[WMR 2015] Final Term

1. Community Discovery
2. Node Similarity
4. Local Diffusion

UNIVERSITÀ DI PISA

Lorenzo De Mattei
Andrea Meini
Vincenzo Rizza

La seconda parte del lavoro ha riguardato l'analisi più approfondita della rete ricavata da LAST FM, già analizzata per altri aspetti per il Mid Term Project. Le analisi hanno riguardato questa volta la community discovery, la node similarity e la local diffusion.

1. Community Discovery

Lo scopo dell'analisi di community discovery è quello di identificare all'interno della rete sociale sottoreti di utenti su base topologica.
Gli algoritmi che sono stati utilizzati a tale scopo sono k-cliques e DEMON.
Sono state effettuate analisi statistiche sulle comunità ottenute e successivamente sono state effettuate delle analisi per capire se gli utenti di ogni comunità condividono abitudini musicali simili.

Statistiche sulle comunità

Una volta generate le comunità con i due algoritmi sopracitati abbiamo calcolato alcuni indici statistici sulla distribuzione dei nodi nelle varie comunità.
Per entrambi i metodi di generazione delle comunità abbiamo una distribuzione secondo legge di potenza (come si può vedere in fig. 2 e fig. 4).
Abbiamo infatti molte comunità con pochi nodi, e poche comunità con molti nodi.
I box-plot (fig. 1 e fig. 3) mostrano che le varie community hanno dimensioni abbastanza uniformi ad eccezione delle poche community principali che risultano essere nettamente più grandi rispetto alle altre. Notiamo che il numero di comunità per DEMON è decisamente più alto che per K-cliques. Anche la densità delle comunità è maggiore in DEMON, segno che molti nodi appartengono a più comunità in DEMON da un lato, e che molti nodi vengono esclusi dalle comunità generate con K-cliques in quanto il numero minimo di nodi per ogni comunità è stato fissato a 5.

DEMON

Communities : 322

Average density of communities: 62.5745341615

Density standard deviation: 123.903698219

K-cliques

Number of communities: 56

Average density: 9.89285714286

Density standard deviation:14.5414169228

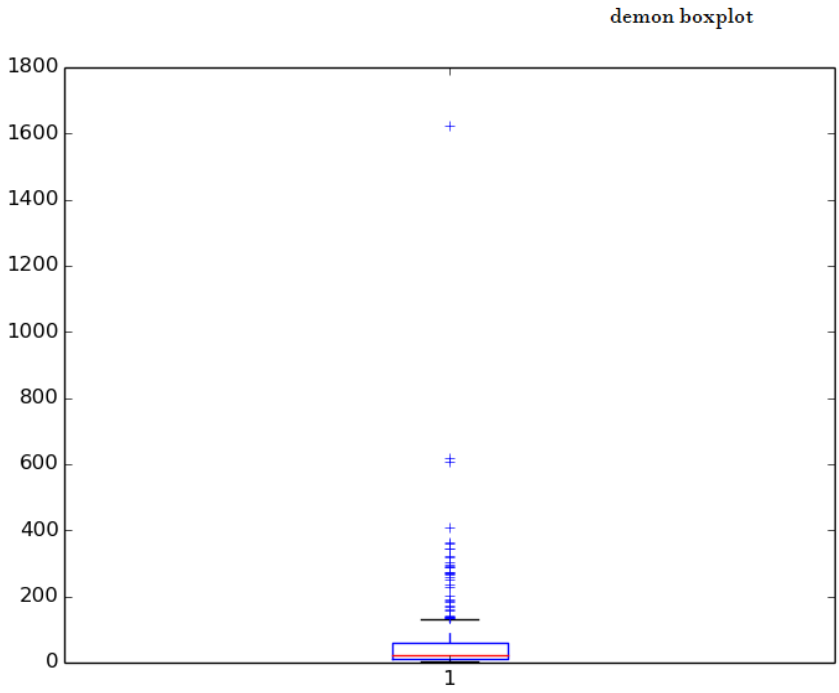

Fig 1

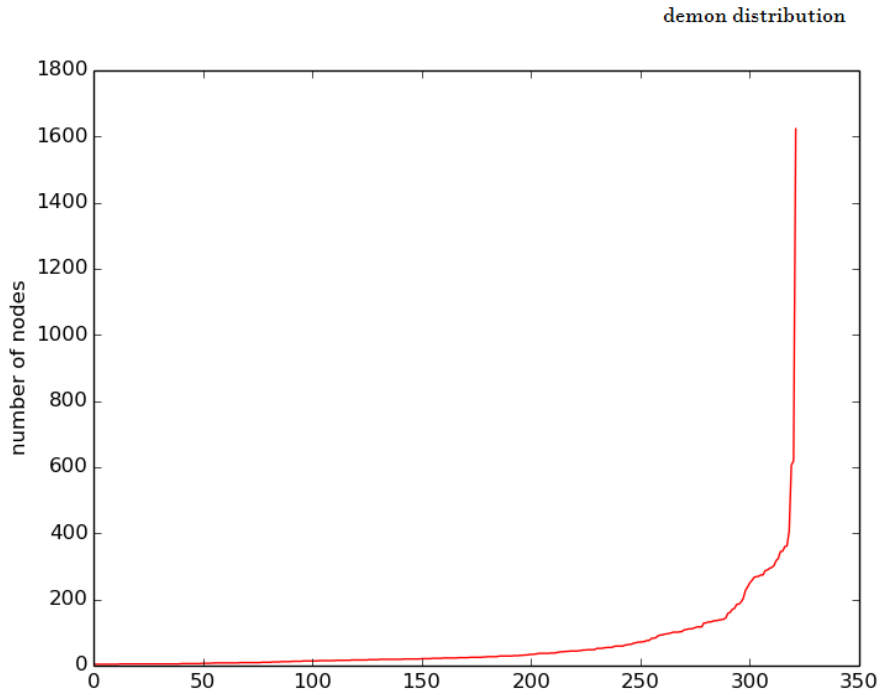
demon distribution
number of nodes
1800
1600
1400
1200
1000
800
600
400
200
0
0
50
100
150
200
250
300
350

Fig 2

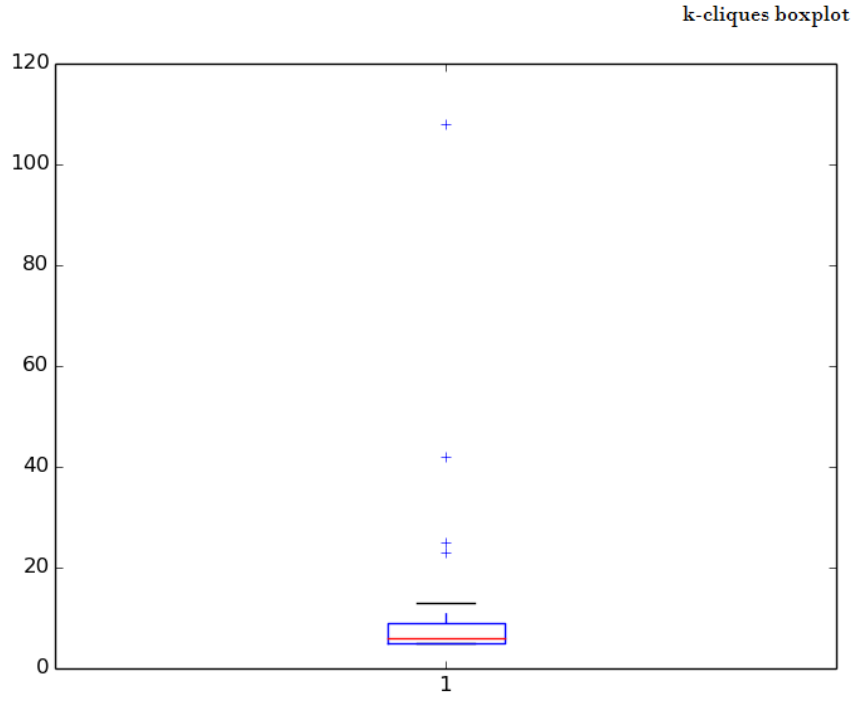
k-cliques boxplot
120
100
80
60
40
20
0
1

Fig 3

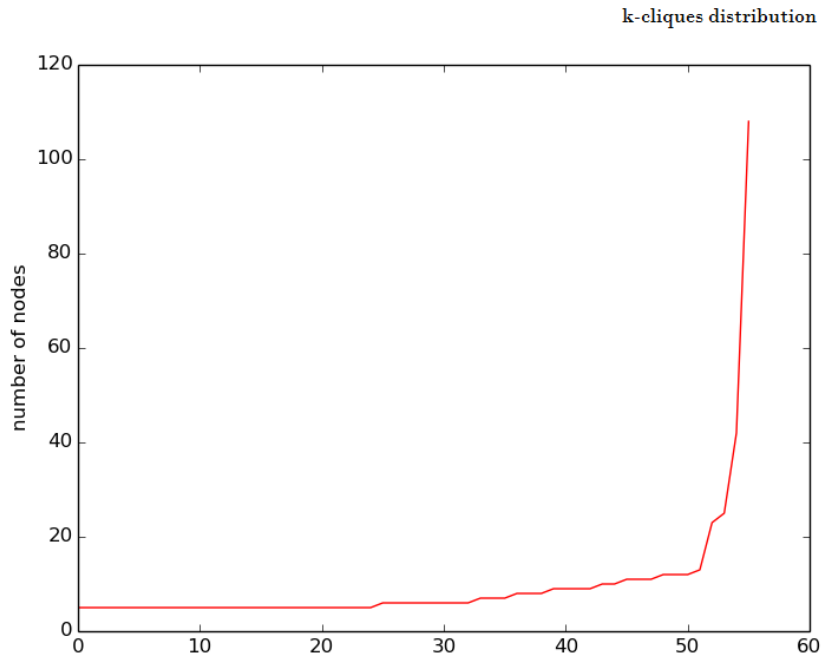

Fig 4

Comunità e abitudini musicali

Nella analisi successiva abbiamo voluto verificare che gli ascolti relativi alle diverse comunità fossero omogenei all'interno delle stesse.

Per fare ciò per ogni comunità e per il grafo totale sono stati calcolati due vettori:

- Vettore 1: ogni elemento corrisponde alla probabilità di ascolto di un artista. Calcolata come:

  $$Pr(a,\ c)\ = \frac{freq(a,\ c)}{numeroAscolti(c)}$$

  dove *freq(a, c)* è la frequenza con la quale un'artista viene ascoltato nella comunità *c* e *numeroAscolti(c)* è il numero totale di ascolti in *c*.
- Vettore 2: ogni elemento corrisponde alla probabilità di ascolto di un genere, calcolata come:

  $$Pr(g,\ c)\ = \frac{freq(g,\ c)}{numeroAscolti(c)}$$

  dove *freq(g, c)* è la frequenza con la quale un genere è ascoltato nella

comunità *c* (la frequenza è calcolata nello stesso modo in cui è stata calcolata nel mid term project) e *numeroAscolti(c)* è il numero totale di ascolti in *c*.

Sul dato totale abbiamo usato gli ascolti di tutta la rete. Gli utenti con meno di 100 ascolti non sono stati considerati. Le comunità con meno di 10 utenti con almeno 100 ascolti non sono state considerate. Per gli utenti con più di 100 ascolti sono stati considerati i primi 100 ascolti.
Abbiamo quindi calcolato la distanza euclidea tra il vettore di ogni comunità e il vettore della rete completa, al fine di osservare se le abitudini di ascolti divergessero da quelle medie. Riportiamo nei grafici in fig 5 e fig 6 la distribuzione delle distanze euclidee. Osserviamo l'andamento lineare delle due distribuzioni, fatta eccezzione per le fasi iniziale e finale in cui la crescita è più rapida. Possiamo osservare comunque che la maggior parte delle comunità diverge in maniera significativa dall'andamento generale. Riportiamo i dati solo per le comunità generate con DEMON in quanto quelle genrate con k-cliques analizzabili sono solamente 6, tuttavia l'andamento sembra essere il medesimo anche per queste comunità.

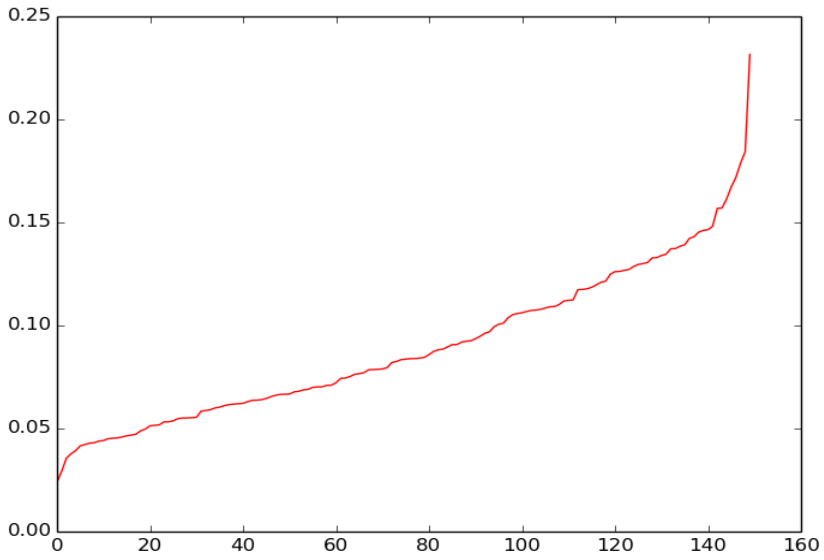

Fig. 5

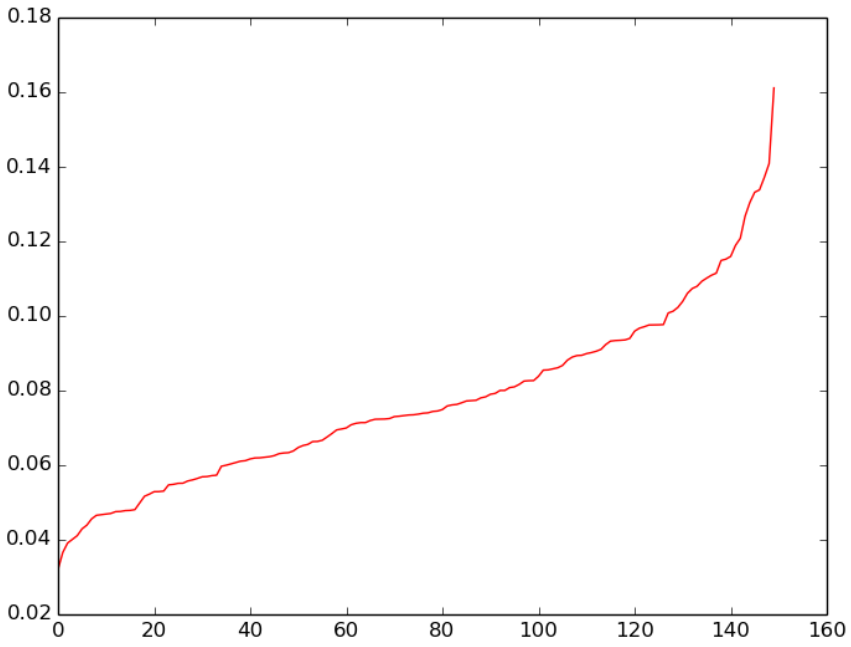

Fig. 6

Abbiamo cercato di evidenziare quali fossero i 5 generi che distinguono al meglio una comunità. Per ogni comunità abbiamo fatto la differenza tra il vettore della rete completa e il vettore della comunità in analisi. Abbiamo quindi estratto per ogni comunità i 5 generi corrispondenti ai valori massimi risultanti dalla differenza.
Troviamo gli output di questa analisi ai link:
https://raw.githubusercontent.com/LoreDema/Social_network_analysis_project/master/community/demon_top5_relevant_genre.txt
https://raw.githubusercontent.com/LoreDema/Social_network_analysis_project/master/community/k-clique_top5_relevant_genre.txt.

Possiamo osservare come per ogni comunità i 5 generi principali sono decisamente correlati tra loro nella maggior parte delle comunità.
Infine, abbiamo tentato di visualizzare quanto queste comunità siano coese internamente. Per farlo abbiamo costruito per ogni genere un vettore che lo

rappresenti, nello specifico ogni elemento di questi vettori contiene quanto un determinato artista appartiene al genere stesso, utilizzando la metrica sviluppata nel mid term project. In seguito abbiamo ridotto a 3 le dimensioni di questi vettori utilizzando la Principal components analysis.
In ultima analisi abbiamo attribuito un colore ad ogni comunità e abbiamo disegnato uno scatter plot tridimensionale, in cui è rappresentato un punto per ognuno dei 5 generi più rilevanti per ogni cluster. Il colore dei punti distingue i cluster di appartenenza. I risultati tuttavia non sono stati soddisfacenti come si può vedere dai grafici sottostanti. Probabilmente questo risultato è in parte dovuto al fatto che è pretenzioso pensare di rappresentare tali vettori in sole 3 dimensioni. Inoltre emerge analizzando i dati emerge come ci siano diverse comunità che condividono generi simili, pur essendo comunità distinte.

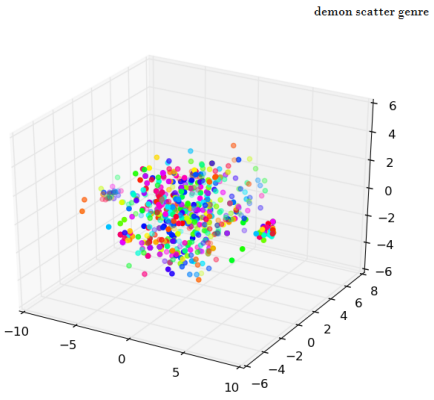

Fig 7

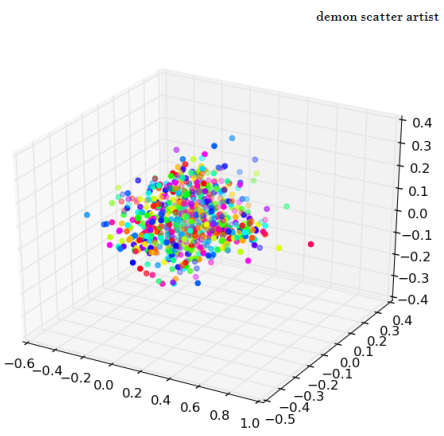

fig. 8

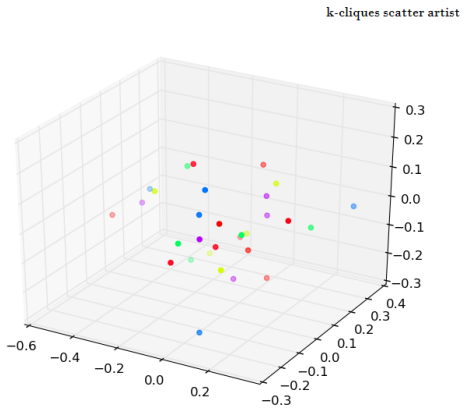

fig. 9

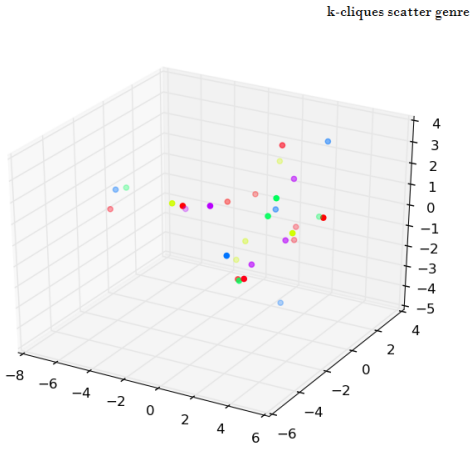

fig. 10

2. Node similarity (tie strength)

Lo scopo di questo task è stato quello di valutare quanto la similirità tra gli ascolti delle coppie di utenti connessi sia rilevante rispetto alla correlazione tra le stesse coppie di utenti calcolata su base topologica.
Per fare questo è stata determinata una misura di similarità tra le coppie di utenti connessi, calcolata sulla base degli ascolti. Innanzitutto è stato calcolato un vettore per ogni utente. Ogni elemento di questo vettore contiene la frequenza di ascolto per un particolare genere. Quindi per ogni coppia di utenti collegati abbiamo calcolato la distanza euclidea tra due vettori corrispondenti come misura di tie strength.

Una volta ottenuti questi risultati li abbiamo confrontati con altre misure di distanza tra i nodi di natura topologica, in particolare con il numero di vicini in comune (fig 12), il jaccard coefficent (fig. 11), l'adamic adar (fig. 14) e l'edge betweenness(fig. 13). Osservando i grafici risultanti emerge una scarsa correlazione tra l'andamento della similarità basata sugli ascolti e quella calcolata con indici di natura topologica. Osserviamo però che al crescere della tie strength aumenta il grado di oscillazione della vicinanza topologica tra i nodi.
Come abbiamo visto durante la fase di community discovery, ci sono comunità differenti che ascoltano gli stessi generi musicali. Questa oscillazione più elevata potrebbe essere dovuta appunto al fatto che ci sono nodi appartenenti a diverse comunità e quindi con alta distanza topologica ma con bassa distanza a livello di tie strength, allo stesso modo ci sono nodi appartenenti alla stessa comunità che risulteranno essere molto vicini sia a livello di tie strength sia a livello topologico.

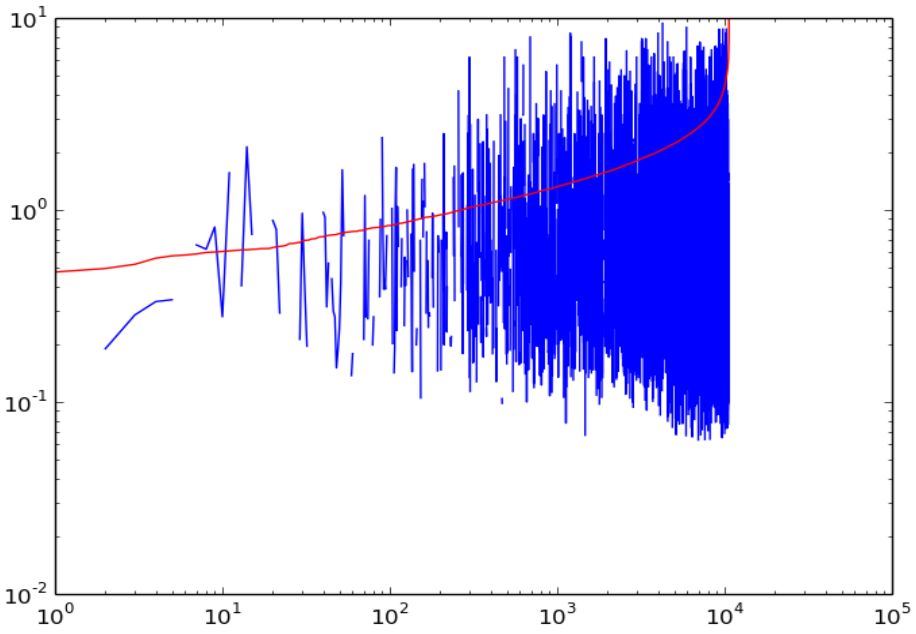

fig. 11 (Jaccard Coefficient)

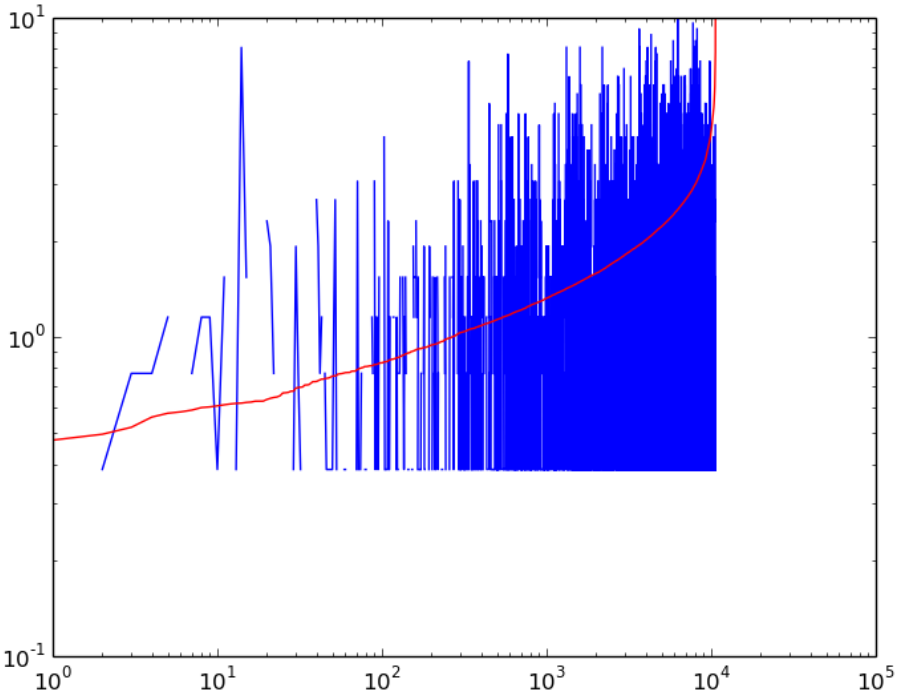

fig. 12 (Common neighbours)

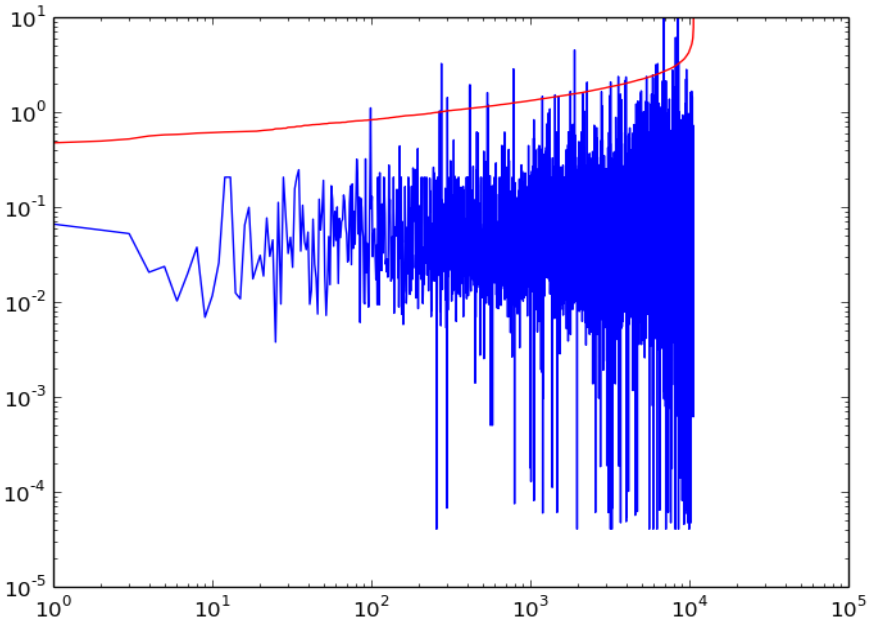

fig.13 (Edge Betweenness)

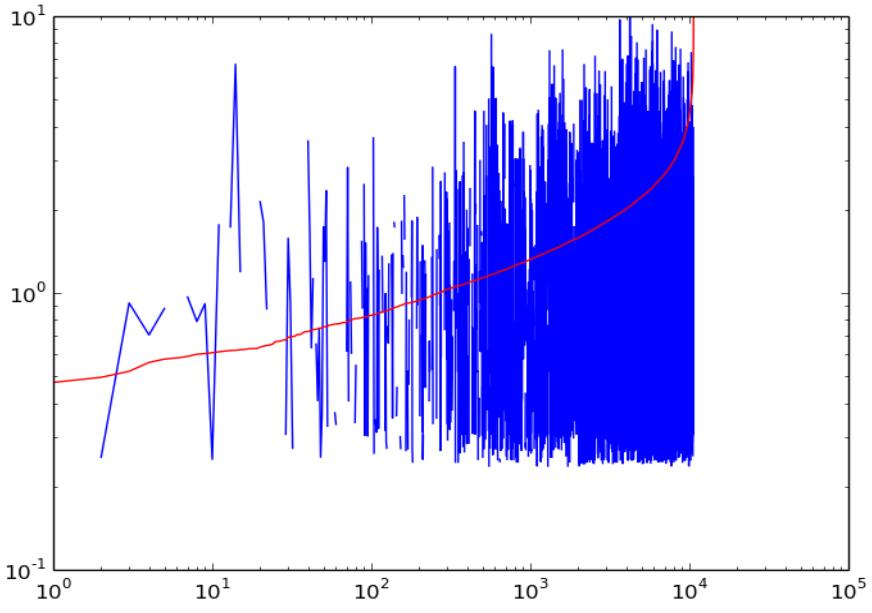

fig. 14 (Adamic Adar)

Quindi abbiamo analizzato come la struttura della rete fosse influenzata dalla rimozione degli archi a seconda del tie strength. La rimozione degli archi è stata eseguita in ordine di tie strength crescente (in rosso) e decrescente (in blu) (fig 15 fig 16 fig 17).
La crescita del numero di componenti (fig 16) e la decrescita della grandezza della componente principale (fig 15) hanno in entrambi i casi un andamento simile.

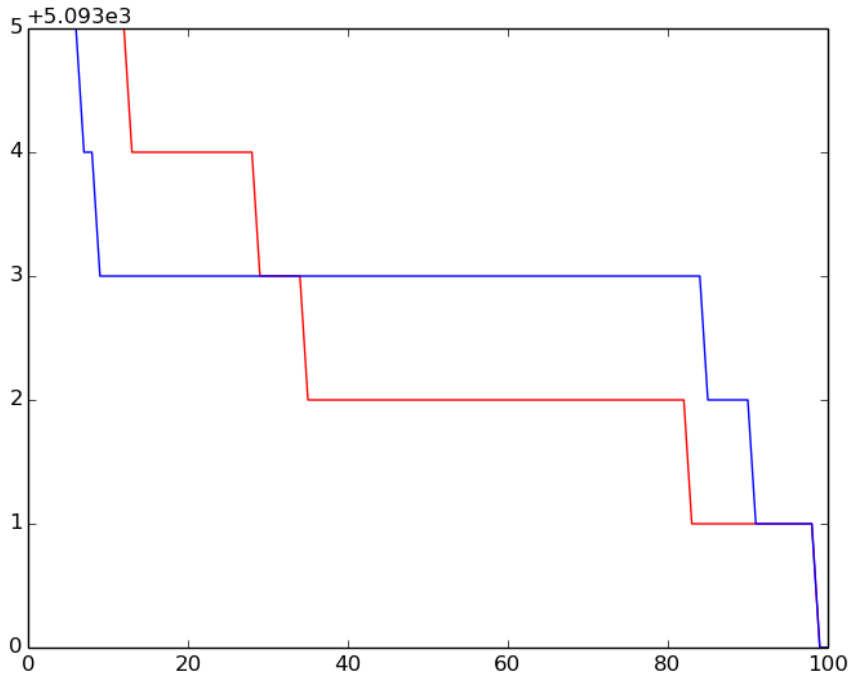

Fig 15

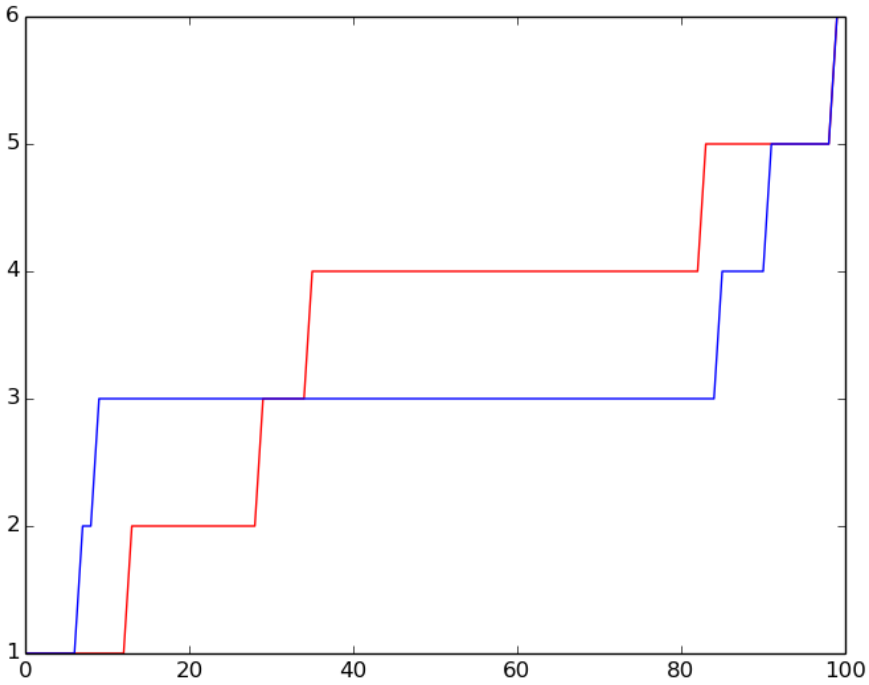

fig. 16

Al contrario la variazione del clustering coefficient (Fig 17) avviene in maniera molto diversa. La rimozione di archi con una tie strength alta comporta un calo graduale del clustering coefficient, in quanto più probabilmente questi archi appartengono a sottografi fortemente connessi. L'eliminazione di questi archi comporta perciò la distruzione di sottostrutture interne, diminuendo sensibilmente l'attitudine della rete a creare sottogruppi.
Viceversa eliminando archi con un basso tie strength è probabile che si vadano ad interessare connessioni tra nodi appartenenti a sottogruppi diversi. Perciò la rimozione di questi archi non causa sempre una diminuzione del clustering coefficient, ma spesso anzi un innalzamento.

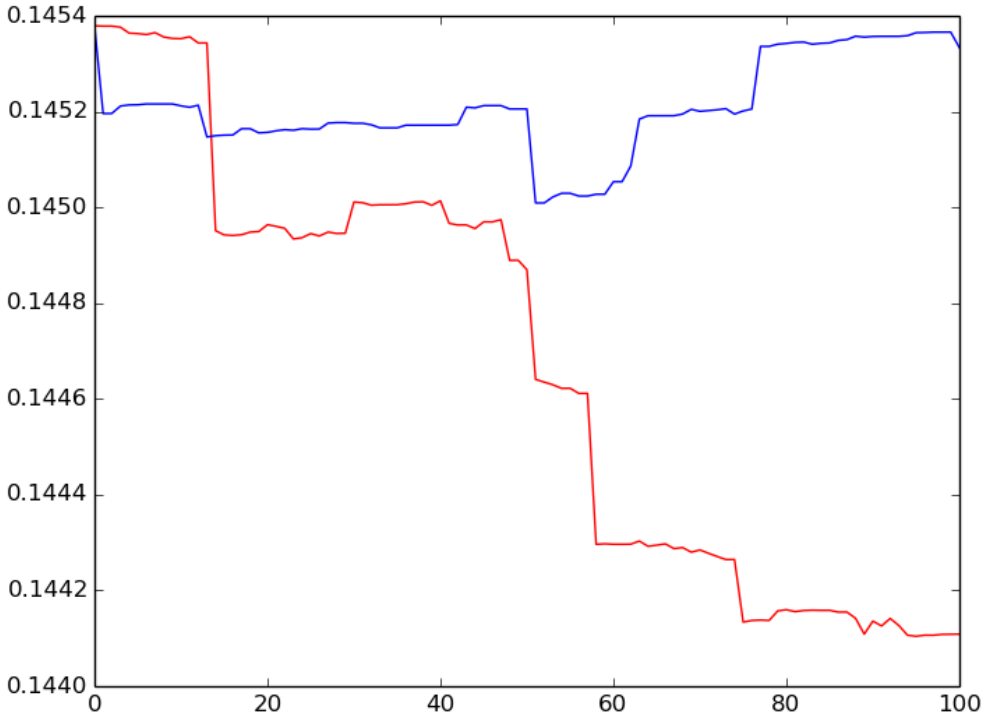
0.1454
0.1452
0.1450
0.1448
0.1446
0.1444
0.1442
0.1440
0
20
40
60
80
100

fig. 17

3. Local Diffusion

Nella struttura della rete è di notevole importanza lo studio dell'influenza che un utente, con le proprie abitudini musicali, esercita sugli altri utenti.
Per poter studiare questo fenomeno è stato necessario individuare un cosiddetto "leader" che fosse identificabile come primo diffusore di un particolare artista o genere musicale.

L'analisi è stata effettuata su cinque artisti con diversi gradi di hotness, suddividendo il totale degli artisti in cinque blocchi e per ciascun blocco è stato selezionato l'artista che possedeva il maggior numero di ascolti nella rete. Questo accorgimento è stato fondamentale per le analisi, in quanto il valore di hotness, da solo, non forniva un dato rappresentativo della presenza di un certo artista nella nostra rete. Gli artisti selezionati sono stati: Arctic Monkeys, Manowar, Megadeth, Metallica, Saviour Machine; non a caso tutti artisti con generi prevalenti simili ai generi più ascoltati dal nostro seed. È stato inoltre apportato un ulteriore filtro sulla hotness minima dell'elenco degli artisti, escludendo quelli aventi un valore inferiore a 0,4.

Per calcolare la local diffusion sono necessari due parametri: leader e intervallo temporale. I leaders rappresentano all'interno del nostro network una sorta di innovatori, in quanto primi ad aver ascoltato e quindi dato in pasto alla rete quel determinato artista e avviato il processo di "contagio sociale", per valutarne la diffusione conseguente dell'artista. L'intervallo temporale è invece un range entro il quale è possibile definire un utente influenzato da un altro.

Per la prima analisi è stato scelto come leader il primo ascoltatore nell'intera lista degli ascolti e sono state fatti vari esperimenti fissando il valore dell'intervallo temporale fino a una settimana. I risultati però non erano soddisfacenti, in quanto, dopo pochi ascolti era frequente riscontrare un gap temporale maggiore o uguale a un mese. Aumentare ulteriormente la soglia

avrebbe portato ad un risultato non significativo, non potendo trattarsi di local diffusion. Si è deciso dunque di cambiare approccio nella scelta dei leader.
All'interno dell'intera lista degli ascolti dell'artista è stata selezionata, fissando un certo intervallo, la sottorete che garantiva il maggior numero di ascolti. In questo modo è stato possibile ricercare, all'interno dei dati, delle sequenze più significative di utenti a partire dai quali poter individuare un nuovo leader. Grazie a questo accorgimento è stato possibile abbassare l'intervallo temporale tra due ascolti a solo 1 giorno, valore per cui è molto ragionevole pensare che, in una stessa rete, un ascolto sia influenzato da un altro.

Il risultato pressocchè nullo iniziale è relativo al fatto che in fase di crawling della rete, per ogni utente, venivano memorizzati solo ed esclusivamente gli ultimi 100 ascolti in ordine temporale. Per confermare questa dinamica abbiamo fatto un analisi sugli alberi ottenuti con questo nuovo approccio e abbiamo notato che gli ascolti corrispondenti a ciascun nodo erano tutti concentrati in un arco temporale recente.
Gli alberi, con eccezione dell'artista col grado minore di hotness (Saviour Machine), risultano molto più profondi dei precedenti.
Di seguito i risultati derivanti dalle analisi dei suddetti alberi, suddivisi per grado di hotness di ogni artista decrescente con i link ai rispettivi alberi:

	Width at first level	Depth	Size
Metallica	3	27	293
Arctic Monkeys	5	38	295
Megadeth	3	17	140
Manowar	2	12	57
Saviour Machine	1	1	2

Si nota che il numero di nodi dell'albero (*Size*) è abbastanza proporzionale alla hotness dell'artista, infatti si nota che i primi due artisti producono alberi di grandezza simile e i successivi artisti producono alberi di dimensioni minori.

Meno influenzata dal valore di hotness è invece la profondità degli alberi. Si nota infatti che il secondo artista produce un albero abbastanza più profondo del primo, mentre il terzo e il quarto hanno valori simili.

L'ampiezza del primo livello invece non è per niente correlata né alla hotness dell'artista né alle dimensioni dell'albero.

Una caratteristica che si nota è che gli alberi sono molto sviluppati in larghezza e poco bilanciati. Nella parte più alta gli ascolti sono meno frequenti, così come il numero dei figli per ciascun nodo sembra essere in media inferiore rispetto ai nodi presenti nella parte più bassa. Questa distribuzione comunque risulta giustificabile con l'andamento degli ascolti della nostra rete, con la scelta del leader e della sua rete discussi in precedenza.