COMPORTAMENTI LOCALI
E
FENOMENI GLOBALI

COMPORTAMENTI LOCALI E FENOMENI GLOBALI

Le reti sono uno strumento utilissimo per descrivere il rapporto tra i *comportamenti locali* dei singoli ed i *fenomeni globali* che si verificano in una popolazione.

Ad esempio, per rispondere alle seguenti questioni fondamentali.

- Come si evolvono i flussi di informazione in una rete sociale?
- Come nodi diversi possono svolgere ruoli distinti nel processo?
- Come ciò può influire sull’evoluzione della rete nel tempo?

Esperimento di Granovetter

- Il sociologo Mark Granovetter per la sua tesi di dottorato intervistò numerose persone che avevano appena trovato lavoro e chiese come erano venuti a conoscenza dell’offerta di lavoro
- Risultò che la maggior parte delle persone avevano saputo le informazioni mediante *contatti personali*
- Inoltre descrivevano questi contatti più spesso come *conoscenti* che come amici

Se gli amici dovrebbbero essere quelli che più ci tengono ad aiutare e fornire notizie utili *perché i conoscenti risultano più utili nella ricerca di un nuovo lavoro?*

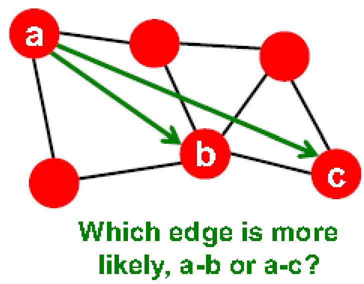
Granovetter stabilisce una connessione tra **ruolo sociale e strutturale** di un edge: gli dge strutturalmente incorporati (**short** range) sono anche *socialmente forti,* mentre gli edge a **lungo** raggio che abbracciano diverse parti del rete sono *socialmente deboli.*

Questa differenziazione si riflette sul flusso di informazioni all'interno della rete: gli edge a lungo raggio consentono di raccogliere informazioni provenienti da diverse parti della rete (es. poer trovare un lavoro); gli edge strutturalmente incorporati sono ridondanti in termini di accesso alle informazioni.

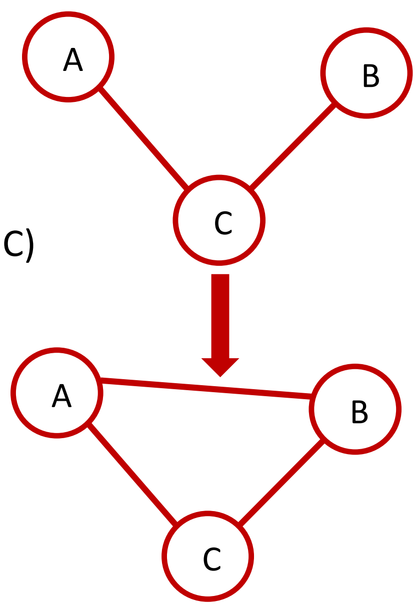
In altri termini, i soggetti aventi legami deboli, fatti cioè di conoscenze amicali non troppo strette, hanno più possibilità di accesso ad informazioni (es. potenziali posizioni lavorative di proprio interesse), rispetto a coloro che investono socialmente soltanto nei legami forti, quali i familiari, i parenti e gli amici intimi.

Quali sono i meccanismi con cui gli edge si formano e svaniscono?

Principio di base (**Triadic Closure**): Se due persone in una rete sociale hanno un amico in comune, allora vi è una maggiore probabilità che essi diventeranno essi stessi amici ad un certo momento futuro.



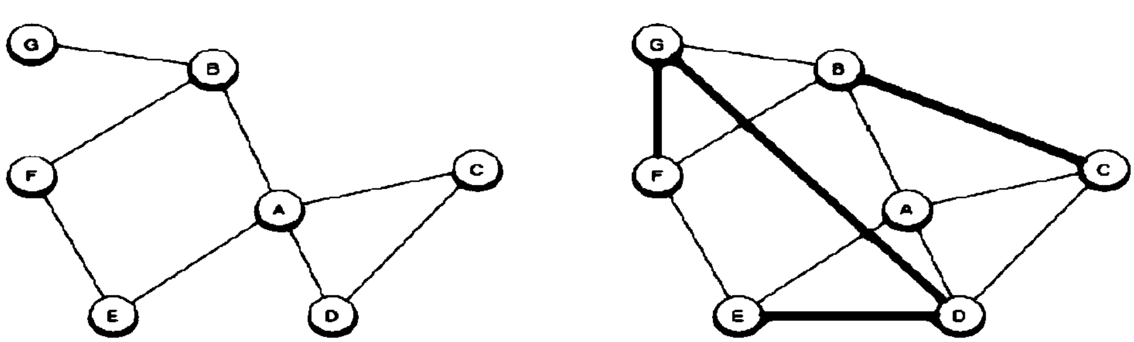
Triadic Closure: Motivazioni sociologiche



Se A e B hanno un amico C in comune, allora:

- è più probabile che A e B si incontrino
  (dal momento che entrambi trascorrono del tempo con C)
- A e B tendono ad avere fiducia l'un l'altro
  (dal momento che hanno un amico in comune)
- C ha un incentivo a far incontrare A e B
  (è più difficile per C mantenere due rapporti disgiunti)

Se osserviamo l'evolversi di una rete sociale nel tempo noteremo la creazione di nuovi edge nel tempo e la tendenza alla chiusura dei triangoli

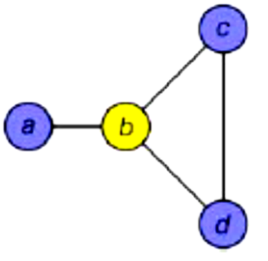


Nota: come si forma l'edge G-D? (per altri motivi)

Coefficiente di clustering

Il coefficiente di clustering di un nodo v è la frazione di coppie di amici di v che sono collegati tra loro da edges

Esempio: b ha 3 amici, quindi potrebbero esister tre possibili coppie di amici, di queste una è collegata.
Di conseguenza il coefficiente di clustering di b è 1/3



Per un nodo u, sia n_u il numero dei suoi vicini e sia e_u il numero di edges tra nodi del suo vicinato

Il **coefficiente di clustering** di u è definito come

$$C_u := \frac{e_u}{\binom{n_u}{2}} = \frac{e_u}{\frac{n_u(n_u-1)}{2}}$$

Il coefficiente di clustering di un nodo è sempre compreso tra 0 e 1:

è 0 se nessuno degli amici del nodo sono amici tra loro ,

è 1 se tutti gli amici del nodo sono amici tra di loro.

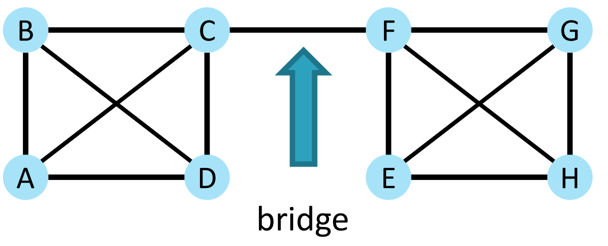
Il coefficiente di clustering può anche essere pensato come la probabilità che due amici scelti a caso di u sono amici tra di loro.

Notiamo che più è forte la triadic closure, maggiore sarà il coefficiente di clustering

BRIDGE

Un arco A-B è un **bridge** se la sua rimozione porrebbe A e B in componenti connesse distinte del grafo.

- Un bridge è un nodo che permette l'accesso tra parti della rete che sarebbero irraggiungibili altrimenti.
- Essere un bridge è una proprietà globale della rete



Note:

- un bridge non appartiene a triangoli (perchè?)
- I bridge sono estrememente rari in reti reali, è stato quindi introdotto il concetto di bridge locale

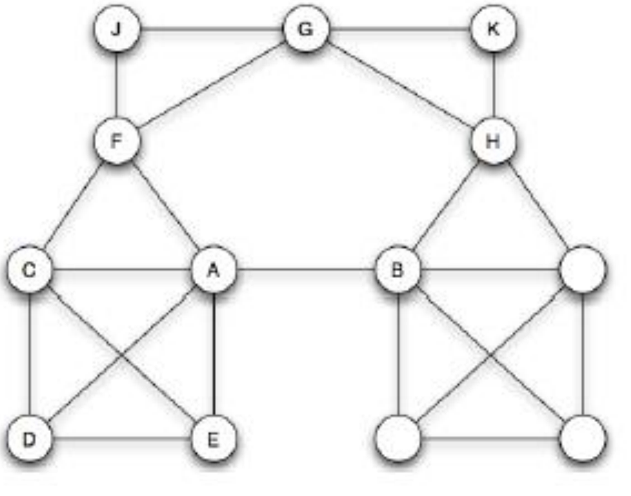
LOCAL BRIDGE

Un arco tra due nodi A e B in un grafo è un **bridge locale** se i suoi estremi A e B non hanno amici in comune

- Per ogni vertice u, l'insieme N(u) rappresenta il vicinato (neighborhood) di u, cioè l'insieme dei vicini di u.
- Arco {A, B} è un bridge locale se e solo se

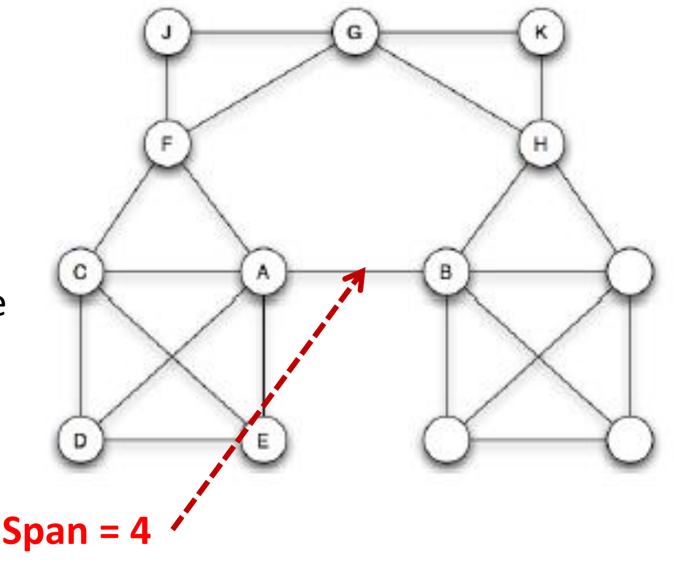
$$(N(A) \setminus \{A\}) \cap (N(B) \setminus \{B\}) = \emptyset$$

Nota: Eliminando A-B, la distanza tra A e B aumenta ad un valore strettamente >2.



Lo **span** di un local bridge è la distanza tra i suoi estremi se togliessimo l'arco.

Maggiore è lo span più importante è il ruolo di coesione dell'arco



Local Bridge ed Esperimento Ganovetter

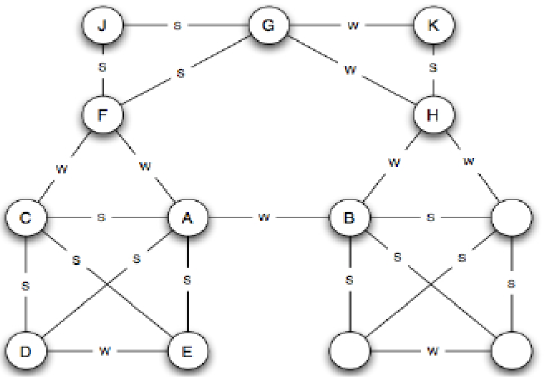
- *Un gruppo stretto (connesso da legami forti) è formato da individui aventi le stesse nostre stesse informazioni*
- Local bridge (legame debole) connette con un altro *gruppo stretto, quindi probabilmente da accesso ad altre informazioni*

STRONG TRIADIC CLOSURE

- **Possiamo etichettare edge come legami Forti (S) o deboli (W).**

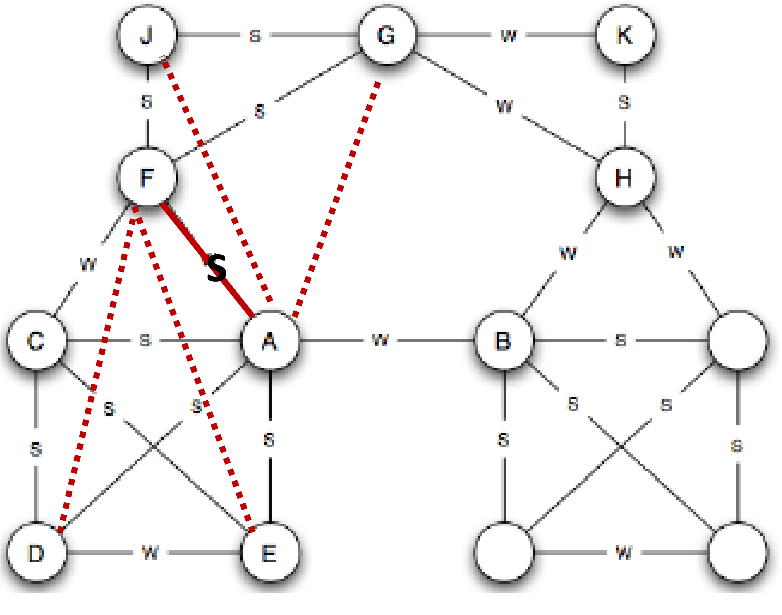
Il nodo A soddisfa la **Strong Triadic Closure** se

per ogni scelta di vicini B e C di A tali che AB e AC sono legami forti risulta che BC è un legame (forte o debole).



Nessun nodo della figura viola la proprietà di **Strong Triadic Closure.**

Se edge A-F dovesse essere un legame forte , piuttosto che un debole, i nodi A e F entrambi violerebbero la proprietà.

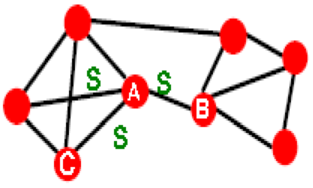


Notiamo che la proprietà di Strong Triadic Closure è troppo estrema per valere per tutti i nodi di una grande rete sociale, ma è un passo utile come astrazione della realtà epermette di ragionare ulteriormente circa le conseguenze strutturali di legami forti e deboli.

WEAK TIES

Assumiamo la Strong Triadic Closure (STC).

Se un nodo ha almeno due strong ties allora ogni local bridge a cui è adiacente è un weak tie



- Supponiamo A soddisfa STC ed ha due strong ties
- Sia AB un local bridge ed anche strong tie
  - BC deve esistere per STC
  - BC non può esistere perchè AB non può stare in un triangolo
  - ➔CONTRADDIZIONE
- Quindi AB non può essere uno strong tie

Quindi

1. Se la rete ha un numero sufficiente di strong ties allora ogni local bridge è un weak tie
2. Le informazioni cruciali viaggiano sui weak tie
3. Un weak tie collega una proprietà globale (local bridge) ad un comportamento locale (strength of tie)

FORZA DEI LEGAMI IN DATA-SET REALI

Per molti anni non si è potuto verificare le teorie di Granovetter (connessioni tra forza dei legami sociali e proprietà strutturali della rete) su network di grandi dimensioni e in contesti realisti per mancanza di dati concreti.

Oggi abbiamo dati per reti “who-talks-to-whom” di grandi dimensioni, quali Facebook, social networks, tabulati telefonici, email,...

Onnela e al. Nel 2007 hanno effettuato un esperimento con dati di un provider di telefonia cellulare che copriva il 20% della popolazione statunitense:

- Periodo di osservazione 18 settimane
- Due utenti sono collegati se hanno scambiato almeno una telefonata in entrambe le direzioni
- La forza del legame è misurata dal numero di minuti di comunicazione

L‘esperimento ha confermato la teoria di granovetter.

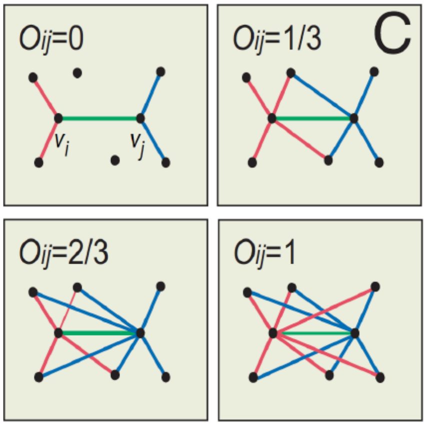
Oltre le dicotomie legame forte-debole e bridge si-no: Neighborhood Overlap

Il neighborhood overlap di due nodi adiacenti i e j è definito

$$O_{ij} = \frac{\#(\text{vicini comuni ad i e j})}{\#(\text{vicini di i o j})}$$

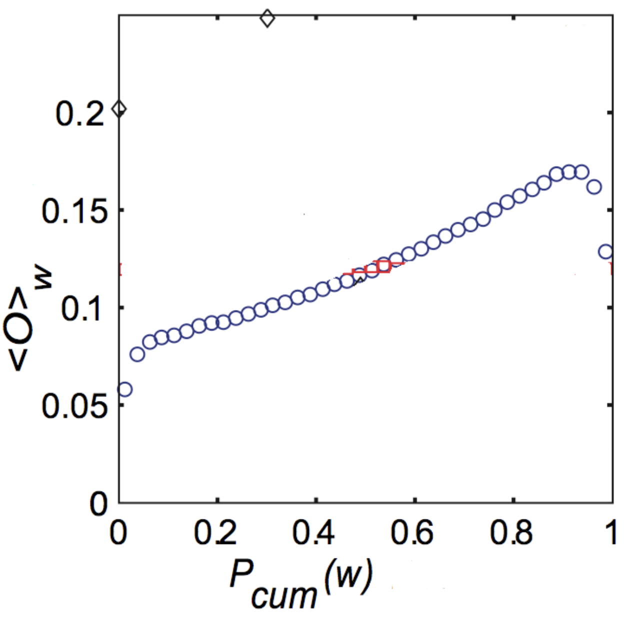
Notiamo che se l'edge (i,j) è un local bridge allora ha neighborhood overlap pari a 0

Possiamo vedere edge con con un *piccolo* neighborhood overlap come "*quasi*" dei local bridge.



Ritornando all' esempio dell'esperimento sulla telefonia cellulare, il grafo in figure mostra il neighborhood overlap degli archi in funzione del percentile tra gli archi ordinati per forza del legame (es. numero chiamate)

Al crescere dei valori sull'asse delle x, otteniamo archi di sempre maggiore forza, e poiché la curva aumenta in modo lineare, abbiamo anche archi aventi sempre maggiore Neighborhood Overlap



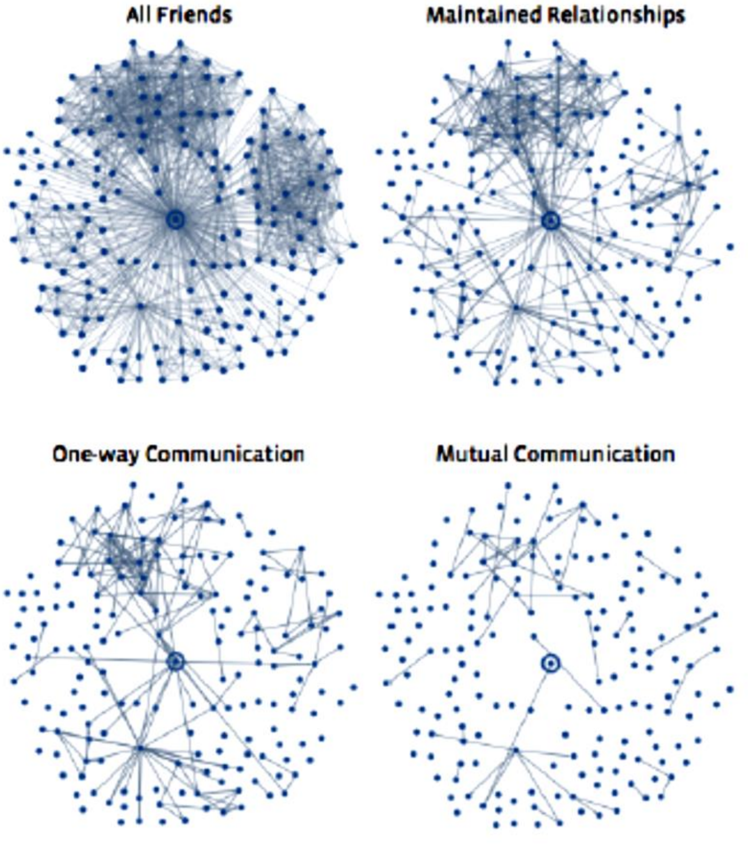
RISULTATI EMPIRICI SU NEIGHBORHOOD OVERLAP

Riportiamo in seguito i risultati di osservazioni su alcune reti reali.

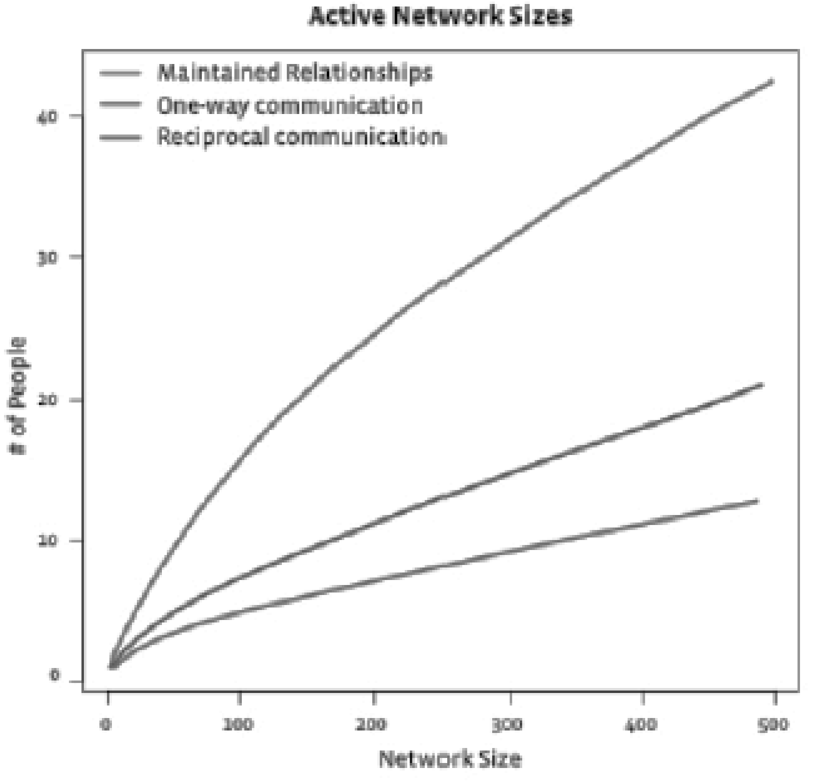
Facebook.
Un link (Facebook) rappresenta

- **Comunicazione reciproca** se ciascun utente ha inviato messaggi all'amico all'altra estremità del collegamento, e ha anche ricevuto i suoi messaggi durante il periodo di osservazione.
- **Comunicazione unidirezionale** se l'utente invia uno o più messaggi all’amico all'altra estremità del collegamento (sia che questi messaggi sono stati ricambiati o meno).
- **Relazione mantenut**a se l'utente ha seguito l’amico all'altra estremità del collegamento, anche in assenza di comunicazione vera e propria (es. visitando il profilo di un amico più di una volta).

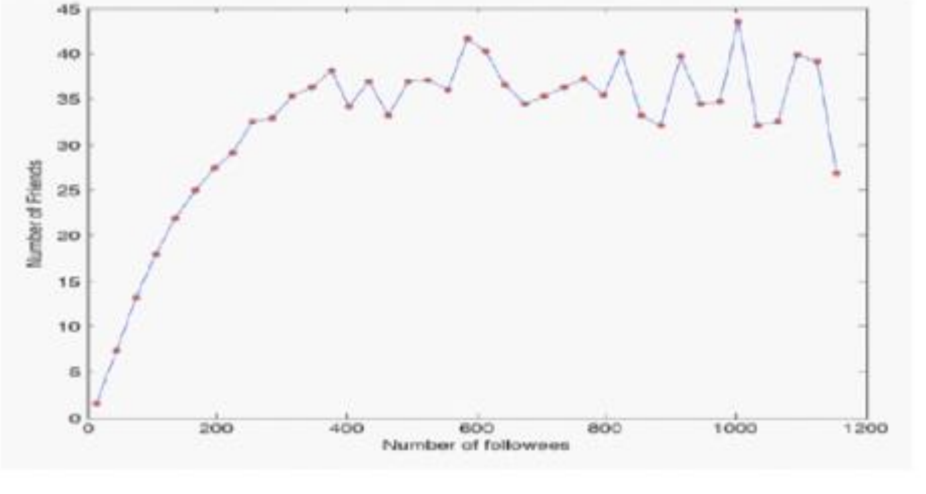
Vicinati degli utenti Facebook da differenti punti di vista



Il numero di link corrispondenti ai rapporti mantenuti, comunicazione unidirezionale e reciproca, come funzione della dimensione del neighborhood per gli utenti su Facebook.



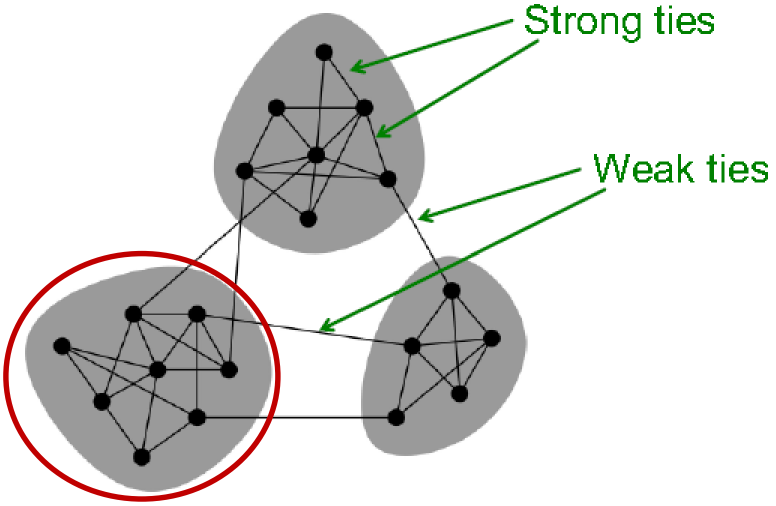
Twitter. La figura riporta Il numero totale di legami forti di un utente (definito da messaggi diretti multipli) in funzione del numero dei followee che ha su Twitter.



PARTIZIONAMENTO DELLE RETI

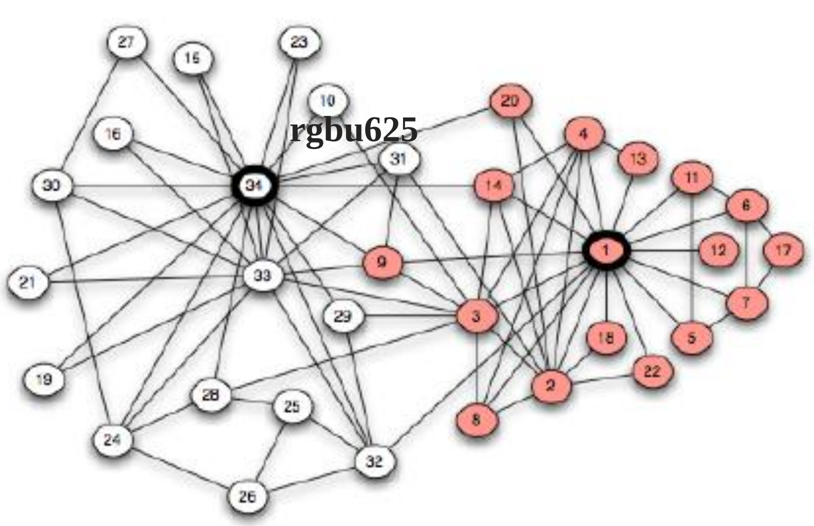
La teoria di Granovetter suggerisce che le reti sono costituite da gruppi di nodi strettamente collegati

Comunità: insieme di nodi con *molti* di collegamenti all'interno e pochi verso l'esterno (il resto della rete)



La figura seguente mostra il partizionamento naturale della rete Karate Club:

Ci chiediamo se i confini dei due club possono essere predetti dalla struttura di rete.



Dalla discussione precedente ricaviamo un'idea di struttura di una rete sociale: Un insieme di comunità fittamente connesse al loro interno e legate tra di loro da legami occasionali.

Il problema è: come possiamo individuare le aree fittamente connesse?

Si tratta di un problema algoritmico non banale che ha un ruolo importante in diversi contesti. Ad esempio, identificare le comunità della rete è vitale per gli esperti di marketing.

Sono stati proposti numerosi metodi euristici per il partizionamento delle reti. Tali metodi possono essere suddivisi in due grandi classi:

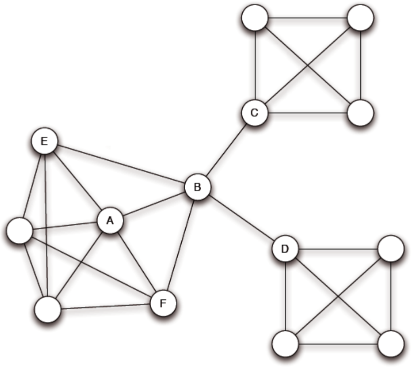
- Metodi agglomerativi: Partendo dai singoli nodi, si cerca di aggregarli in gruppi (le comunità);
- Metodi divisivi: partendo dalla rete nel suo complesso, si cerca di dividerla in sottoreti (le comunità) connesse in modo sparso tra loro.

NODI/EDGE DIVERSI ⇔ RUOLI DIVERSI

L' *embeddedness* di un arco AB è definita come #(vicini comuni ad A e B)

Notiamo che

- L'embeddedness il numeratore della funzione di neighborhood overlap
- Un local bridge ha embeddedness nulla
- La fiducia tra i due nodi adiacenti cresce con l'embeddedness dell'arco che li unisce
- Se un nodo è adiacente ad archi con alta embeddedness ha più facilità di relazione e maggiore fiducia nei suoi vicini



Es. Nella figura A è circondato da archi con alta embeddedness: èincluso in una comunità molto coesa

MISURE DI CENTRALITÀ DI UN NODO

Le misure di centralità di un nodo sono state introdotte per misurare il ruolo svolto dal nodo nella rete

Sono state proposte diverse misure di centralità legate a aspetti diversi della rete, tra cui:

- Degree centrality
- Closeness centrality
- Betweenness centrality

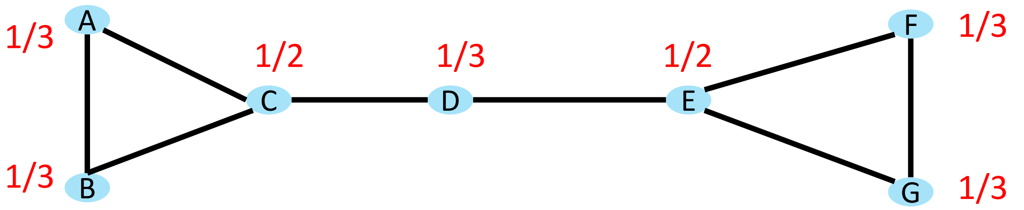
Sono tutte definite in modo da fornire un valore in [0, 1] che dia una misure dell'"importanza" del nodo nella rete

La *degree centrality* di un nodo z è definita come

$$\deg(z)/(n\text{-}1),$$

dove deg(z) è il grado del nodo z e n è il numero di nodi della rete. Essa misura l'importanza del nodo in base al numero dei suoi vicini.

La figura seguente riporta per ogni nodo la corrispondente degree centrality.



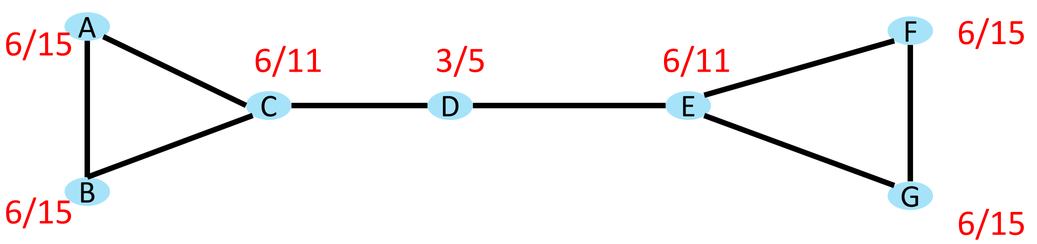
La *closeness centrality* di un nodo z è definita come

$$(n\text{-}1)/\sum_{u\neq z} d(u,z)$$

Dove d(u,z) è la distanza tra u e z nella rete.

Notiamo che essa coincide con l'Inverso della distanza media dei nodi da z, quindi misura la velocità con cui un nodo viene raggiunto da tutti gli altri nodi (o raggiunge tutti gli altri nodi).

La figura riporta per ogni nodo la corrispondente closeness centrality.



La *betweenness centrality* di z è definita come

$$2\,/\,(n\text{-}1)(n\text{-}2)\sum_{u\neq v,\, z\,\neq\, u,v} P_z(u,v)\,/\,P(u,v),$$

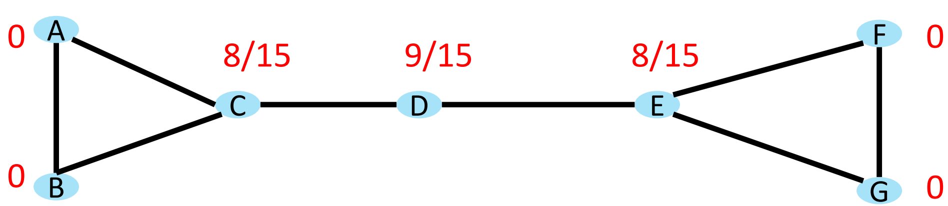
Dove

$P(u,v)$ è il numero cammini minimi tra u e v

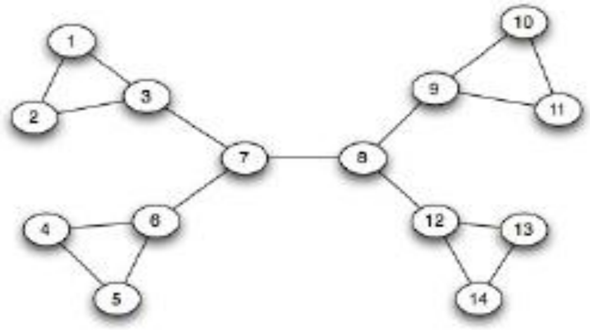
$P_z(u,v)$ = # cammini minimi tra u e v passanti per z

La betweenness centrality di z misura quanto il nodo z è cruciale per la comunicazione tra tutte le coppie di nodi.

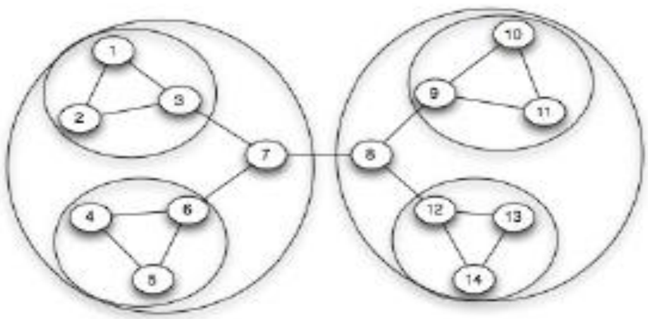
La figura riporta per ogni nodo la corrispondente betweeness centrality.



PARTIZIONAMENTO DELLE RETI. DOVE PARTIZIONARE?



(a) *A sample network*



(b) *Tightly-knit regions and their nested structure*

Bridge e local bridges spesso collegano parti della rete debolmente interagenti; si può provare a rimuovere prima questi bridges

Basta individuare i bridge? La risposta è no. I bridge e bridge locali spesso collegano parti della rete debolmente interagenti; si può quindi provare a rimuovere prima questi ponti.

Idea per il partizionamento: Definire il concetto di "traffico" sulla rete e cercare gli archi che portano la maggior parte di questo traffico.

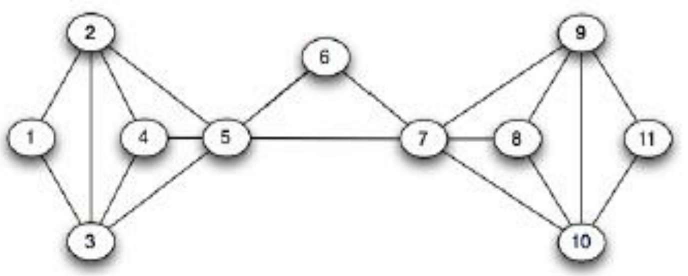
- Ci aspettiamo che questi archi colleghino nodi che si trovano in differenti regioni (densamente collegate), e quindi sono buoni candidati per la rimozione.

Traffico: Per ogni coppia di nodi A e B nel grafo che sono collegati da un cammino, immaginiamo vi sia un'unità di *''flusso*" lungo gli archi da A a B.

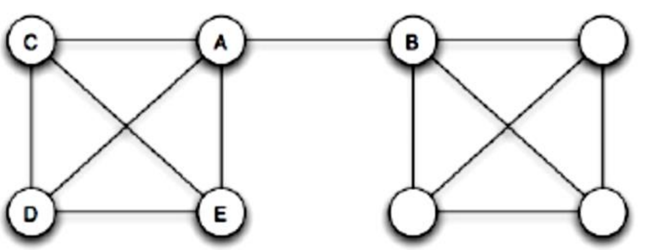
Se A e B appartengono a diverse componenti connesse, nessun flusso scorre tra di loro.

Il flusso tra A e B si divide in modo uniforme lungo tutto il possibili percorsi più brevi da A a B:

➔ se ci sono k cammini minimi da A e B, 1 / k di unità passa lungo ciascuno.



La betweeness di un arco è la quantità totale di flusso che esso porta, contando il flusso sui cammini minimi tra tutte le coppie di nodi che utilizzando questo arco.



Come si fa a calcolare la betweeness di un/ogni arco?

Eliminazione archi : Algoritmo di Girvan-Newman

1. **Trovare** l’arco di massima betweenness (o gli archi se c’è un pareggio) e rimuovere questi archi dal grafo.

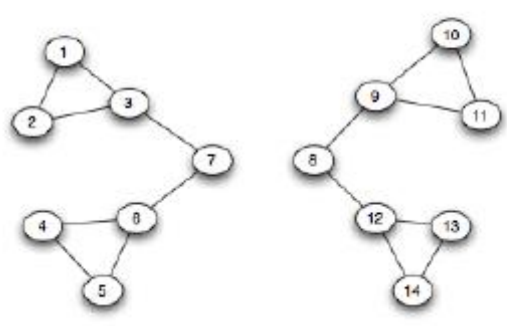
   / * Questo può disconnettere il grago. Se è così, questo è il primo livello di regioni del partizionamento del grafo. * /

   ricalcolare le betweenness, e ancora una volta togliere l’arco o archi di massima betweenness.

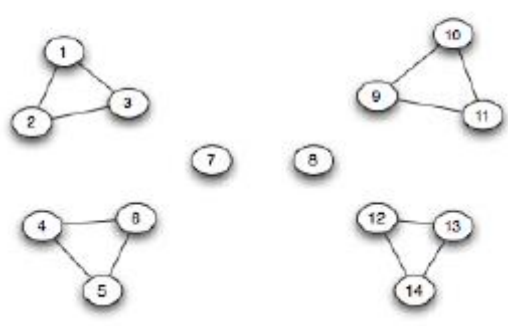
   / * Questo può decompporre alcune componenti esistenti in componenti più piccole; in tal caso, queste sono le regioni annidate all'interno delle regioni più grandi. * /

2. **Iterare:** Procedere in questo modo finché rimangono archi nel grafo, in ogni fase ricalcolando tutte le betweenness e rimuovendo l’arco o gli archi di betweness massima

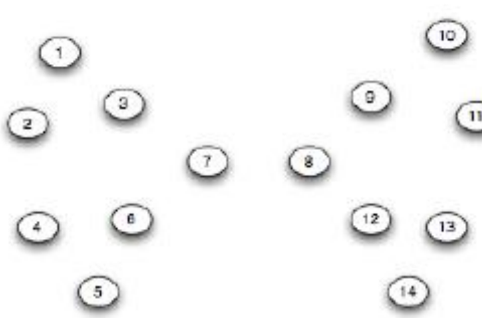
Esempio.



(a) *Step 1*

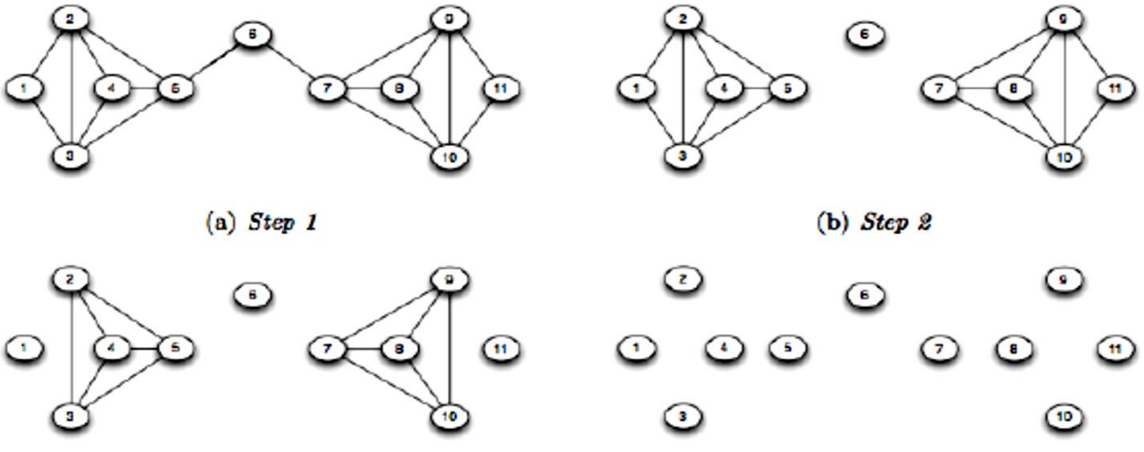


(b) *Step 2*



(c) *Step 3*

Esempio.



(a) *Step 1* (b) *Step 2*

(c) *Step 3* (d) *Step 4*

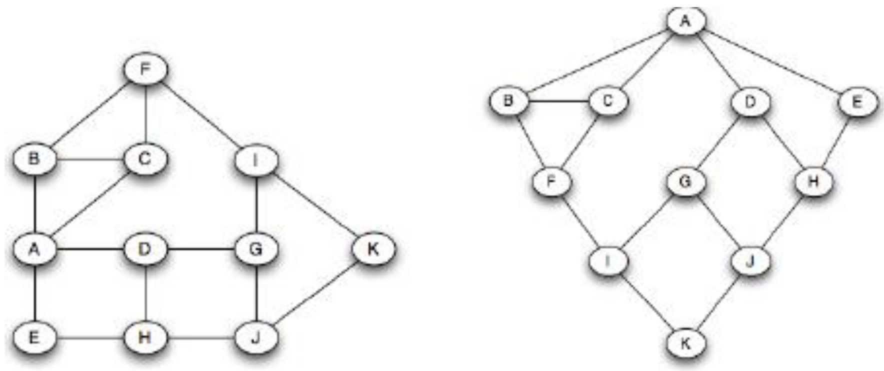
COME SI FA A CALCOLARE I VALORI DI BETWEENESS?

L'algoritmo di Girman-Newman ad ogni passo deve ricalcolare la betweenness di tutti gli archi rimanenti nel grafo. Si può fare efficientemente? L'idea per farlo efficientemente è di utilizzare una BFS da ogni nodo *u* della rete. In particolare calcoliamo come un'unità di flusso si distribuisce nella rete a partire da un nodo *u.*

Algoritmo

- A partire da ogni nodo u esegui una BFS e costruisci l'albero dei percorsi minimi
- Per ogni nodo *v*, calcola quanti cammini minimi ci sono da *u* a *v*
  - Se il nodo *v* si trova al *k*-imo livello dell'albero i cammini minimi da *u* a *v* hanno lunghezza *k* e passano per i padri di *v* nell'albero
  - Il numero dei cammini minimi per v è la somma del numero di cammini minimi per i suoi padri
- Determina quanto flusso attraversa ogni arco del grafo
  - Bottom up
  - Ad ogni nodo assegna un'unità di flusso più tutto quello che gli arriva dai figli
  - Ogni nodo distribuisce il proprio flusso tra i padre in proporzione al numero di cammini che passano per ogni padre.

In figura sono visualizzati i risultati della BFS dal nodo A;
la ricerca viene eseguita in ampiezza da ogni nodo a turno.

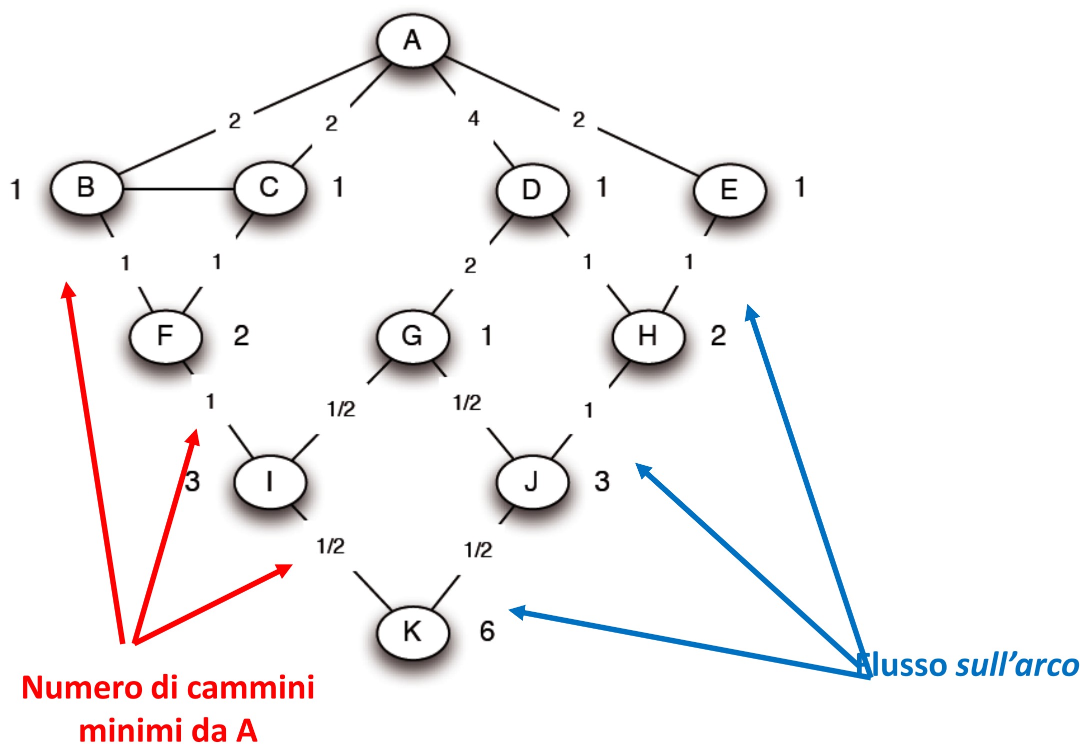


PRIMA FASE: TROVA IL NUMERO DI PERCORSI MINIMI DA A A CIASCUNO ALTRO NODO.

Può essere fatto sommando valori per percorsi più brevi, muovendosi verso il basso attraverso la struttura di ricerca in ampiezza.

SECONDA FASE: DETERMINARE LA QUANTITÀ DI FLUSSO

Può essere fatto partendo dagli strati più bassi della BFS,
dividendo il flusso che arriva dall'alto ad ogni nodo in proporzione al numero di cammini minimi che entrano nel nodo su ogni arco.



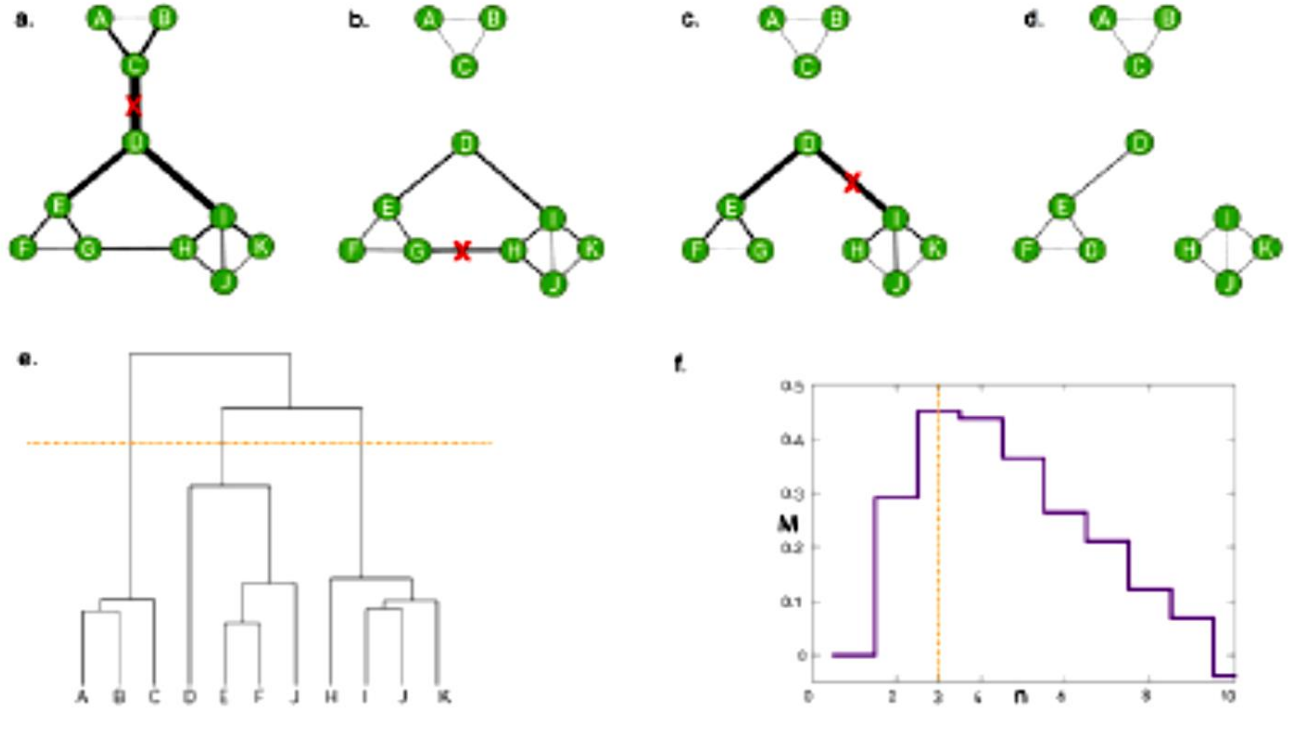
Complessità: L'algoritmo richiede di ricalcolare ad ogni passo la betweenness di tutti gli m archi. Quindi si hanno O(m) passi. Ad ogni passo occorre un tempo O(m) per computare la BFS e su di essa la betweeness, per ognuno degli N nodi . Di conseguenza si hanno in totale $O(m^2N)$ passi.

Quale partizione scegliere?

Poiché **l**'algoritmo lavora per partizionamenti successivi, in generale è necessario un criterio per stabilire quale livello di partizione è quello giusto.

Una possibilità è quella di valutare i vari livelli di partizionamento utilizzando una misura di bontà delle partizioni.

Es. Per la rete in figura (punto a.), applicando l'algoritmi di Girvan-Newman, si ottengono le seguenti partizioni successive. La modularità M (che vedremo in seguito) ci dice che la partizione migliore è al livello 3 (l'ultima riportata in figura).



Modularità. La modularità è una quantità scalare che misura la densità di archi all'interno delle comunità individuate.

Data una partizione dei vertici C = $\{C_1, . . . , C_t\}$ in cluster/comunità

la modularità Q (C) di C è definita come

Nota: C' varia su tutti i cluster → tutti edges sono contati due volte

$$Q(\mathcal{C}) = \sum_{C \in \mathcal{C}} \left[\frac{|E(C)|}{m} - \left(\frac{|E(C)| + \sum_{C' \in \mathcal{C}} |E(C, C')|}{2m} \right)^2 \right]$$

dove E (C, C ') indica l'insieme di edges tra vertici dei cluster C e C ' ed E (C) =E (C, C).

Riscrivendo in forma più conveniente, otteniamo

$$Q(\mathcal{C}) = \sum_{C \in \mathcal{C}} \left[\frac{|E(C)|}{m} - \left(\frac{\sum_{v \in C} \deg(v)}{2m} \right)^2 \right]$$

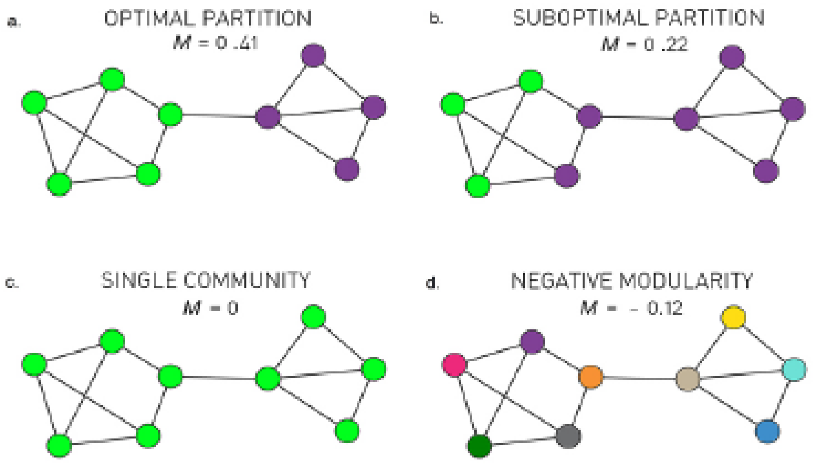
Notiamo che per massimizzare il primo termine, dovrebbero essere contenuti molti edge nei cluster, mentre la minimizzazione del secondo termine si ottiene dividendo il grafo in molti cluster di grado totale ''piccolo''.

Notiamo inoltre che

la Modularità prende valori tra -1/2 e 1

Il valore minimo Q(C) = - 1/2, lo assume quando tutti gli edge sono tra cluster
Ad esempio in un grafo bipartito con insieme dei nodi V=X U Y con 2 cluster corrispondenti a X e Y

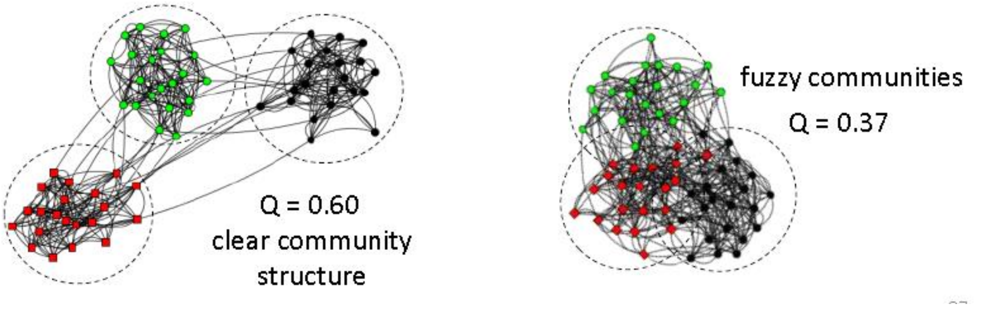
Il valore massimo Q(C) = 1, lo assume quando non vi sono edges tra cluster (ad esempio, tutti cluster formati da 2 nodi connessi da un edge) . In tal caso si può vedere che Q(C) → 1 al crescere del numero dei nodi.



La figura mostra la variazione della modularità al variare della partizione, in un grafo che ammette una chiara partizione in due comunità (fig. a.).

L'esperienza su reti reali mostra che

- valori >0.5 indicano la presenza effettiva di una struttura in comunità,
- valori prossimi allo zero indicano che la distribuzione degli archi tra intra- e inter-comunità non si discosta dalla casualità (in grafi ER il valore atteso è 0)



Ipotesi di lavoro. Un' ipotesi di lavoro, comunemente accettata (anche se non dimostrata formalmente) è che

la partizione ottimale in comunità è quella che massimizza la modularità.

Trovare la partizione corrispondente al valore massimo di modularità per un dato grafo è quindi la migliore partizione possibile per individuare le comunità.

Purtroppo si dimostra che

Teorema. Il problema della massimizzazione della modularità è NP-completo.

Sono state quindi sviluppate varie euristiche, ma anche queste spesso non riescono a far fronte in tempi ragionevoli alla dimensione attuale delle reti reali.

Completezza
Per formulare il nostro risultato consideriamo il seguente problema decisionale alla base della massimizzazione della modularità.

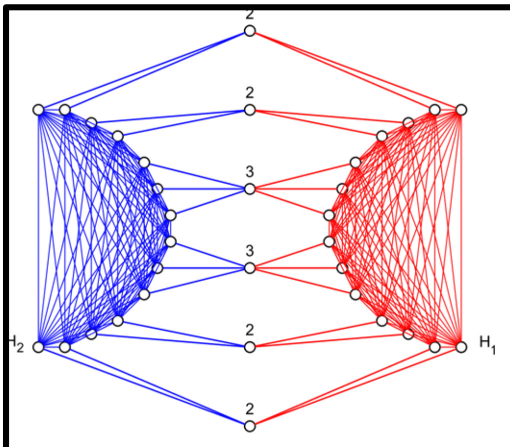
Problema (Modularità) *Dati un grafo G e un numero K, esiste una partizione C di G, tale che Q(C) ≥ K?*

Si noti che la massimizzazione della modularità non può essere più semplice del problema decisionale, poiché la determinazione della modularità massima di un grafo fornisce immediatamente una risposta alla domanda di decisione.

Da un'istanza A di 3-Partition, costruiamo un grafo G con k clique (sottografi completamente connessi) $H_1, \ldots, H_k$ di dimensione ciascuno $a=a_1+ \ldots + a_{3k}$. Per ogni $a_i \in A$ introduciamo un vertice singolo e lo colleghiamo ad a_i vertici in ciascuna delle k clique in modo tale che ciascun membro della clique sia collegato esattamente ad un vertice singolo.

È facile vedere che

- ogni vertice della clique ha il grado a
- il vertice dell'elemento corrispondente all'elemento $a_i \in A$ ha grado ka_i.
- Il numero di edge in G è $m=k(a(a-1)/2)+ak=a(a+1)k/2.$



Es. grafo G per l'istanza A = {2,2,2,2,3,3} di 3-Partition.

I colori degli edge indicano gli edge all'esterno e all'interno delle k = 2 clique H1 (rosso) e H2 (blu).

Le etichette dei vertici indicano i numeri corrispondenti ai ∈ A.

Lemma 1. In un partizionamento di modularità massima di G,

- nessuna delle clique è divisa.
- ogni cluster contiene al una delle cliques
- ogni vertice singolo appartiene esattamente ad un cluster

Dal lemma deduciamo che ogni cluster è composto da una clique e dei vertici singoli.

Quindi, il numero di edge tra cluster diversi *è*

$$a_1 (k-1)+ \ldots + a_{3k} (k-1))=a (k-1)$$

notiamo che è indipendente dai valori a_i

Da questo e dalla definizione di modularità, deduciamo quindi che il clustering ottimale per G deve minimizzare

$$\deg(C_1)^2 + \deg(C_2)^2 + \ldots + \deg(C_k)^2.$$

Notiamo che il minimo valore di $\deg(C_1)^2 + \deg(C_2)^2 + \ldots + \deg(C_k)^2$ si ottiene quando tutti gli addendi sono uguali

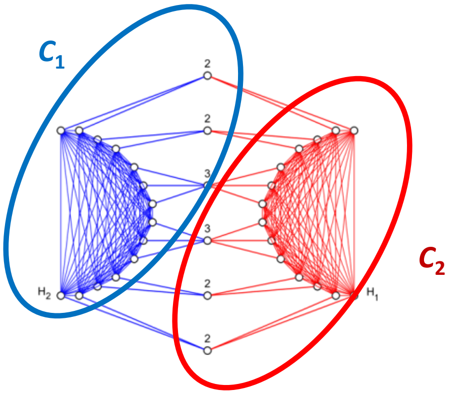
(infatti, In generale una somma $x1^2+\ldots+xk^2$ con $x1+\ldots+xk=S$ minima quando $x1=\ldots=xk=S/k$).

Ne consegue che la *somma* dei valori dei vertici singoli associati ad ogni clique per formare il cluster sono uguali, cioè pari a b (ricorda la somma dei valori associati ai vertici singoli è a=kb)

Cioè,

$\deg(C_1)^2 + \deg(C_2)^2 + \ldots + \deg(C_k)^2 \geq k(a(a-1)+2kb)^2= k(a(a-1)+2a)^2= k\,(a(a+1))^2$

con uguaglianza sse tutti gli addendi sono uguali.



Se 3-partition è soddisfacibile allora possiamo dividere gli a_i in k gruppi ognuno di somma b ed avere

$$\deg(C_1)^2 + \deg(C_2)^2 + \ldots + \deg(C_k)^2 = k\,(a(a+1))^2$$

cioè il valore minimo possibile. Ricordando

$$Q(\mathcal{C}) = \sum_{C\in\mathcal{C}} \left[\frac{|E(C)|}{m} - \left(\frac{\sum_{v\in C} \deg(v)}{2m} \right)^2 \right]$$

$$\sum_{C\in\mathcal{C}^*} \frac{|E(C)|}{m} = 1 - \frac{2k-2}{k(a+1)}$$

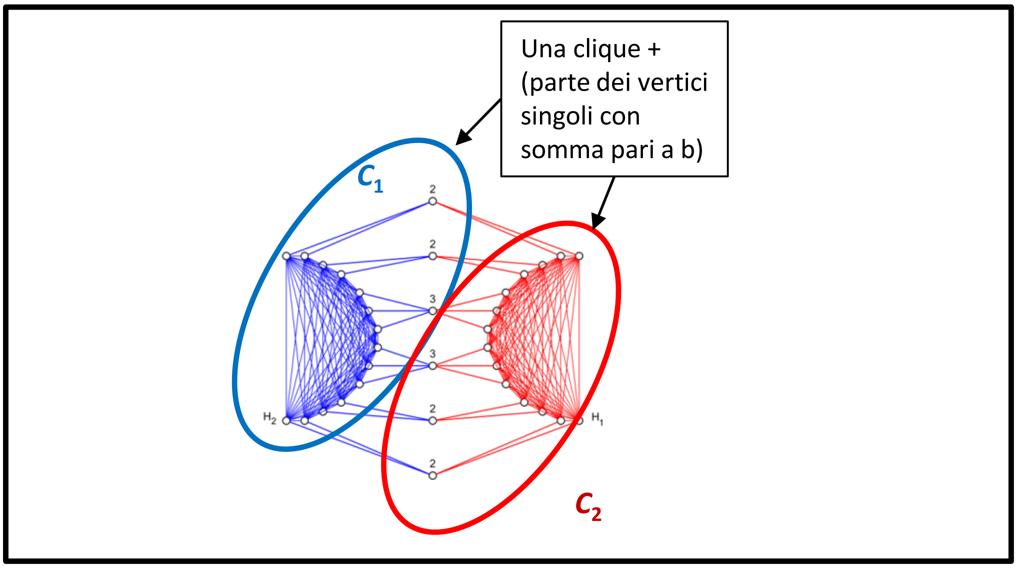
Il numero di edge in G è $m=k(a(a-1)/2)+ak=ka(a+1)/2$

Otteniamo che esiste partizionamento di G in cluster con Q(C) pari a K, infatti

$$1 - \frac{2k-2}{k(a+1)} - \frac{ka^2(a+1)^2}{k^2a^2(a+1)^2} = \frac{(k-1)(a-1)}{k(a+1)}$$

Supponiamo ora che esista una partizione in k cluster aventi modularità Q(C)≥K

Dove K $= \dfrac{(k-1)(a-1)}{k(a+1)}$



Sapendo che

$\deg(C_1)^2 + \deg(C_2)^2 + \ldots + \deg(C_k)^2 \geq k\,(a(a+1))^2$ *con uguaglianza sse tutti gli addendi sono uguali*

E ricordando
$$Q(\mathcal{C}) = \sum_{C\in\mathcal{C}} \left[\frac{|E(C)|}{m} - \left(\frac{\sum_{v\in C} \deg(v)}{2m} \right)^2 \right]$$

$$\sum_{C\in\mathcal{C}^*} \frac{|E(C)|}{m} = 1 - \frac{2k-2}{k(a+1)}$$

Il numero di edge in G è *m=k(a(a-1)/2)+ak=ka(a+1)/2*

Otteniamo che Q(C) ≤ $1 - \dfrac{2k-2}{k(a+1)} - \dfrac{ka^2(a+1)^2}{k^2a^2(a+1)^2} = \dfrac{(k-1)(a-1)}{k(a+1)}$

Con uguaglianza sse la somma dei vertici singoli in ogni cluster è la stessa

Quindi Q(C)≥K , implica Q(C)=K e la somma dei vertici singoli in ogni cluster è la stessa. Di conseguenza, la partizione in cluster fornisce una 3-partition (vertici singoli in ogni cluster). Quindi l'istanza di 3-Partition è soddisfacibile.

Questo completa la riduzione e dimostra il teorema.