Statistical Network Data Analysis
Laboratorio di Data Science in Economia CLEBA

Roberto Benedetti

Dipartimento di Economia, email benedett@unich.it

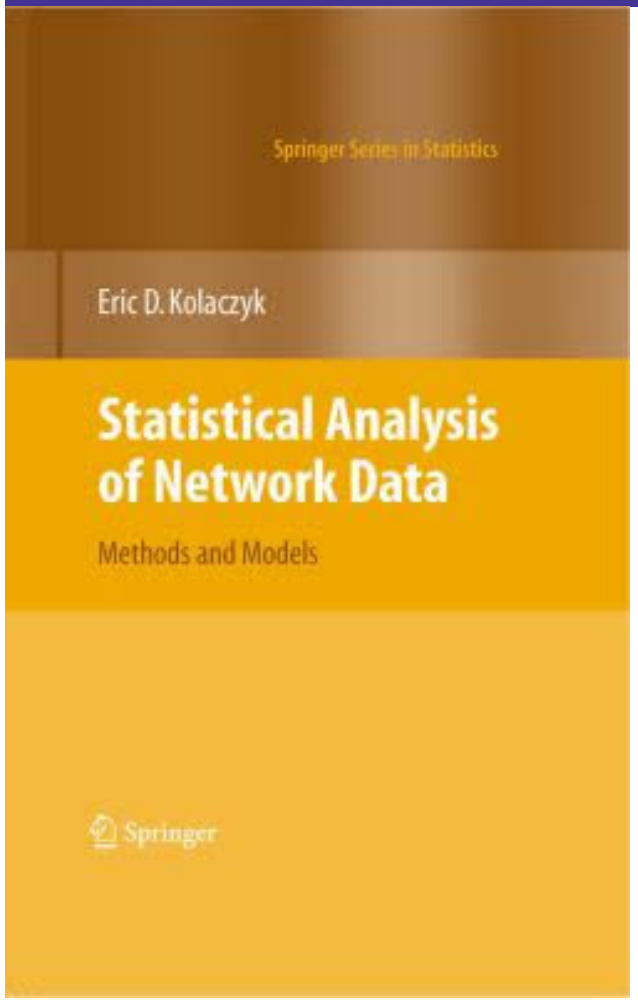
Il termine «Network» viene comunemente utilizzato per descrivere un insieme di elementi e le loro inter-relazioni.

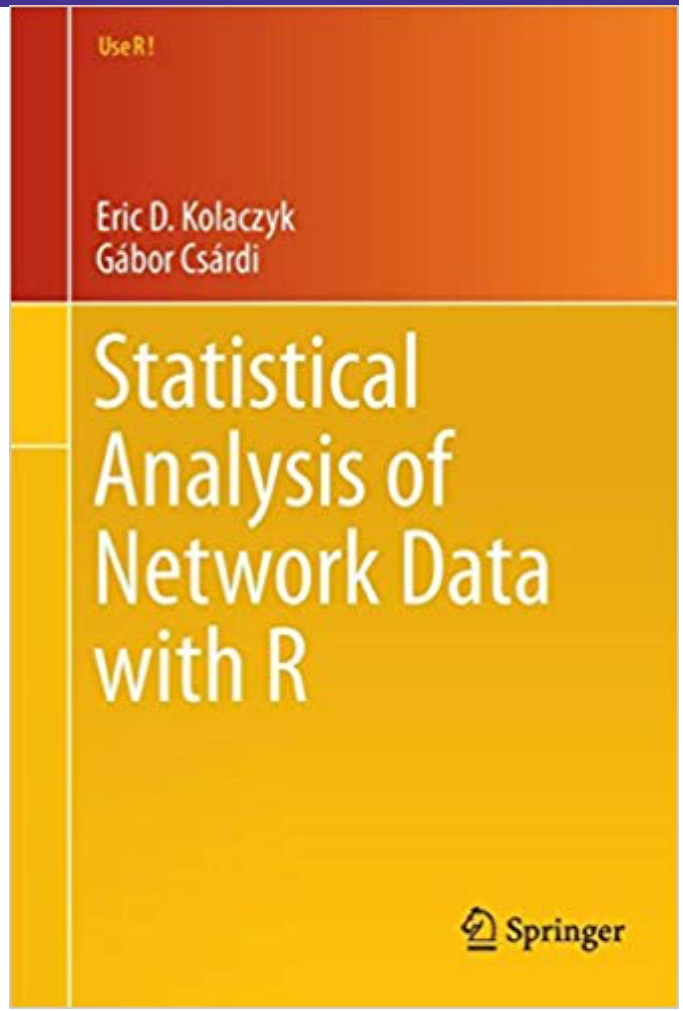
La *Statistical Network Data Analysis* nasce in relazione alla cosiddetta «Teoria dei Grafi», che fornisce definizioni, strumenti, tecniche e formalizza risultati riguardanti l'analisi dei grafi e delle loro proprietà.

Argomanti trattati

- Preliminari, concetti e strutture dei dati
- Mappatura dei Grafi e loro rappresentazione grafica
- Analisi descrittiva delle caratteristiche dei grafi di rete
- Campionamento e stima nei grafici di rete (per ora NO)
- Modelli per grafi di rete
- Inferenza sulla topologia
- Modellazione e previsione per i processi sui grafi di rete
- Analisi dei dati di flusso
- Modelli grafici
- Applicazioni

Riferimento principale





Come formalizzare le relazioni: il **Grafo**

- Per formalizzare graficamente una relazione fra due o più elementi c'è bisogno di una struttura chiamata *Grafo*:

Un grafo **G = (V, E)** è una struttura matematica consistente in un set di *V* vertici, comunemente detti *nodi (nodes),* ed un set di *E* spigoli (*edges*), comunemente detti *links,* dove ogni elemento appartenente ad *E* è rappresentato da una coppia $\{u,v\} \in V$, di vertici distinti.

Il numero di vertici $N_v = |V|$ ed il numero di spigoli $N_e = |E|$ costituiscono rispettivamente l'*ordine* e la *grandezza* del grafo.

Alcune definizioni…

- Un grafo H = (Vh, Eh) è un *sottografo* di G = (Vg, Eg), se Vh ⊆ Vg e Eh ⊆ Eg;
- Un grafo non ha spigoli (edges) per i quali entrambi i capi conducono ad un singolo vertice (si dice che il grafo non ha *loops*), e non ha coppie di vertici con più di uno spigolo fra loro (detti *multi-edges*);
- Un grafo G dove ogni spigolo è caratterizzato da un ordine dei suoi vertici tale che {*u,v*} ≠ {*v,u*} per u,v ∈ V, è denominato «Grafo diretto» (*Digrafo),* ed i suddetti spigoli sono chiamati «archi» (*o directed edges),* con la direzione dell'arco {*u,v*} letta da sinistra a destra, ovvero dalla coda *u* alla testa *v.*
- Un «Multi-digrafo» è un'estensione naturale di un digrafo, dove archi multipli condividono la stessa coda *u* e la stessa testa *v.*

Connettività di un Grafo

- La nozione principale di connettività di un grafo riguarda i concetti di «adiacenza»:
  - Due vertici u,v ∈ V sono adiacenti, se sono uniti dallo stesso spigolo e ∈ E.
  - Due spigoli e_1 , e_2 ∈ E sono adiacenti, se sono uniti dallo stesso punto di destinazione v ∈ V.
- E di «incidenza»:
  - Un vertice v ∈ V è detto incidente su uno spigolo e se v è un punto finale di e.

Da ciò deriva la nozione di *grado* di un vertice v, (dv), definita come il numero di spigoli incidenti su v.

La *sequenza* di gradi di un grafo G, è la sequenza formata dall'insieme dei gradi dei vertici dv, ordinati in maniera non decrescente.

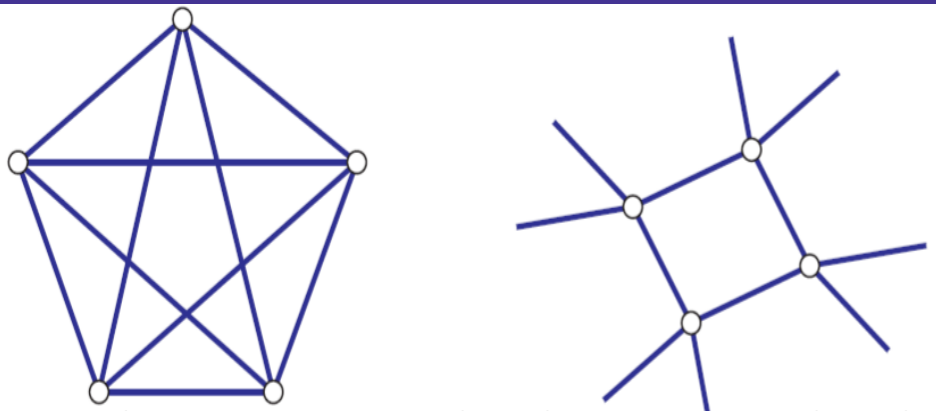
Movimenti all'interno di un Grafo

- All'interno di un grafo G, un *cammino* (walk) da v_0 a v_1, è rappresentato da una sequenza alternata $\{v_0, e_1, v_1, e_2, \ldots, v_{l-1}, e_l, v_l\}$, dove i punti di arrivo di e_i sono $\{v_{i-1}, v_i\}$. La lunghezza di tale cammino sarà uguale ad l.
- Una pista (*trail*) è un cammino senza spigoli ripetuti, mentre un sentiero (*path*) è una pista senza vertici ripetuti. Una pista nella quale il vertice di inizio e fine è il medesimo è detta circuito *(circuit).*
- Un cammino di lunghezza almeno 3, per il quale il vertice di inizio e fine è lo stesso, ma con tutti gli altri vertici distinti, è chiamato *ciclo (cycle).* Se un grafo non contiene cicli, è denominato *aciclico (acyclic).*
- Un vertice v in un grafo G è detto raggiungibile, da un altro vertice u, se esiste un cammino da u a v. Un grafo è quindi *connesso,* se ogni vertice è raggiungibile da tutti gli altri.

Movimenti all'interno di un Grafo

- La comune nozione di *distanza* tra vertici in un grafo, è definita come la lunghezza del più corto sentiero tra essi, se esso non esiste, la distanza sarà uguale ad ∞.
- Il valore della più lunga distanza in un grafo è detto *diametro* dello stesso.
- È comune assegnare ad ogni spigolo o vertice dei valori numerici: nel caso di spigoli e ∈ E, essi rappresentano i cosiddetti *pesi.* L'insieme *E* di tutti gli spigoli può essere infatti rappresentato attraverso un set di pesi $\{w_e\}$, e la corrispondente lunghezza di un percorso è misurata come la somma dei valori dei pesi degli spigoli attraversati.
- Tutti i concetti analizzati si estendono naturalmente ai digrafi.

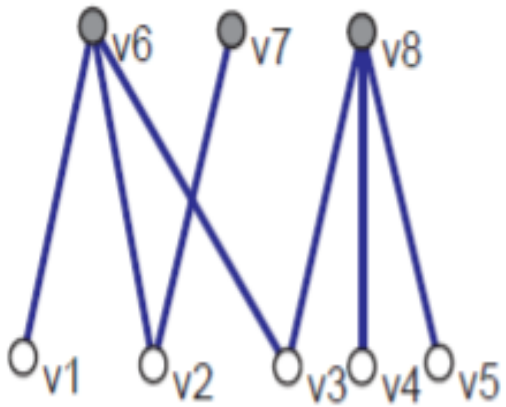
Tipologie di Grafo

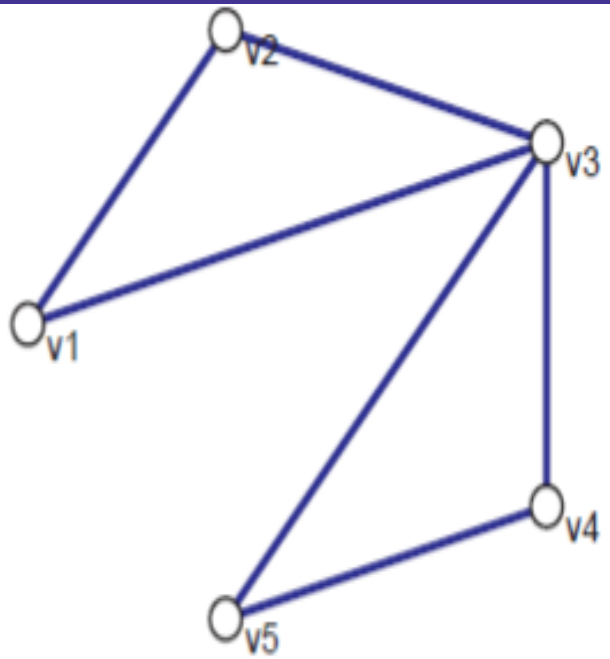


In base alle diverse relazioni tra gli elementi che lo compongono, un grafo può apparire di ogni forma e grandezza. I più comuni sono:

- **Grafo completo**: (a sinistra). Dove ogni vertice è collegato a tutti gli altri da uno spigolo. Un sottografo H di G, anch'esso completo, si definisce *clique.*
- **Grafo regolare**: (a destra). Dove ogni vertice ha lo stesso grado (numero di connessioni/relazioni); un grafo regolare con lo stesso grado (d) per ogni vertice, è detto *d-regolare.* In figura, un grafo 4-regolare.

Tipologie di Grafo





- **Grafo bipartito**: è un grafo G=(V,E), in cui il set di vertici V, può essere scomposto in due set disgiunti (V_1, V_2) ed ogni spigolo in E ha un punto finale in V_1, ed un altro in V_2. (figura a sinistra)
- **Grafo planare:** è un grafo che può essere disegnato su un piano, in modo che nessuno spigolo si incontri, tranne che sui vertici dove sono comunemente incidenti (figura a destra)

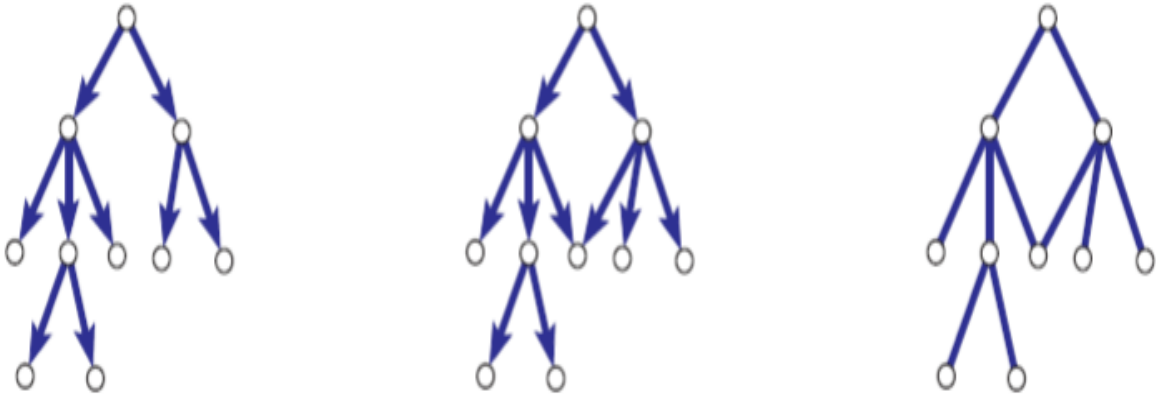
Una struttura comune in economia e statistica, l'«albero»

- **Albero**: è un grafo connesso, senza cicli al suo interno. L'insieme disgiunto di questo tipo di grafi è chiamato *foresta*.
- **Albero diretto**: è un grafo diretto, il cui grafo sottostante è un albero.

Nell' analisi dei network, gli alberi sono di fondamentale importanza, poiché essi possono sintetizzare complesse strutture di dati, e fornire algoritmi efficienti per l'analisi delle relazioni all'interno del network. (*Machine learning, decision tree…).*

L'albero è una struttura utilizzata anche in molti rami dell'economia, come l'economia comportamentale, la teoria dei giochi etc..

Ulteriori tipologie di alberi



Spesso gli alberi diretti si distinguono da tutti gli altri per la presenza, al loro interno, della cosiddetta *radice (root*), ovvero un vertice dal quale è possibile raggiungere tutti gli altri vertici nel grafo. Tale albero si dice *albero radicato* (figura a sinistra).

Un' altra importante struttura è il *grafo aciclico diretto (DAG),* che come il nome suggerisce, non ha cicli al suo interno (figura al centro) ma, a differenza di un albero diretto, il grafo sottostante non è un albero (struttura in figura a destra).

Teoria algebrica dei Grafi

Unisce l'analisi dei network (e quindi dei grafi) all'algebra matriciale.

Matrice di adiacenza: è una matrice simmetrico-binaria N_v x N_v, contenente al suo interno:

$$\mathbf{A}_{i,j} = \begin{cases} 1, & se\ \{i,j\} \in E, \\ 0, & altrimenti, \end{cases}$$

dove ogni spigolo e ∈ E è rappresentato con una coppia non ordinata di vertici i, j ∈ V. **A** è quindi una matrice composta da 1 per gli elementi del grafo che hanno vertici collegati da uno spigolo, e 0 per quelli che non possiedono questa proprietà.

Matrice di incidenza: è una matrice binaria N_v x N_e composta da:

$$\boldsymbol{B}_{i,j} = \begin{cases} 1, & se\ il\ vertice\ i\ incide\ sullo\ spigolo\ j \\ 0, & altrimenti, \end{cases}$$

«Mappare» i Network

Con questo termine si intende un processo finalizzato alla rappresentazione grafica, della struttura di dati (e quindi del network) oggetto di analisi, e delle loro relazioni.

La visualizzazione grafica di un network non è il risultato di un processo banale, per due ragioni fondamentali:

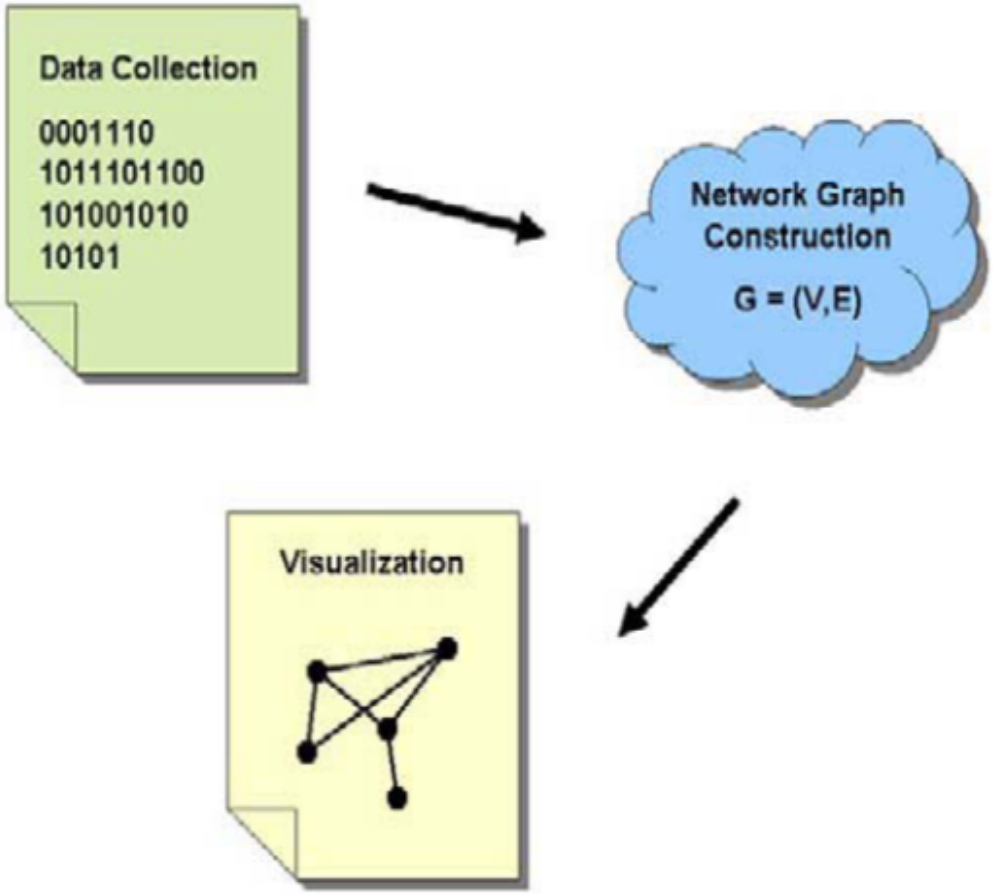
- In molti contesti è necessario più di un singolo grafo per rappresentare un singolo sistema;
- Anche dopo la rappresentazione grafica, l'effettiva sfida è riuscire a comunicare la totalità di informazioni fornite dal network, su un piano bi (o tri)-dimensionale.

Il grafo del network si sostanzia nell'accoppiamento di due set: uno di vertici, ed uno di spigoli, e da ciò deriva un'enorme flessibilità nel mappare diversi tipi di informazioni, qualitative e quantitative, fornite (in taluni casi) da dataset altamente multidimensionali.

Step fondamentali per la produzione di Network

1. Raccolta dati riguardanti un sistema oggetto d'analisi;
2. Creazione di una rappresentazione di essi mediante uno o più grafi (*network graph representation);*
3. Traduzione della network graph representation in una raffigurazione visiva (*visual image*).

Key Steps


Data Collection
0001110
1011101100
101001010
10101
Network Graph Construction
G = (V,E)
Visualization

Raccolta dati

Non esisterebbero network senza dati.

Ovviamente, tipologie di dati, nonché strumenti e tecniche di rilevazione variano da disciplina a disciplina, ma c'è una peculiarità che accomuna aree di indagine e metodi differenti, e che rende l'analisi dei network realizzabile:

I dati devono essere **relazionali**

Cosa significa?

Che oltre alla misurazione e analisi dei singoli elementi che compongono il network, bisogna individuare legami, relazioni, rapporti ed interazioni fra essi.

Individuazione di elementi e interazioni

Per produrre una visualizzazione del network, come visto, è necessaria una costruzione e rappresentazione del grafo. Ciò che rende complessa questa fase è il fatto che i dati a disposizione, spesso riguardano misurazioni elementari di un set di unità, e le «interazioni» fra esse non sono esplicitamente specificate.

Da ciò deriva la *soggettività* nella scelta degli *elementi* e delle *interazioni* del nostro network.

La scelta è fondamentale poiché essa influenza inevitabilmente non solo la costruzione grafica, ma soprattutto l'interpretazione del network e le informazioni e conclusioni che esso fornirà.

Categorie di dati

La prospettiva della raccolta di qualsiasi tipo di dato sia di nostro interesse, è senza dubbio ottimistica. Per cui distinguiamo tre fondamentali tipologie di dati da cui possiamo attingere per creare ed analizzare il nostro network:

1. **Enumerated Data**: Sono dati rilevati in maniera esaustiva sulla totalità della popolazione, intesa con lo stesso concetto statistico di «popolazione finita» (ad es. la quantità di importazioni ed esportazioni dei paesi UE)
2. **Partial Data**: Sono dati derivanti esclusivamente da un subset della popolazione di riferimento. Dal contesto e dalla prospettiva di conduzione dell'analisi, può accadere che alcuni dataset che apparentemente sembrano «enumerated», siano di contro «partial».
3. **Sampled Data**: O dati campionari, sono dati rilevati su unità selezionate dalla popolazione mediante metodi di campionamento. Sebbene, come nel caso dei partial data, siano rilevati solo su un subset della popolazione, essi non forniscono a priori una visione esaustiva della sotto-popolazione che sarà oggetto d'indagine.

Categorie di dati

Tutte le tre categorie citate hanno un loro grado di convenienza, ma la realtà della rilevazione potrebbe presentarsi in maniera più sfumata, combinando aspetti dei tre tipi di dati analizzati.

Per cui, si possono alternativamente utilizzare i concetti di:

- Observed data (dati osservati)
- Missing data (dati mancanti)

Rappresentazione grafica del Network

Si esplica nella costruzione di un grafo G = (V,E) formato da un set di vertici (V) e di spigoli (E).

Tuttavia possono anche essere inseriti all'interno del grafo sia set di pesi per gli spigoli $\{w_e\}$ e ∈ E, che forniscono indicazioni sulla variabilità delle connessioni tra vertici, che set di variabili $\{x_v\}$ v ∈ V riguardanti attributi (quantitativi o qualitativi) dei vertici stessi.

Step 1: creazione di *v* ed *e*

Con un set di dati del network a nostra disposizione, i primi step verso la costruzione di G sono:

1. Assegnare vertici *v*, agli elementi oggetto di misurazione (le unità d'analisi);
2. Assegnare spigoli $e \in E$ alla misurazione delle interazioni fra le unità.

Problemi nella trattazione di *v*

Nel processo di rappresentazione grafica di una qualsiasi mappa su un qualsiasi schermo, sia lo spazio che la risoluzione disponibili sono finiti.

Per grafi con svariate centinaia di vertici, la rappresentazione può divenire scomoda, a causa dell'enorme numero di unità da raffigurare.

Un modo per ovviare a questo problema è quello di rappresentare esclusivamente i sottografi maggiormente rilevanti. Se ciò non è possibile, poiché è necessario rappresentare l'intero grafo ai fini dell'analisi, c'è bisogno di una riduzione del numero di vertici, ad esempio modificando la scala di analisi. (es. non analizzo le inter-relazioni fra i comuni, ma fra le province)

Problemi nella trattazione di *e*

Il numero di spigoli Ne in un Grafo, spesso si relaziona in maniera lineare al numero di vertici Nv, da ciò deriva la connessione delle due problematiche.

In particolare, nel caso di relazioni più grandi (ad esempio quadratiche fra vertici e spigoli: $N_e \leq \binom{N_v}{2}$) si può facilmente incorrere nel problema di avere un numero spropositato di spigoli nonostante quello di vertici sia contenuto. Ciò si traduce in una eccessiva densità di relazioni, poco funzionali all'analisi interpretativa.

Si può ovviare a questo problema stabilendo una soglia sotto la quale la relazione non è significativa per la nostra analisi: eliminando un numero sufficiente di collegamenti, la densità del grafo diminuirà.

Visualizzazione del Network

È ovvio che ci sono infiniti modi differenti per utilizzare un insieme di punti e linee al fine di rappresentare un grafo G; ciò che è importante è che la scelta effettuata, sia in grado di comunicare le informazioni relazionali desiderate.

Ad esempio, un approccio frequente è quello di posizionare i vertici del grafo in maniera circolare ed equidistante, e di inserire linee continue tra vertici connessi. Sebbene di semplice realizzazione, questo metodo tende ad accentuare la densità delle relazioni verso il centro.

Per ridurre le problematiche della visualizzazione, sono state formalizzate delle linee guida che possono essere riassunte in:

Linee guida per la rappresentazione

- ***Convenzioni grafiche***: sono linee guida riguardanti il tipo di linea utilizzato per le relazioni (ad es. linee rette piuttosto che linee multiple o curve, o la non intersezione delle stesse) od il posizionamento di vertici e spigoli (ad es. lungo una griglia immaginaria). Per gli *alberi* o per grafi aciclici, spesso si usa un approccio ascendente (upward) o discendente (downward), rappresentando gli spigoli in maniera crescente o decrescente rispetto l'asse verticale.
- ***Estetica:*** sono meno stringenti delle convenzioni grafiche, e riguardano l'estetica del grafo (cercare di minimizzare l'area usata per la rappresentazione, il numero o la lunghezza delle linee)
- ***Restrizioni:*** sono requisiti riguardanti più i sottografi $H \subset G$, che i grafi stessi. Possono riguardare il posizionamento di un vertice nel disegno, o di un cluster di essi, la direzione di un percorso, ecc.

Macro-Categorie di Grafi

Sebbene possano essere utilizzati molteplici metodi per tradurre visivamente il grafo, la maggior parte di essi può essere racchiusa nella creazione di due macro categorie di Grafi:

- **Grafi con struttura particolare**

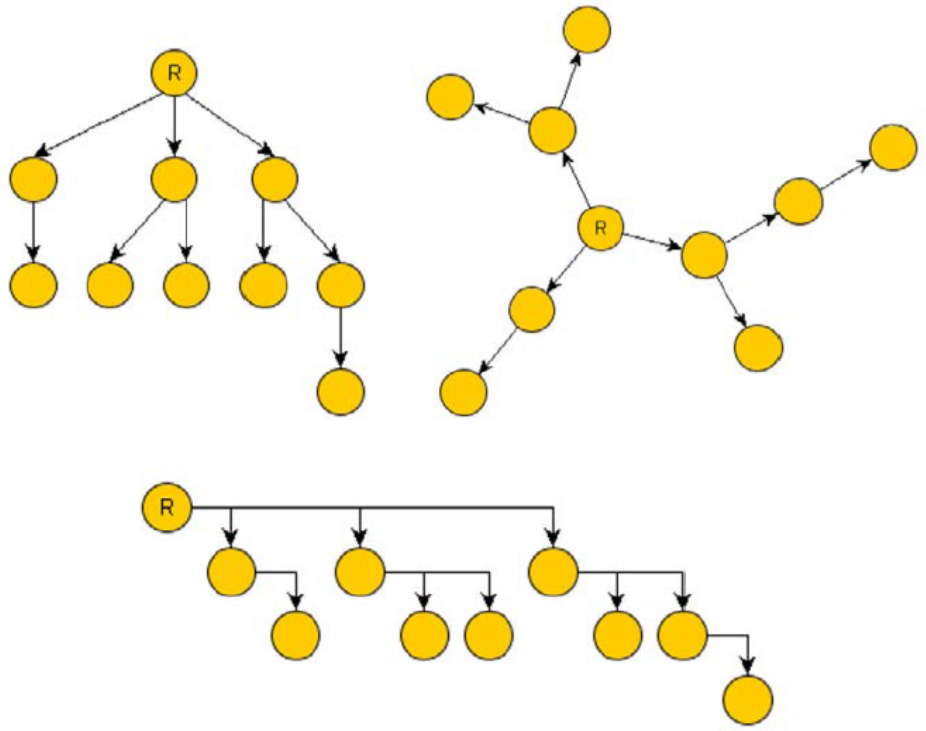
Vi sono due strutture principali nell'analisi dei Grafi: oltre ai già citati alberi, una struttura fondamentale è quella del grafo planare.

Il «Grafo planare», è una particolare struttura di grafo associata ai network con un forte legame con la geografia terrestre, come ad esempio quelli riguardanti l'energia o i trasporti.

Due tecniche comuni per la sua realizzazione sono l'utilizzo di percorsi ortogonali per creare gli spigoli, e l'uso di poligoni regolari di k lati, per rappresentare un percorso di lunghezza k. In altre parole, blocchi e poligoni diventano le strutture geometriche di fondo del network.

Macro-Categorie di Grafi

Gli «alberi», sono utilizzati nella rappresentazione di strutture gerarchiche, come ad esempio organigrammi o alberi genealogici, poiché l'obiettivo principale dell'analisi di questo tipo di network è quello di comunicare l'informazione riguardante le relazioni all'interno della gerarchia.



Grafi con analogie tra la struttura e le relazioni

Se il grafo G non ha una struttura particolare, un metodo per rappresentarlo è quello di basarsi su analogie tra la struttura relazionale del grafo e le forze in gioco fra gli elementi.

In questo ambito, due approcci comuni sono:

- Spring-embedder method: è un approccio che introduce forze attrattive e repulsive rappresentando vertici come palline, e spigoli come molle, definendo la nozione di forza che dipende dalla posizione e dalla distanza di coppie di vertici.
- Energy-placement method: identificando col termine «energia», una funzione delle posizioni dei vertici, con l'obiettivo di ottimizzare quest'ultime, per minimizzare l'energia del sistema.

Analisi strutturale descrittiva delle caratteristiche del Network

Una volta costruita la rappresentazione grafica, si passa all'esplorazione delle caratteristiche strutturali e delle proprietà del network. Tale analisi, è stata trattata principalmente con un approccio descrittivo, piuttosto che inferenziale, con strumenti e tecniche fornite non solo dalla statistica, ma anche dalla matematica, fisica e computer science.

Le macro aree che compongono questa analisi descrittiva sono:

1. Caratterizzazione di vertici e spigoli;
2. Caratterizzazione della connessione del network.

Caratteristiche di vertici e spigoli

- **Grado di un vertice**: un grado dv di un vettore v, in un network G = (V,E), è rappresentato dal numero di spigoli in E, incidenti su v. Questo concetto ci permette di fornire una misura della connessione di v agli altri vertici del grafo.

Se i gradi dei singoli vertici vengono considerati in maniera aggregata, attraverso la sequenza $\{d_1, \ldots, d_{N_v}\}$, possono essere proposte misure della connettività totale nel grafo riguardanti i gradi, come:

a. Distribuzione: In un grafo G, data la frazione f_d di vertici v con grado d_v = d, l'insieme $\{f_d\}$ d ≥ 0 è chiamato distribuzione dei gradi di G, ovvero l'istogramma formato dalla sequenza dei gradi dei vertici.

b. Correlazione: fornisce informazioni su come i vertici sono connessi fra loro, poichè due grafi possono avere identica sequenza di gradi, ma differente connessione fra vertici.

Centralità

È il concetto che definisce «l'importanza» di uno o più elementi all'interno di un network, obiettivo spesso primario dell'analisi.

Il grado di un vertice è un esempio di misura di centralità dello stesso, ma sono state proposte negli anni altre 3 misure per questo tipo di relazione:

- **Closeness**: vicinanza di un vertice ad altri;
- **Betweenness**: considera l'importanza di un vertice in relazione alla sua posizione rispetto ai sentieri/percorsi nel grafo, e quindi ad altre coppie di vertici;
- **Eigenvector centrality**: si basa sul concetto di «rank» (prestigio), poiché più centrali sono i vertici vicini ad un vertice, più centrale è esso stesso.

Indici di centralità

Closeness centrality: In questo indice, la misura della centralità varia inversamente alla distanza totale di un vertice da tutti gli altri.

$$c_{CI}(v) = \frac{1}{\sum_{u \in V} dist(v,u)}$$

Dove dist(v,u) è la distanza geodetica tra i vertici u,v ∈ V. Spesso, per fornire comparazioni fra grafi, e con altre misure di centralità, questa misura è normalizzata affinchè sia nell'intervallo [0,1], attraverso la moltiplicazione per il fattore $(\mathrm{N_v} - 1)$.

Indici di centralità

Betweenness centrality: Introdotta da Freeman, misura la centralità come:

$$c_B(v) = \sum_{s \neq t \neq v \in V} \frac{\sigma(s,t|v)}{\sigma(s,t)}$$

Dove $\sigma(s,t|v)$ è il numero di percorsi minimi fra s e t passanti per v, mentre $\sigma(s,t) = \sum_v \sigma(s,t|v)$. Questa misura può essere ristretta all'intervallo unitario attraverso la divisione per il fattore $[(N_v - 1)(N_v - 2)/2]$.

Il calcolo di tutti gli indici di *betweenness centrality* $c_B(v)$ richiede quindi:

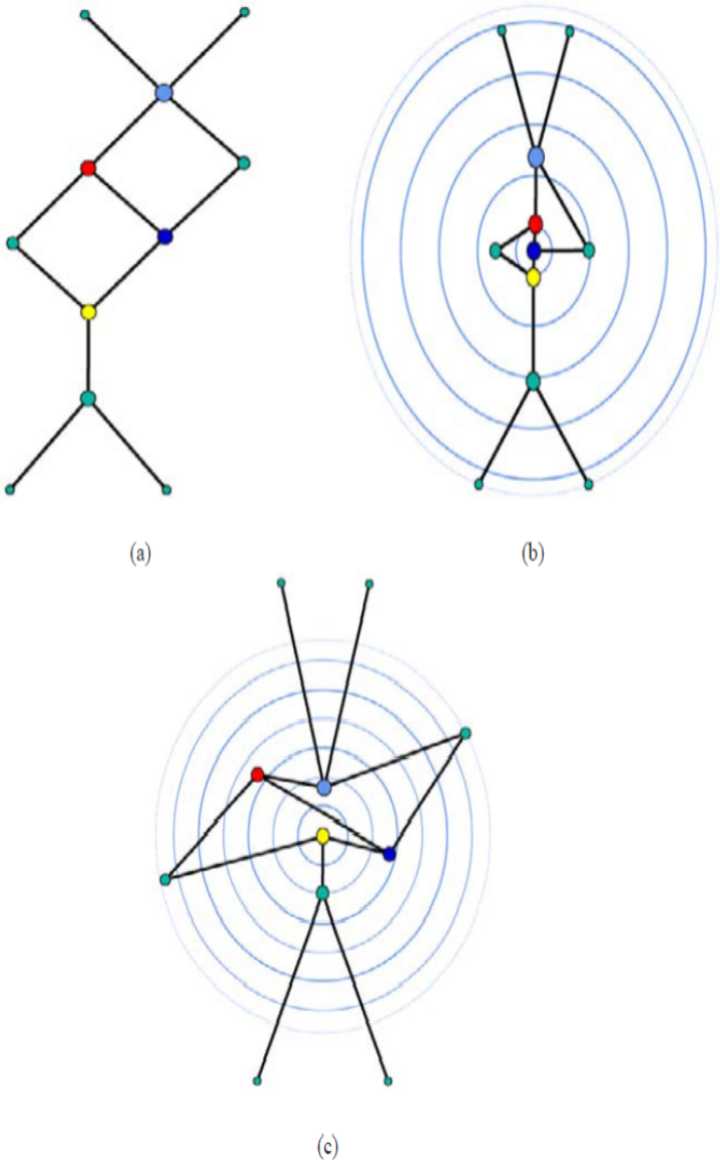
- Il calcolo di tutte le lunghezze dei percorsi minimi tra tutte le coppie di vertici;
- La computazione della sommatoria sopradescritta.

Confronto tra indici di centralità

La figura (a) mostra un grafo di $N_v = 11$ vertici ed $N_e = 12$spigoli. Quale vertice, tolti i 4 estremi, possiamo considerare più centrale?

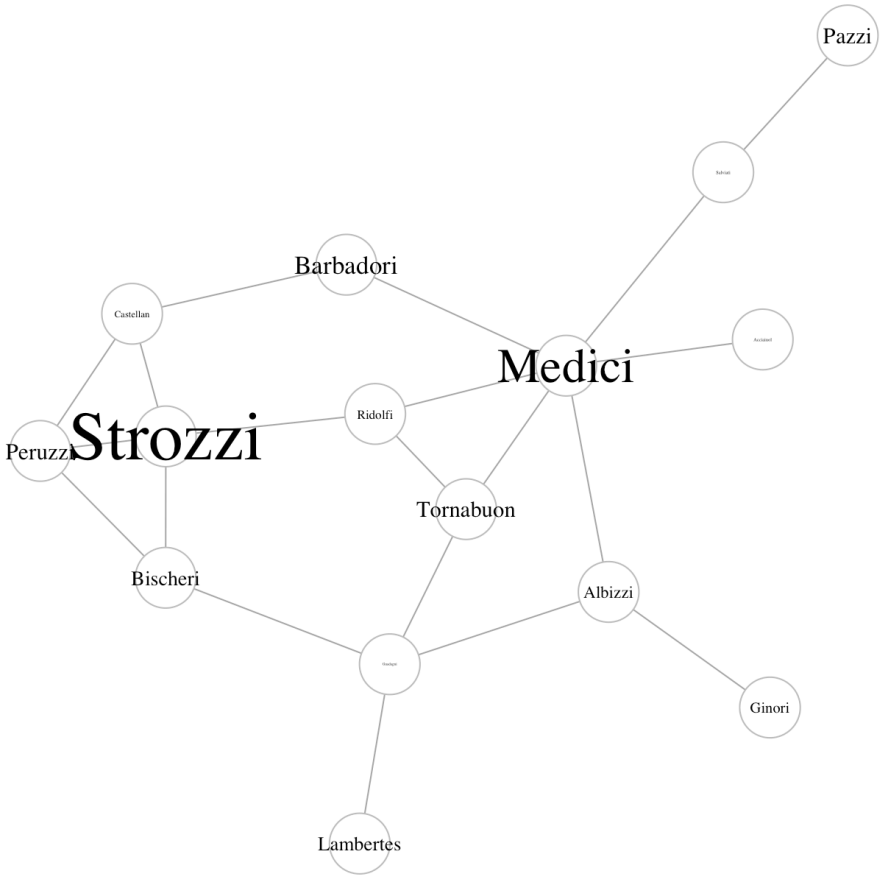
Le figure (b) e (c) mostrano graficamente i risultati del calcolo, rispettivamente, degli indici di Closeness e Betweenness centrality.

Si può notare che, considerando l'indice di Closeness centrality $c_{CI}(v)$, il vertice blu è considerato il più centrale, subito seguito dal rosso e dal giallo. Tuttavia, per l'indice di Betweenness centrality $c_B(v)$, il vertice giallo è considerate maggiormente centrale. Il motivo è che, il fatto che due vertici giacciono vicini, non implica che giacciano all'interno di percorsi minimi fra altri paia di vertici.

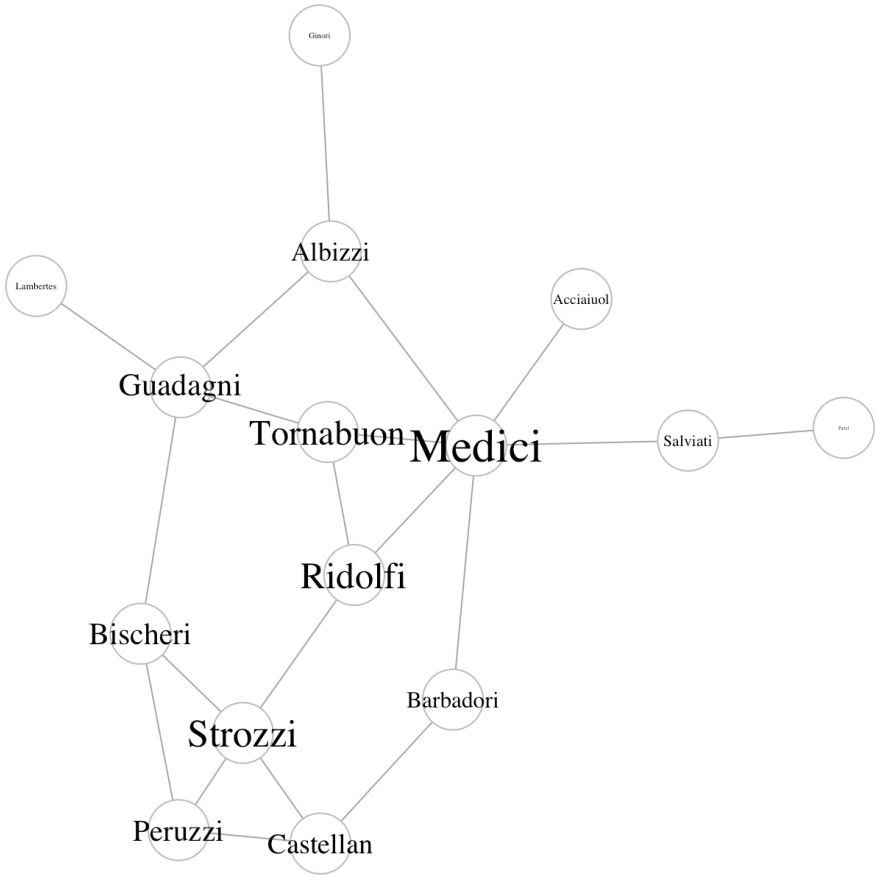


Esempio: famiglie fiorentine

Nome prop. alla ricchezza



Nome prop. alla centralità



Esempio: famiglie fiorentine

	ricchezza	degree	betweenness	closeness	eigenvector	subgraph
Acciaiuol	10	1	0.000000	0.02631579	0.3071155	1.849348
Albizzi	36	3	19.333333	0.03448276	0.5669336	3.592664
Barbadori	55	2	8.500000	0.03125000	0.4919853	2.672602
Bischeri	44	3	9.500000	0.02857143	0.6572037	4.151361
Castellan	20	3	5.000000	0.02777778	0.6019551	4.026544
Ginori	32	1	0.000000	0.02380952	0.1741141	1.656598
Guadagni	8	4	23.166667	0.03333333	0.6718805	4.399924
Lambertes	42	1	0.000000	0.02325581	0.2063449	1.707042
Medici	103	6	47.500000	0.04000000	1.0000000	7.275216
Pazzi	48	1	0.000000	0.02040816	0.1041427	1.598035
Peruzzi	49	3	2.000000	0.02631579	0.6407743	4.564129
Ridolfi	27	3	10.333333	0.03571429	0.7937398	4.275104
Salviati	10	2	13.000000	0.02777778	0.3390994	2.500726
Strozzi	146	4	9.333333	0.03125000	0.8272688	5.632056
Tornabuon	48	3	8.333333	0.03448276	0.7572302	4.318346

Esempio: famiglie fiorentine

```
library(igraph)
library(netrankr)
library(magrittr)
library(ggplot2)
data("florentine_m")
plot(florentine_m,vertex.label.cex=V(florentine_m)$wealth*0.01,vertex.label.color="
black",vertex.color="white",vertex.frame.color="gray")
#Delete Pucci family (isolated)
florentine_m <- delete_vertices(florentine_m,which(degree(florentine_m)==0))
par(mar=c(1,1,1,1),mfrow=c(1,2))
plot(florentine_m,vertex.label.cex=V(florentine_m)$wealth*0.02,vertex.label.color="
black",vertex.color="white",vertex.frame.color="gray",main="Nome prop. alla
ricchezza")
risorse <- data.frame(ricchezza=V(florentine_m)$wealth,degree =
degree(florentine_m),betweenness = betweenness(florentine_m),
  closeness = closeness(florentine_m),eigenvector =
eigen_centrality(florentine_m)$vector,subgraph = subgraph_centrality(florentine_m))
risorse
plot(florentine_m,vertex.label.cex=eigen_centrality(florentine_m)$vector*2,vertex.l
abel.color="black",vertex.color="white",vertex.frame.color="gray",main="Nome prop.
alla centralità")
```

Connessione e coesione del Network

L'analisi della coesione di un network, è volta a verificare se subset di vertici dello stesso, tendono a rispettare la relazione che definisce la connessione all'interno della rete.

Ad esempio, gli amici di una persona centrale nel network, tendono anche loro ad essere amici di qualcun'altro? etc.

Le principali misure della coesione di un network sono:

- Densità locale: riguarda la densità di un subset del network principale. Ad esempio, un clique è totalmente coeso, poiché essendo un sottografo completo, ogni vertice è collegato da uno spigolo (e quindi da una relazione);
- Connettività: riguarda la possibilità di separare un grafo in sottografi distinti fra loro. Un «componente connesso» di un grafo, è un sottografo dove tutti i vertici sono connessi fra loro.

Partizioni del Network

Una partizione $C = \{C_1, \ldots, C_K\}$ di un set S, è una decomposizione di S in K disgiunti subset C_k tali che $\cup K_k=1 C_k = S$.

Nell'analisi dei network, un subset di vertici *coeso*, è un subset che possiede due caratteristiche:

- I vertici al suo interno sono ben connessi fa loro;
- Sono relativamente separati fa gli altri vertici del network.

Il metodo più comune di individuare partizioni è quello della **Clusterizzazione gerarchica**, la quale produce una serie di subset del network, organizzati in maniera gerarchica e presentati graficamente con il cosiddetto *dendrogramma.*

Partizioni del Network

I metodi di clusterizzazione gerarchica possono essere classificati in:

- *Metodi agglomerativi*
- *Metodi divisivi*

I primi basano il procedimento su agglomerazioni successive di partizioni seguendo un processo di fusione *(merging),* mentre i secondi attuano un processo di perfezionamento delle partizioni, mediante successive divisioni (*splitting*).

Ad ogni fase, la partizione analizzata è modificata minimizzando una specifica «misura di costo». Nei metodi agglomerativi, viene quindi eseguita l'unione meno costosa di due singole partizioni, mentre in quelli divisivi è eseguita la divisione meno costosa di un singolo elemento in due distinte partizioni.

Un esempio di clustering agglomerativo

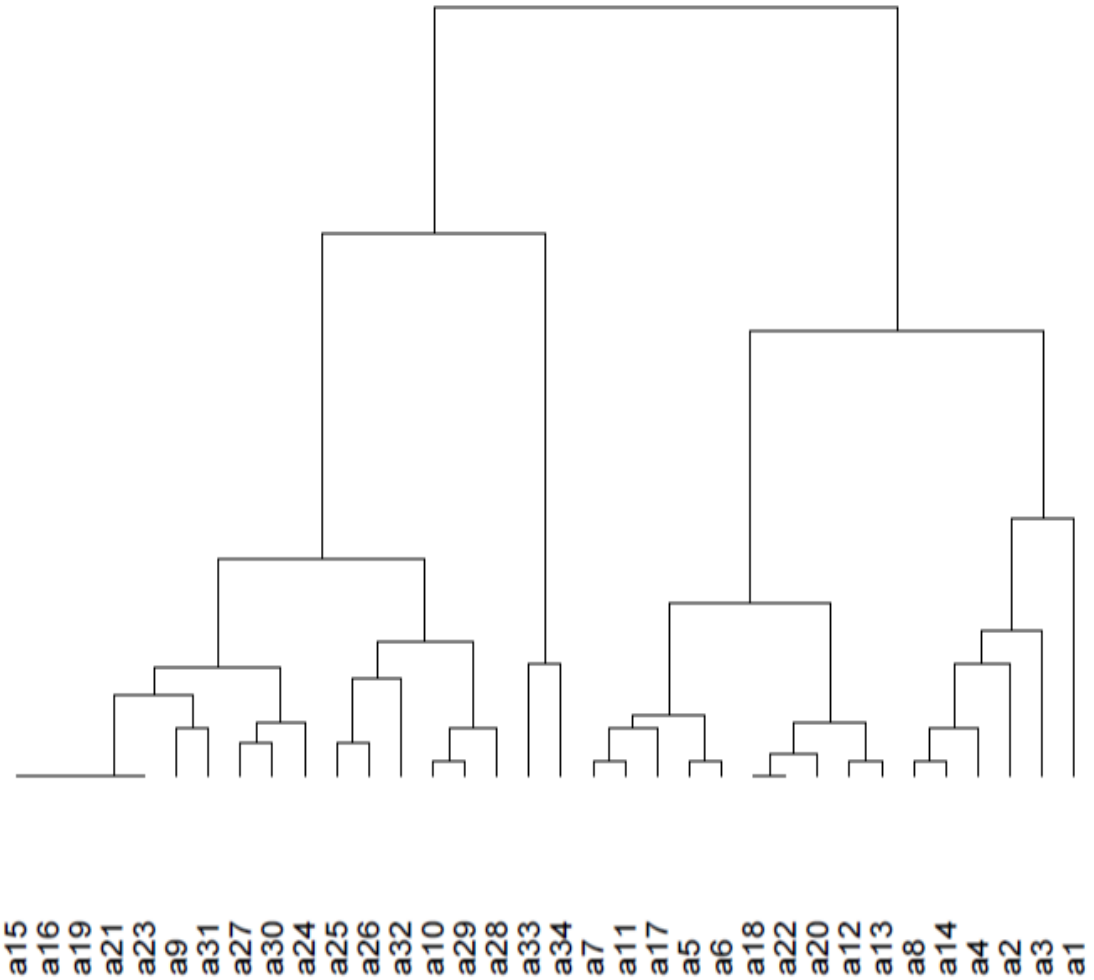
Il dendrogramma che segue, presenta il risultato di un'analisi agglomerativa del network precedente. I risultati sono basati sulla dissimilarità x_{ij}, per vertici v_i e v_j , definita come:

$$x_{ij} = \frac{|N_{vi} \,\Delta\, N_{vj}|}{d_{(N_v)} + d_{(N_v-1)}},$$

Dove N_v è il set dei vertici adiacenti ad un vertice *v*, 'Δ' indica la differenza simmetrica dei due set, ovvero il set di elementi che sono in uno, o in un altro, ma non in entrambi; d_i è l'i-esimo elemento più piccolo nella sequenza dei gradi.

In altre parole, x_{ij} è il numero di vertici vicini a v_i e v_j, ma non condivisi, normalizzato nell'intervallo [0,1] dove 0 ed 1 indicano, rispettivamente, perfetta similarità o dissimilarità.

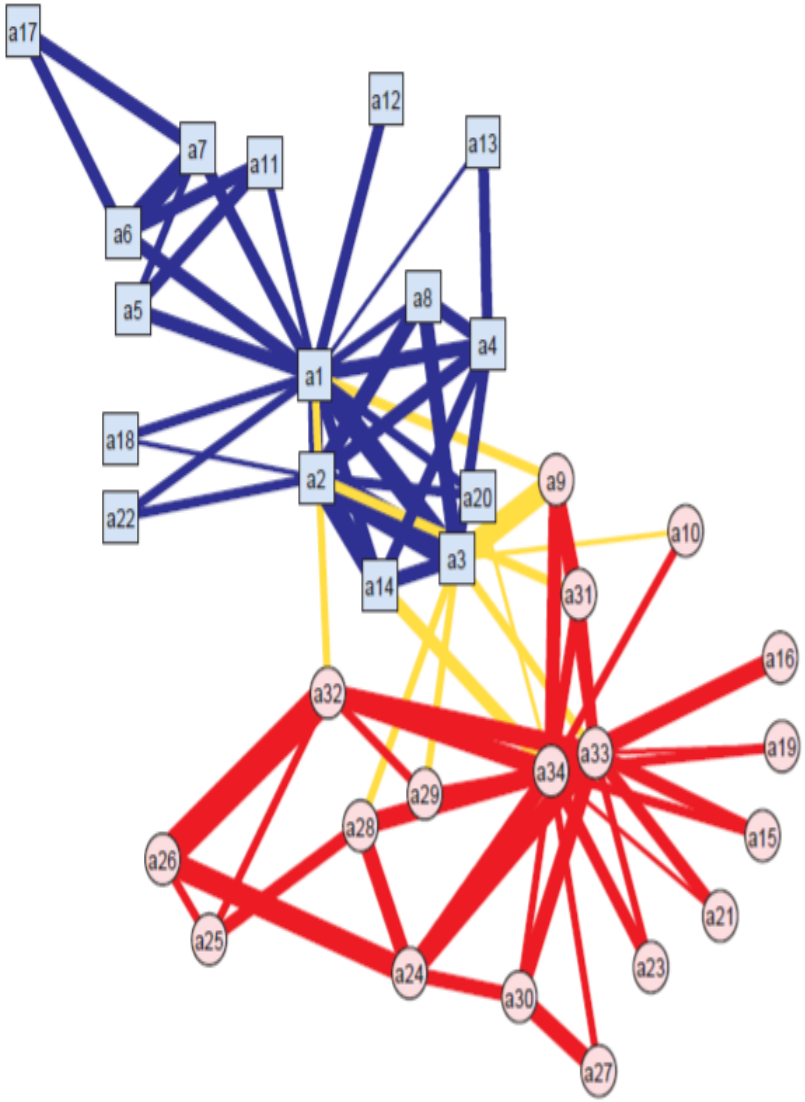
Dendrogramma del Network di club di Karate



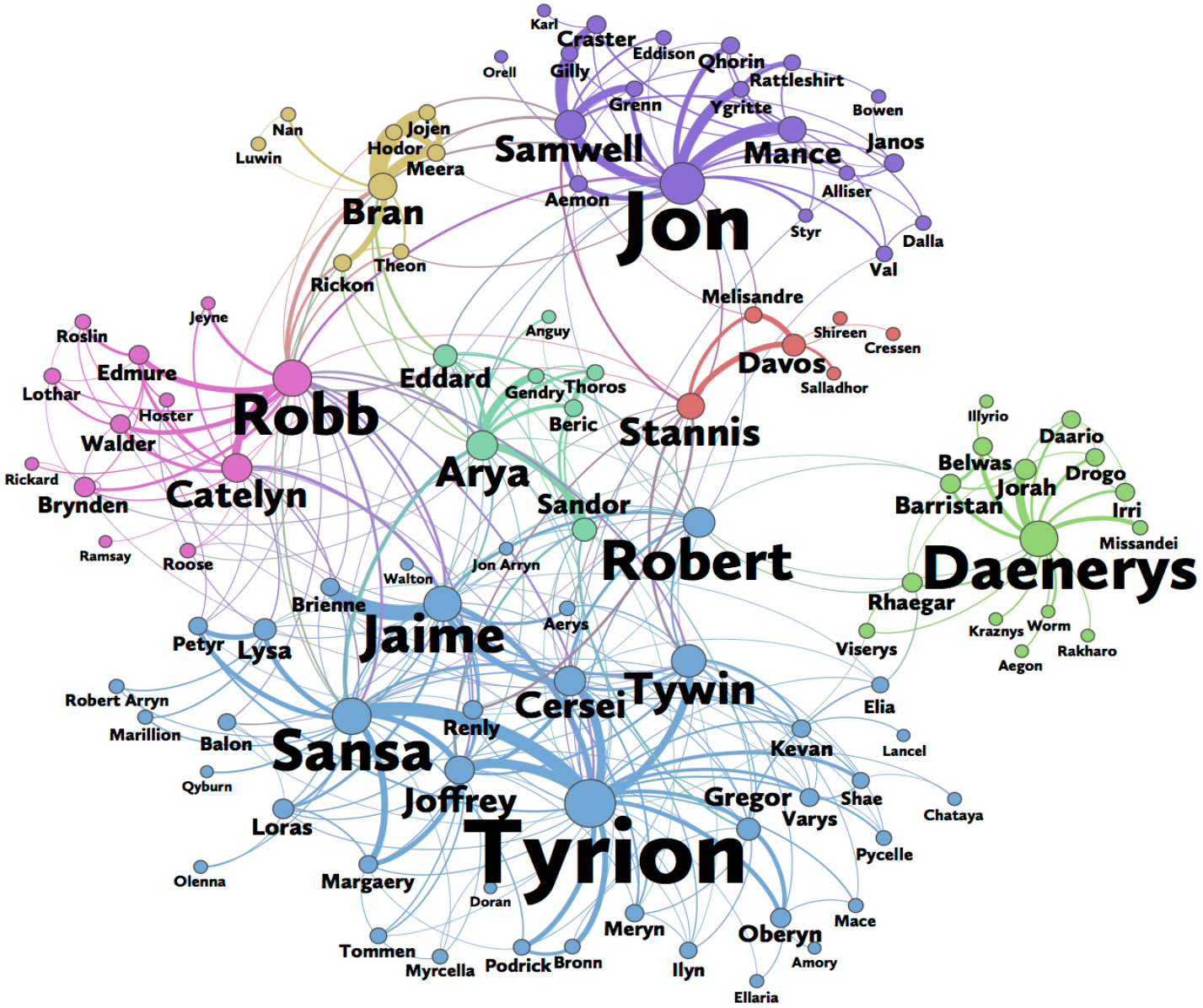
I social Network

Il Network mostra un gruppo di partecipanti ad un club di Karate, osservato da Zachary nel 1970. Le connessioni indicano relazioni sociali tra i componenti, rappresentati dai vertici del grafo. Si possono notare due sottogruppi principali, centrati attorno agli elementi 1 e 34, e composti da elementi raffigurati con forme e colori differenti (cerchi rossi e quadrati blu).

Le connessioni fra elementi dello stesso sottogruppo sono dello stesso colore (rosso e blu), mentre connessioni fra elementi dei diversi sottogruppi sono raffigurate in giallo

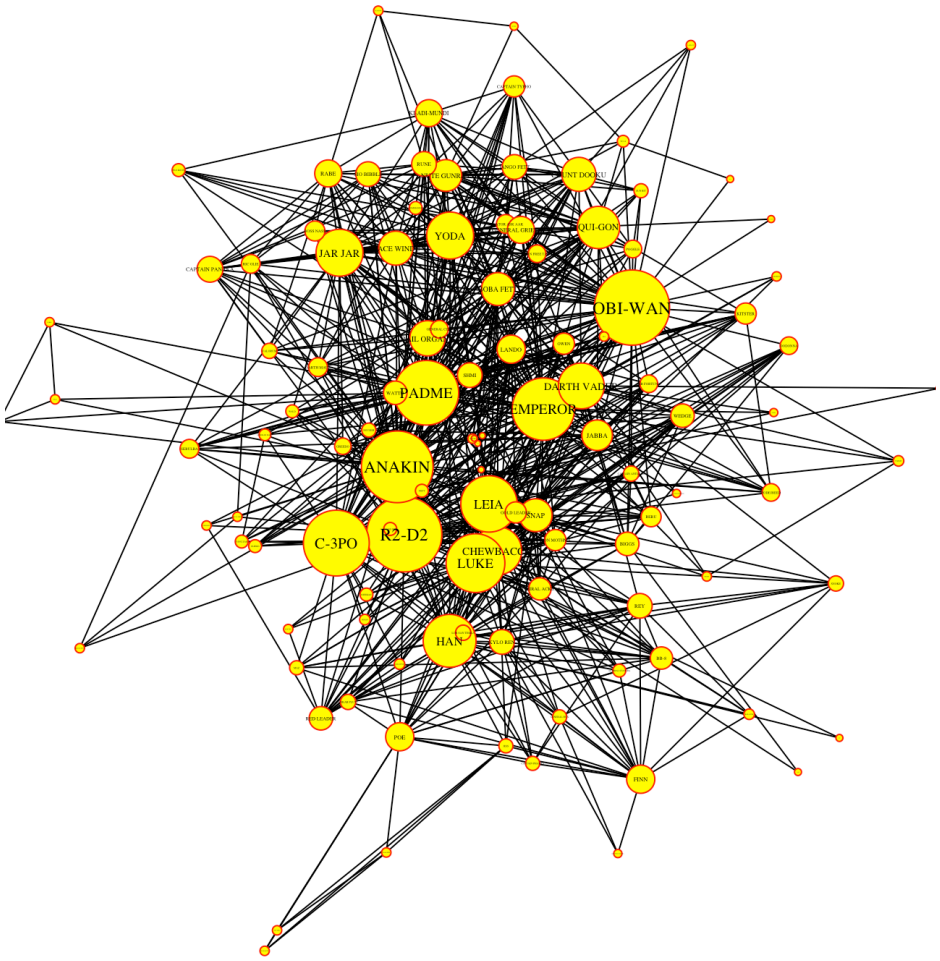


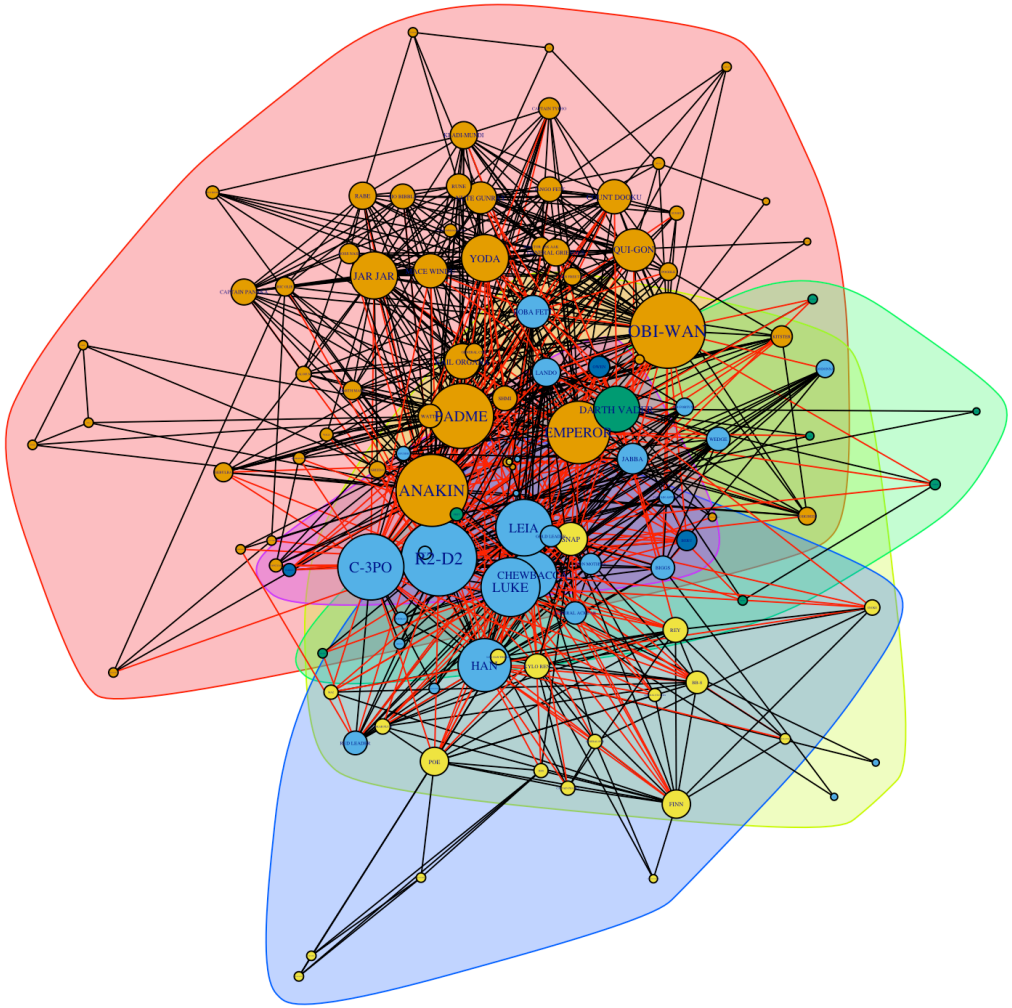
Relazioni sociali complesse: Game of Thrones Network



Esempio: personaggi di Guerre Stellari

Nome prop. alla centralità





Esempio: personaggi di Guerre Stellari

	degree	betweenness	closeness	eigenvector	subgraph
HAN	41	273.3905321	0.002857143	1.000000e+00	2.144158e+09
C-3PO	52	237.3491943	0.002873563	9.638816e-01	4.169590e+09
CHEWBACCA	37	453.9678644	0.003134796	9.567137e-01	1.930895e+09
LUKE	46	682.2037778	0.003144654	9.188490e-01	2.732553e+09
LEIA	44	481.2867719	0.003194888	9.097087e-01	2.741573e+09
R2-D2	60	793.5831053	0.003154574	8.765661e-01	5.207872e+09
OBI-WAN	60	640.1329640	0.002976190	6.513951e-01	5.110530e+09
ANAKIN	57	542.5774404	0.003289474	5.173688e-01	4.733398e+09
PADME	51	513.4290564	0.003154574	4.177036e-01	3.999679e+09
LANDO	20	65.0100567	0.002976190	3.438390e-01	1.174334e+09
EMPEROR	49	362.7540474	0.003076923	2.844444e-01	3.848670e+09
DARTH VADER	35	500.9380354	0.003095975	2.590184e-01	2.198761e+09
QUI-GON	32	31.9139000	0.002604167	2.445354e-01	1.989038e+09
JAR JAR	36	173.9023195	0.002747253	2.114518e-01	2.392227e+09
REY	17	94.2251140	0.002604167	2.092167e-01	4.291558e+08
FINN	20	57.9099231	0.002262443	2.091625e-01	3.560521e+08
YODA	36	300.8675804	0.002873563	1.878921e-01	2.946740e+09
BB-8	15	2.0861472	0.002457002	1.537799e-01	3.812018e+08
JABBA	22	156.6366916	0.002976190	1.459943e-01	1.311605e+09
COUNT DOOKU	25	170.1754435	0.002659574	1.025377e-01	1.588279e+09