lo fa tornare dalla sua famiglia e da Cosette. È quando è al suo letto di morte che Valjean è in grado di smettere di nascondersi e rivelare la sua vera identità, e finalmente trovare la pace.
Come avrete intuito, Jean Valjean, è il personaggio centrale del romanzo

Cerchiamo di capire ora se la sua importanza sociale é rispecchiata nelle misure del grafo.

Il Grafo

Un grafo è idealmente rappresentato da cerchi, chiamati nodi e linee che connettono questi nodi, chiamati archi.



Ogni label rappresenta un personaggio che appare nel romanzo. Ogni linea rappresenta un' associazione tra i personaggi. Non ci sono cappi, quindi é un grafo semplice.

Plottando il grafo non si riesce a vedere nulla, l' unica cosa che balza agli occhi è che il grafo è **un unica componente connessa forte**, cioè ogni nodo può essere raggiunto da un altro tramite collegamento diretto.

Proviamo a capire qualcosa in più dai numeri:

```
graph G has 77 nodes and 254 edges
```

Quindi ci sono almeno 77 personaggi e ogni volta che due personaggi di Les Miserables compaiono in un capitolo (per la precisione 254), viene creata una relazione simmetrica tra loro: se il personaggio x appare nello stesso capitolo di y, allora y appare anche nello stesso capitolo di x, quindi il grafo é **undirected.**

Centralità

Our view is that centrality is only a descriptive property of a network. An area of future research should be concerned with innovative uses of centralities to describe how networks may change over time or to determine the consequences of new scenarios when nodes or lines are added or deleted.

Karen Stephenson and Marvin Zelen

Centrale rispetto a cosa? in base a ciò che vogliamo vedere abbiamo 3 metriche per rappresentare la centralità.

Grado : *Il grado di un nodo è il numero di archi adiacenti ad esso, cioè il numero dei suoi vicini, cioè avere "tanti amici"*
Nel caso del grafico corrispondente a Les Miserable, avremmo potuto supporre prima di effettuare la misurazione che il protagonista corrisponda al nodo del grado più alto. Questo nodo è quello associato a Jean Valjan, che è il personaggio principale del libro il grado di questo nodo è 32. Il secondo personaggio più popolare è Gavroche che ha grado 22, un ragazzo che vive per le strade di Parigi e svolge un ruolo breve ma significativo.
Questi sono i primi 20 personaggi con grado piu elevato, mentre **il grado medio é 6,574.**

```
-- distribution of first 20 nodes --
  [+] Valjean has degree 36
  [+] Gavroche has degree 22
  [+] Marius has degree 19
  [+] Javert has degree 17
  [+] Thenardier has degree 16
  [+] Fantine has degree 15
  [+] Enjolras has degree 15
  [+] Courfeyrac has degree 13
  [+] Bossuet has degree 13
  [+] Joly has degree 12
  [+] Bahorel has degree 12
  [+] MmeThenardier has degree 11
  [+] Mabeuf has degree 11
  [+] Feuilly has degree 11
  [+] Eponine has degree 11
  [+] Cosette has degree 11
  [+] Combeferre has degree 11
  [+] Myriel has degree 10
  [+] Gueulemer has degree 10
  [+] Grantaire has degree 10
```

```
-- closeness of first 20 nodes --
  [+] Valjean (closeness level: 0.6440677966101694)
  [+] Marius (closeness level: 0.5314685314685315)
  [+] Thenardier (closeness level: 0.5170068027210885)
  [+] Javert (closeness level: 0.5170068027210885)
  [+] Gavroche (closeness level: 0.5135135135135135)
  [+] Enjolras (closeness level: 0.4810126582278481)
  [+] Cosette (closeness level: 0.4779874213836478)
  [+] Bossuet (closeness level: 0.475)
  [+] Gueulemer (closeness level: 0.4634146341463415)
  [+] Babet (closeness level: 0.4634146341463415)
  [+] MmeThenardier (closeness level: 0.46060606060606063)
  [+] Fantine (closeness level: 0.46060606060606063)
  [+] Montparnasse (closeness level: 0.4578313253012048)
  [+] Claquesous (closeness level: 0.4523809523809524)
  [+] MlleGillenormand (closeness level: 0.4418604651162791)
  [+] Gillenormand (closeness level: 0.4418604651162791)
  [+] Myriel (closeness level: 0.4293785310734463)
  [+] Bamatabois (closeness level: 0.42696629213483145)
  [+] Simplice (closeness level: 0.4175824175824176)
  [+] MmeMagloire (closeness level: 0.41304347826086957)
```

Closeness: stima il grado di vicinanza di un nodo con altri.Intuitivamente misura gli step necessari per diffondere un informazione, fornisce una misura della rapidità di propagazione dell'informazione da un certo nodo. *misura la distanza media di un vertice rispetto ad altri vertici.*

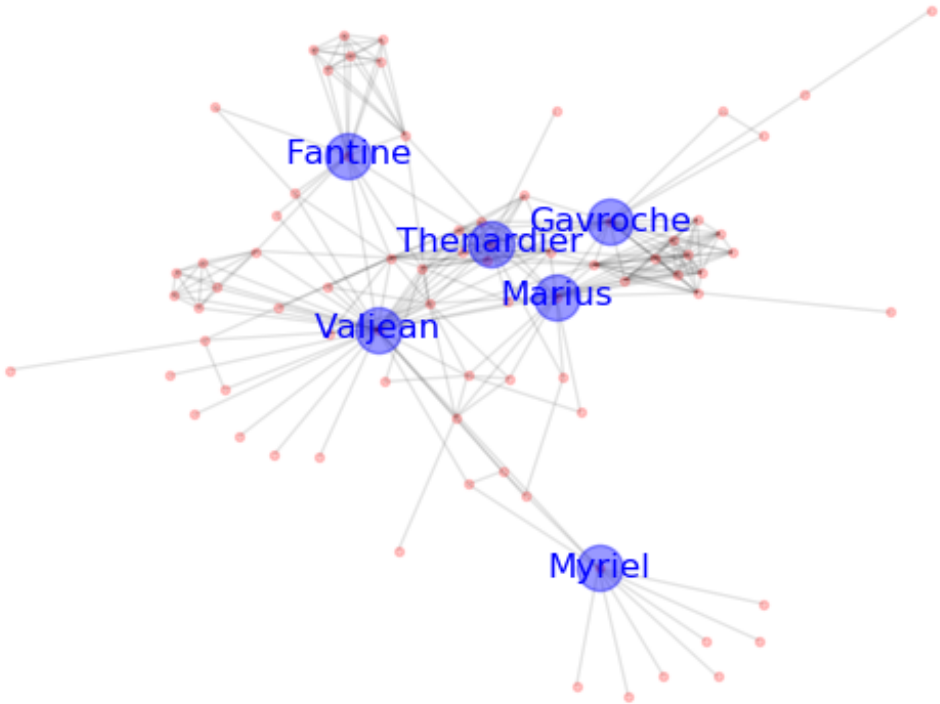
I nodi con alta Closeness hanno un migliore accesso alle informazioni o un'influenza più diretta su altri vertici

Betweewnness. *Il miglior intermediario, sta dentro a molti cammini minimi; Questa metrica puo indicare che l'attore funge da collegamento tra comunita che lavorano su ` argomenti diversi).*

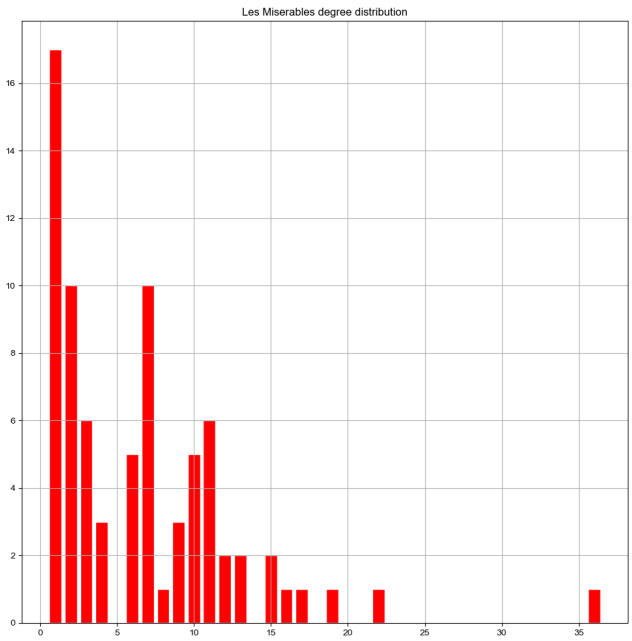
```
-- betweenness of first 20 nodes --
 [+] Valjean (betweenness level: 0.5699890527836184)
 [+] Myriel (betweenness level: 0.17684210526315788)
 [+] Gavroche (betweenness level: 0.16511250242584766)
 [+] Marius (betweenness level: 0.132032488621946)
 [+] Fantine (betweenness level: 0.12964454098819422)
 [+] Thenardier (betweenness level: 0.07490122123424225)
 [+] Javert (betweenness level: 0.05433155966478436)
 [+] MlleGillenormand (betweenness level: 0.047598927875243675)
 [+] Enjolras (betweenness level: 0.0425533568221771)
 [+] Tholomyes (betweenness level: 0.04062934817733579)
 [+] Bossuet (betweenness level: 0.03075365017995782)
 [+] MmeThenardier (betweenness level: 0.02900241873046176)
 [+] Mabeuf (betweenness level: 0.027661236424394314)
 [+] Fauchelevent (betweenness level: 0.026491228070175437)
 [+] MmeBurgon (betweenness level: 0.02631578947368421)
 [+] Cosette (betweenness level: 0.023796253454148188)
 [+] Gillenormand (betweenness level: 0.02021062158319776)
 [+] Eponine (betweenness level: 0.011487550654163002)
 [+] Simplice (betweenness level: 0.008640295033483888)
 [+] Bamatabois (betweenness level: 0.008040935672514621)
```

I nodi con questa metrica elevata possono avere un'influenza considerevole in una rete in virtù delle informazioni che controllano **Come vedremo in seguito questi nodi sono punti deboli della rete!!**

Nell'esaminare la visualizzazione della centralità Betweewnness, ho trovato sorprendente che nodi come Marius e Fantine abbiano una maggiore centralità rispetto a personaggi come Cosette. Questo è stata una piccola sorpresa perché Cosette (nella storia) ha il rapporto più stretto nella storia con Valjean (figlia adottiva) e proprio per questo mi sarei aspettato che il suo nodo avrebbe dovuto avere una maggiore Betweenness.
I personaggi con maggior Betweenness hanno un'influenza considerevolmente maggiore all'interno del "social network dei Miserabili" in virtù del loro controllo sulle informazioni che passano tra altri personaggi. Idem per la Closeness perché anch'essa indica anche che personaggi come Marius e Javert sono più centrali di Cosette.



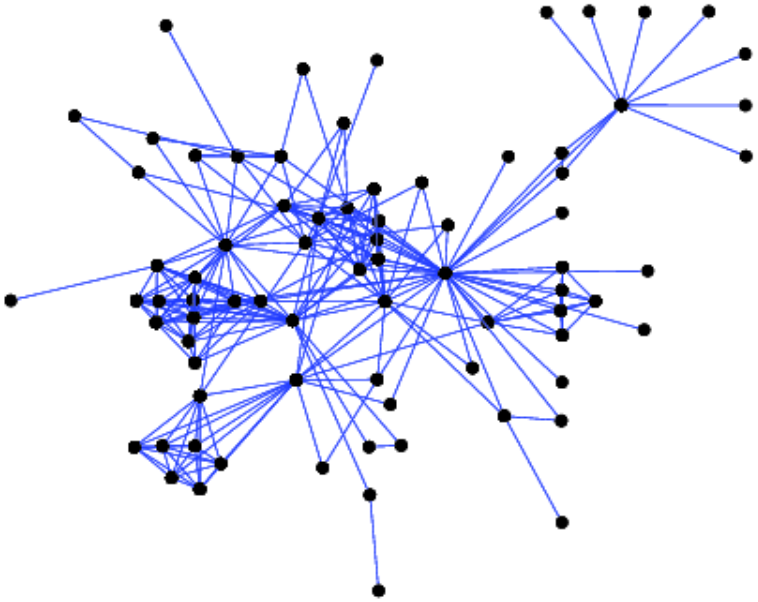
Plottando i 6 nodi con Betweewnness più elevata balza agli occhi, come un personaggio secondario(wikipedia :)) come Myriel abbia il secondo valore più elevato e sopratutto sia connesso a un gruppo di personaggi con collegato a nessuno dei protagonisti, però lo trovo coerente con la definizione della metrica stessa in quando Myriel appare in almeno 7 cammini minimi.



Distribuzione del grado: La distribuzione del grado) è il rapporto tra il numero dei nodi che hanno grado X e il numero totale dei nodi , in poche parole, **misura la probabilità che un nodo estratto a caso abbia esattamente X connessioni.** Dopo averlo già intuito dalla forma plot del grafo possiamo riscontrare un altra caratteristica dei grafi Scale-Free che la distribuzione ricalca una Power Law **cioè molti nodi con grado basso e pochi hub.**

Mi sono reso conto solo in momento di stesura, che anziché un diagramma a barre proprio per il numero esiguo di nodi sarebbe stato più idoneo un diagramma a punti

Quanto è connesso?



Nella teoria dei grafi, una componente connessa (o semplicemente una componente) di un grafo indiretto è un sottografo in cui: qualsiasi coppia di vertici è connessa da cammini. **Come si può vedere a occhio c'è un unica componente connessa (anche detta SCC)**

La Clustering Analisys è la ricerca di gruppi di oggetti tali che gli oggetti appartenenti a un gruppo siano “simili” tra loro e differenti dagli oggetti negli altri gruppi.

Coefficiente di Cluster è un numero compreso tra 0 e 1, **misura la probabilità che due vertici, connessi ad un vertice comune, siano anche connessi tra loro** ed è quindi legata al livello di addensamento medio dei vertici e quindi alla “robustezza” della rete. Si parla di “robustezza” perché misura quanto un grafo continua a restare connesso anche quando vengano rimossi alcuni dei vertici. Un alto valore del coefficiente C indica che sono presenti molte connessioni tra nodi vicini. Al limite, per una rete totalmente connessa C è pari a uno.

Average Coefficiente di Clustering: è un alternativa al coefficiente di clustering globale, il livello complessivo di clustering in una rete è misurato (da Watts e Strogatz) come media dei coefficienti di clustering locali di tutti i vertici, nel nostro caso risulta… **0.574**

Short Path Length: (o cammino minimo) tra due nodi X e Y è il percorso con il minor numero di collegamenti (per una rete con milioni di nodi calcolare il percorso più breve tra due nodi può essere piuttosto dispendioso in termini di tempo).
Essendo un grafo molto piccolo ho potuto calcolarlo agevolmente utilizzando il metodo in networkX.
Average Path Length: distanza media tra tutte le coppie del grafo e’ **2.4**

Diametro, ovvero la distanza massima tra i nodi del grafico e la distanza media è bassa rispetto alla rete, i nodi intermedi sono chiamati gradi di separazione, nel caso specifico del nostro Grafo il diametro è **5** (mentre come abbiamo appena visto)la distanza media è 2.4. Nel caso di Facebook nel 2011 (nodi 721.1M e bordi 68.7G) il diametro era 41 e il APL erano circa 4.
Molto importante sottolineare che in caso di rimozione nodi nei Grafi Scale Free, come il nostro, in caso di rimozione nodi il diametro aumenta.

Assortativita: -0.16 una metrica che indica quanto i nodi sono associati ad altri nodi simili o opposti, ha un assortativita molto simile a Internet (-0.18). Se una rete ha il coefficiente Assortativo negativo significa che gli Hubs tendono a connettersi con i non hubs.
Guardando wikipedia, questo e´ vero sopratutto per le network biologici in quanto sono reti statiche. Un altra cosa importante da dire ´che un forte valore aumenta la robustezza.
Per la precisione questa misura coefficiente negativo viene detta Disortativita´.

Robustezza

La robustezza della rete è una questione molto importante in molti contesti: nelle reti di comunicazione, i guasti al hardware possono disturbare la rete e impedire agli utenti di comunicare; nelle reti di distribuzione (come la distribuzione di energia o acqua), i guasti possono impedire il corretto svolgimento della vita umana; inoltre, le malattie possono diffondersi nelle reti di contatto e vaccinare le persone (quindi in un certo senso rimuoverle dalla rete di diffusione della malattia) può impedire che l'infezione raggiunga un numero elevato di persone.

Molti documenti hanno studiato la robustezza considerando la dimensione del componente connessa più grande (ovvero il più grande nucleo di nodi) come criterio per valutare la robustezza di una rete: maggiore è la dimensione di questo componente, maggiore è il numero di utenti che possono comunicare (o il numero di persone che una malattia può infettare), e quindi più robusta è la rete.

La maggior parte delle reti del mondo reale hanno distribuzioni Power Law (cioè hanno un gran numero di nodi con piccoli gradi, un piccolo numero di nodi con un grado molto alto, e tutti i casi intermedi).

I guasti sono considerati eventi casuali che posso essere simulati mediante l'eliminazione di un nodo casuale, gli attacchi riguardano la mirata eliminazione di nodi importanti.

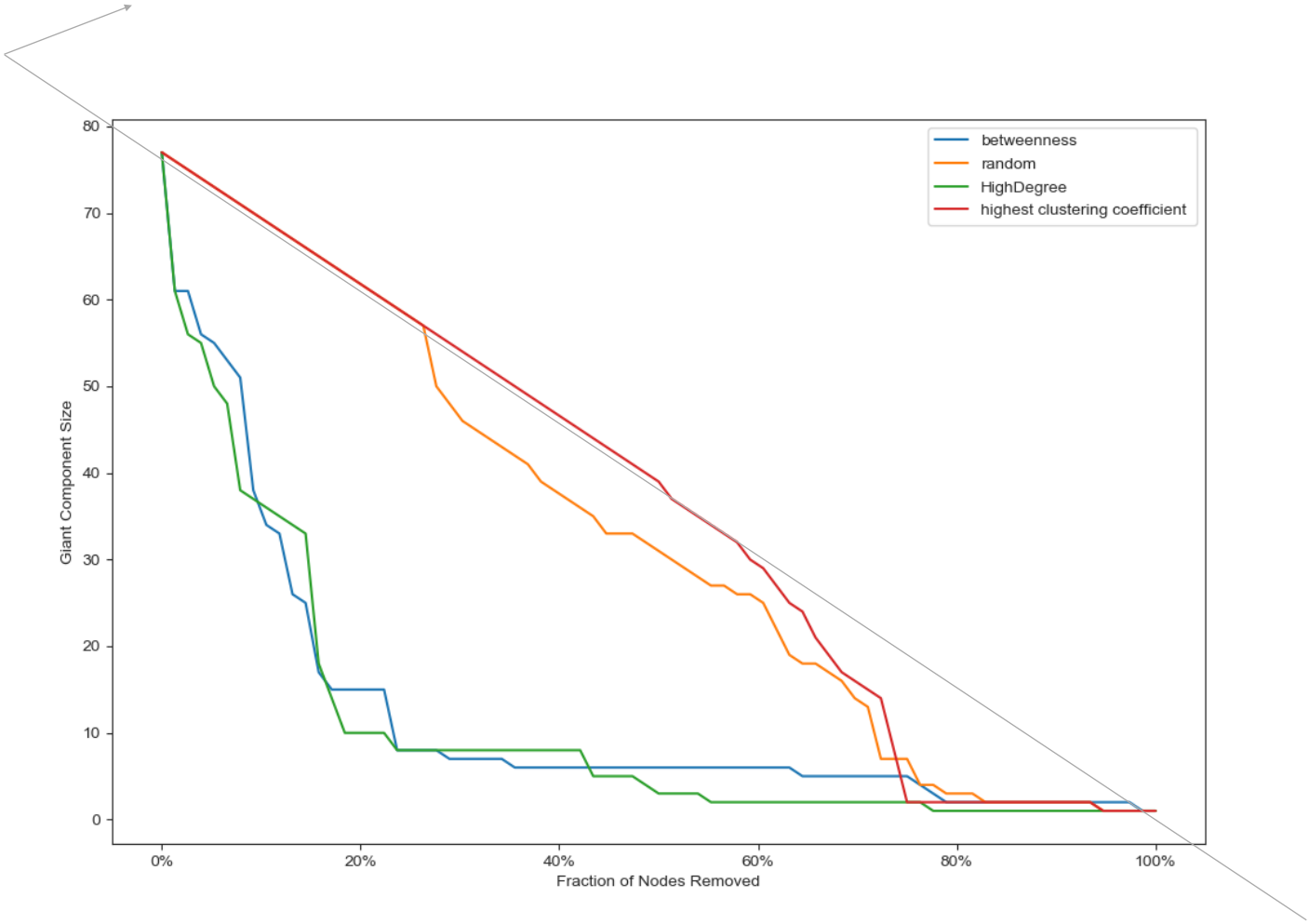
Ora abbiamo visto che cos'é la robustezza, è ora di vedere quanto è solido il nostro grafo.

Per questo attaccheremo i nodi delle reti con più approcci:

- Elimineremo i nodi in base alla **Betweeness,** passando da nodi con un punteggio elevato a quelli con punteggio basso
- Eliminemo i nodi in base al **Coefficiente di Clustering**, passando da nodi con un punteggio elevato a quelli con punteggio basso
- Eliminiamo nodi **in modo casuale**, simulando un guasto.
- Eliminiamo i nodi **in base al Grado** passando da nodi con un punteggio elevato a quelli con punteggio basso

Eliminando i nodi vedremo le conseguenze sulla componente gigante (il componente connesso più grande nel grafico), si ridurrà e alcuni nodi potrebbero avere un ruolo specifico in questo processo che causerebbe un drastico restringimento del componente gigante.

Siccome e' un grafo piccolo posso permettermi provare a eliminare tutti i 77 nodi, per ogni tipologia di attacco e come potete vedere nel grafico sotto li paragonerò al "Ipotetica line ideale di rimozione" **cioè il rapporto ideale tra il numero di nodi e come decresce la componente gigante, va da 77 a 0, un nodo alla volta.**



La rimozione dei nodi con Clustering elevato si comporta molto bene perche´ fino a circa il 55% sovrasta l'ipotetica linea ideale, in parole povere, **io rimuovo un nodo e la GCC decresce di uno.** Quindi il mio grafo e'molto resistente al clustering e probabilmente dopo il 55% un po per mancanza di nodi (perche´il 55% del totali di odi e' 35 nodi) e c'e'un picco dopo il 70%.

La simulazione di guasti, ogni volta che la plotto ovviamente e' diversa in questo specifico caso sono stato fortunato perche´ fino al 30% ricalca il clustering.

Le rimozione del Grado più elevato e la Betweenness sono entrambe un disastro, perche' come si può osservare in figura con un solo nodo eliminato (Valjean) si passa da 77 nodi a 60, ad ogni eliminazione perdo una quantita' significativa di nodi nella mia componente gigante.

In conclusione posso affermare che questi attacchi rispettano assolutamente ciò che mi sarei aspettato essendo uno Scale-Free cioe' in genere resistono bene ai guasti random (proprio perché con molti nodi e pochi hubs e´ più probabile che becco i primi).

Anche per le rimozioni mirate nessuna sorpresa, perché seguendo la power law, eliminando subito i pochi Hubs la maggior parte degli archi salta distruggendo i poco la componente gigante.

Sicuramente posso affermare che Grado e Betweennees sono 2 attacchi molto potenti per il grafo.

Perche´ e´ utile studiare il Contagio Sociale?

I social media hanno rivoluzionato il modo in cui le persone creano e utilizzano le informazioni. A differenza delle trasmissioni dei media tradizionali, che sono utilizzate "passivamente", i social network dipendono dagli utenti per propagare deliberatamente le informazioni che ricevono ai tramite utenti selezionati (follower, Friends ecc). Questo processo, chiamato contagio sociale, può amplificare la diffusione delle informazioni in un social network. Comprendere i meccanismi del contagio sociale è importantissimo per molte applicazioni: un esempio possono essere le campagne di marketing virale, valutare la qualità delle informazioni e prevedere fino a che punto si diffonderà.

Mentre la diffusione delle informazioni è spesso paragonata a una malattia contagiosa il contagio sociale differisce nel fatto che gli utenti dei social media cercano attivamente informazioni e decidono consapevolmente di propagarlo.

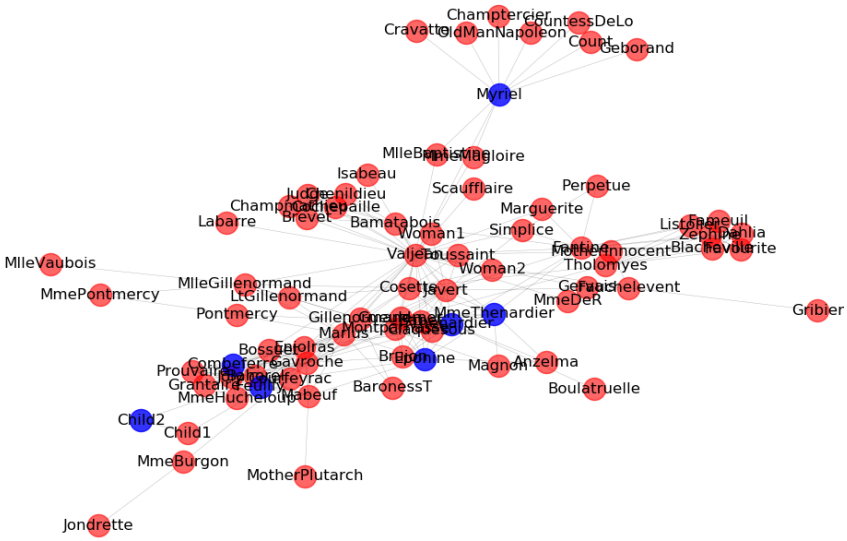
L'enorme flusso di contenuti dei social media disponibili spesso satura la capacità dell'utente di elaborare le informazioni. Nella maggior parte degli studi sulla propagazione delle informazioni sulle reti, gli utenti sono considerati esposti se hanno ricevuto un messaggio, indipendentemente dal fatto che lo vedano o meno, il che può portare a risultati controintuitivi che suggeriscono che esposizioni aggiuntive inibiscono la risposta. In realtà, l'utente che vede un messaggio dipende da come il sito organizza i contenuti, il flusso di informazioni in entrata e lo sforzo che l'utente è disposto a spendere nella scoperta delle informazioni.

il contagio sociale è piuttosto semplice e le risposte delle persone possono essere accuratamente previste.

La prima convenzione grafica e' che ho imposto il label Rosso per l'idea A (inizialmente già radicata) e il Blu per la nuova idea.

Inizialmente, i nodi nel grafico sono impostati sullo stato Rosso.

Dopodiché inizierà la **compartimentazione casualmente** cioe'verranno selezionati alcuni nodi centrali in cui cambiamo la loro label da rosso a blu.



Da qui parte il contagio…

Il contagio si evolve dinamicamente visitando i vicini dei nodi blu e indagando sui cambiamenti necessari nello stato.

Cosa si intende per cambiamenti di stato? Da A e B o nulla!

Il contagio si fermerà quando si troverà in una situazione di stallo (senza alcun cambiamento).

Entrando nel tecnico, ad ogni ciclo, il risultato viene tracciato con i colori appropriati, entrambi plottati sullo schermo (ho semplicemente colorato le label dei nodi) e per comodità salvati su una directory locale.

Invece di una matrice ho semplicemente deciso di due variabili, A e B scelte appropriatamente da me dopo varie prove.

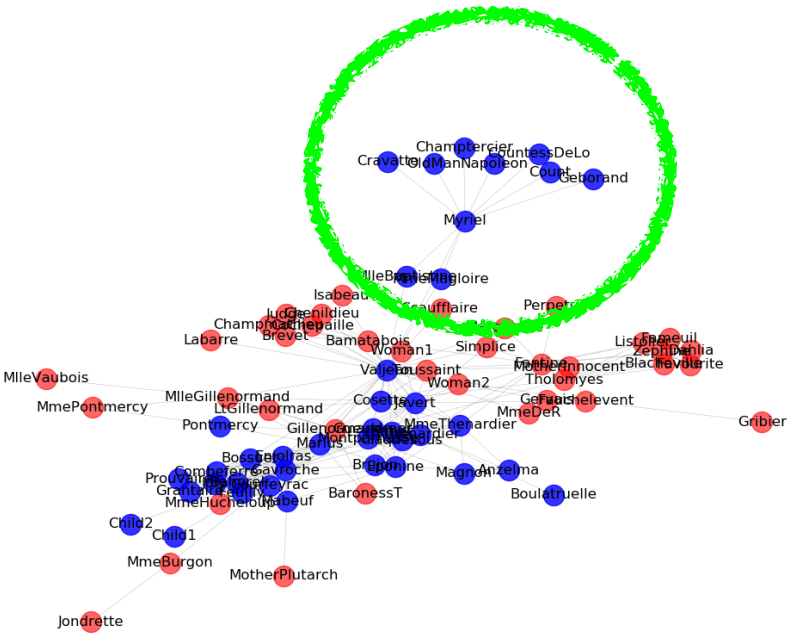
La prima prova e' stata A = 1 e B = 4 ma dipendeva molto dalla randomicita' con cui venivano scelti i primi nodi blu .(ho caricato una gif di questa prova) . Invece aumentando il payoff a **A = 2 e B = 7** sono sicuro che il 90% delle volte li contagio tutti, proprio perché lo stato B (anche se in minoranza iniziale) e' molto più forte nel 1vs1 perché ho fatto un incremento del 250%.

In questo esempio si possono vedere applicate 2 strategie per la diffusione di un idea:

- **Aumentare il payoff per far si che l'idea si più attraente**
- **Sfruttare celebrita' per influenzare più persone, come ad esempio in questo caso Valjean, nodo leader in quasi tutte le metriche e protagonista assoluto del racconto**

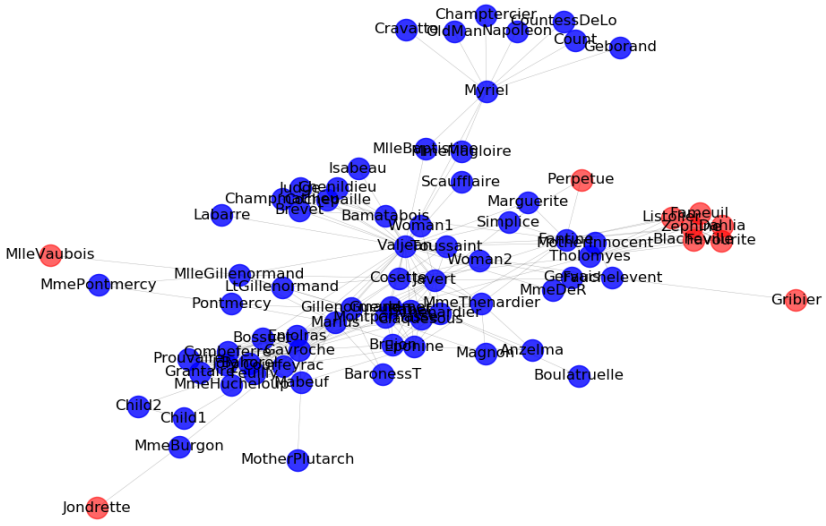
(Seconda ondata)

Perche Myriel oltre ad avere p più alta perché riesce a prendere tutti?



E' il nodo guardiano della comunità cioè il nodo che un numero di archi interni alla comunità e un numero di archi esterni, quelli interni sono in numero maggiore. Detto anche nodo Critico.

Se il nodo critico viene preso sicuramente la comunità viene contagiata, se la comunità 2 nodi vicini vengono presi contagio se il viene preso non e' sempre detto con myrial si ma se fossero connessi ciao.



(Terza ondata)

(quarta e ultima ondata)

CASCATA COMPLETA