Grafi e Comunità - Struttura Disomogenea parte 3

1 Partizionare un grafo in comunità: approccio euristico

Sono stati proposti numerosi metodi "euristici" per partizionare un grafo in comunità, dove con "comunità" si intende ora, genericamente, un insieme coeso di nodi. Perché, in genere, le definizioni combinatoriche che sono state proposte portano a problemi intrattabili - come abbiamo visto con le web-communities.

Per grandi linee, possiamo classificare le tecniche per il partizionamento di grafi in metodi partitivi (o divisivi) e metodi agglomerativi.

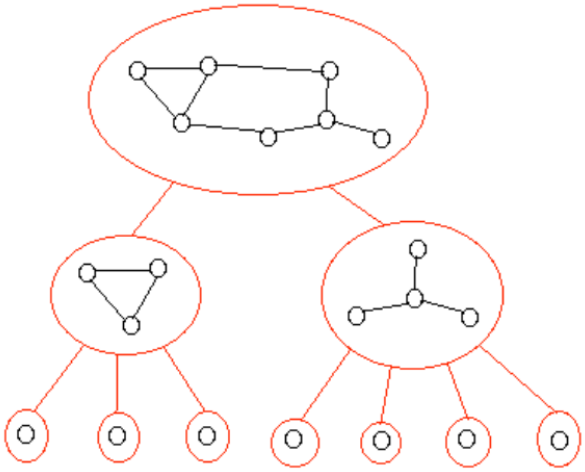
- In un metodo partitivo si inizia considerando l'intero grafo come un'unica grande comunità e poi, man mano, si rimuovono gli archi fino a quando il grafo risulta partizionato in componenti connesse; il procedimento viene poi iterato su ciascuna componente, fino a quando si ottiene un livello di granularità ritenuto adeguato o un insieme di comunità di dimensioni ritenute adeguate.
- In un metodo agglomerativo si inizia considerando ciascun nodo come una piccola comunità e poi, man mano, si aggiungono gli archi del grafo fino a quando si ottengono un numero di comunità ritenuto adeguato o comunità di dimensioni ritenute adeguate.

Partizionare un grafo in comunità: approccio euristico (segue)

Sia i metodi partitivi che quelli agglomerativi permettono di ottenere partizionamenti nidificati.

- In un metodo partitivo: ad ogni passo otteniamo comunità contenute in quelle ottenute al passo precedente.
- In un metodo agglomerativo: ad ogni passo otteniamo comunità che contengono quelle ottenute al passo precedente.

Così che otteniamo uno schema di partizionamento ad albero.



I diversi metodi partitivi / agglomerativi proposti si distinguono per il criterio utilizzato per scegliere ad ogni passo

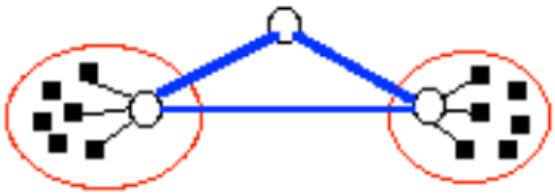
- quale arco del grafo rimuovere (in un metodo partitivo)
- quale arco del grafo aggiungere (in un metodo agglomerativo).

2 Betweenness di un arco

Vediamo ora un particolare criterio per rimuovere gli archi in un metodo partitivo: il criterio è basato sul concetto di betweenness di un arco, a sua volta basato sui concetti di bridge e local bridge:

- un bridge (per definizione) connette due regioni del grafo altrimenti non connesse
- un local bridge connette due regioni che, senza di esso, sarebbero connesse in modo meno efficiente.

Perciò, possiamo dire che bridges e local bridges connettono regioni che, senza di loro, avrebbero difficoltà ad interagire. Inoltre, abbiamo visto che bridges e local bridges sono weak ties e gli archi che rimangono dopo la loro rimozione sono gli strong ties - quelli delle relazioni forti. E da queste considerazioni nasce l'idea: rimuovendo bridges e local bridges il grafo viene partizionato in componenti che bene possono essere considerate comunità. Ma se la rete appare come in figura?



Nessun local bridge ma due regioni dense!

Betweenness di un arco (segue)

Proviamo a utilizzare una proprietà diversa da quella di (local) bridge, anche se, come il (local) bridge, vuole ancora descrivere una crescente difficoltà di collegamento indotta

dalla rimozione di un arco che la soddisfa. Essa è basata sulla nozione di traffico: gli archi che possiamo considerare i "nuovi ponti" sono quelli attraverso i quali passa più traffico. Ragionevole! E il traffico lo misuriamo con una sorta di flusso di un qualche fluido:

- per ogni coppia di nodi s e t assumiamo che s voglia inviare a t una unità di fluido
- che, viaggiando nella rete, per raggiungere t, si suddivide equamente fra tutti gli shortest paths che collegano s a t.

La betweenness di un arco è la quantità totale di fluido che lo attraversa, ottenuta sommando le frazioni di fluido per tutte le coppie $\langle s, t\rangle$. Un arco è tanto più un "nuovo ponte" quanto maggiore è la sua betweenness.

Betweenness di un arco (Definizione)

Formalmente: dato un grafo $G = (V, E)$ (non orientato).

- Per ogni coppia di nodi $s, t \in V$ e per ogni arco $(u, v) \in E$ definiamo: $\sigma_{st}(u, v) =$ numero di shortest paths fra s e t che attraversano (u, v).
- La betweenness relativa di (u, v) rispetto alla coppia $\langle s, t\rangle$, $b_{st}(u, v)$, è la frazione degli shortest paths fra s e t che attraversano (u, v):

  $$b_{st}(u, v) = \frac{\sigma_{st}(u, v)}{\sigma_{st}}$$

  dove σ_{st} è il numero totale di shortest paths fra s e t.
- Infine, la betweenness $b(u, v)$ di un arco $(u, v) \in E$ è la semi-somma delle betweenness relative ad ogni coppia di nodi:

  $$b(u, v) = \frac{1}{2} \sum_{s,t \in V} b_{st}(u, v)$$

NB: quella che abbiamo definito è la edge-betweenness; analogamente si può definire la node-betweenness.

3 Il metodo di Girvan-Newman

Il metodo di Girvan-Newman è un metodo partitivo basato sulla betweenness:

- si inizia considerando l'intero grafo come un'unica grande comunità
- poi si calcola l'arco di betweenness massima e si rimuove: se il grafo residuo è non connesso allora è stata ottenuta una prima partizione in comunità
- il procedimento viene poi iterato calcolando gli archi di betweenness massima nel grafo rimanente e rimuovendoli
- il procedimento ha termine quando si è ottenuto un livello di granularità ritenuto adeguato.

Sì, ma come si fa a calcolare le betweenness degli archi? Non possiamo mica enumerare tutti gli shortest paths di un grafo... (in generale, il loro numero è esponenziale nelle dimensioni del grafo!).

4 Calcolo delle betweenness degli archi

Lo schema dell'algoritmo per il calcolo delle betweenness degli archi è il seguente: per ogni $s \in V$ esegui i seguenti tre passi:

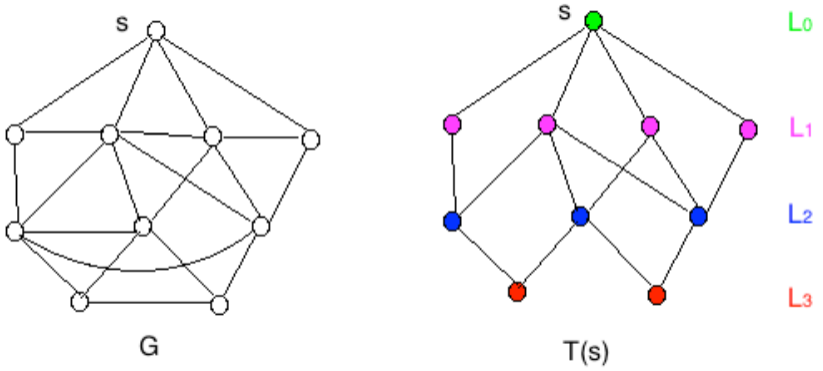
1. calcola il sottografo $T(s)$ degli shortest paths uscenti da s mediante una Breadth First Search (BFS)
2. mediante una visita top-down di $T(s)$, per ogni $v \in V$ calcola σ_{sv}
3. mediante una visita bottom-up di $T(s)$, e usando quanto calcolato al punto 2), per ogni $(u, v) \in T(s)$ calcola $b_s(u, v) = \sum_{t \in V - \{s\}} b_{st}(u, v)$

Per ogni $(u, v) \in E$ calcola $b(u, v) = \frac{1}{2} \sum_{s \in V} b_s(u, v)$.

Osservazione: al punto 3) calcoliamo $b_s(u, v)$ per i soli archi (u, v) in $T(s)$. Infatti, gli archi che non sono in $T(s)$ non fanno parte di alcuno shortest path uscente da s. E ora vediamo in dettaglio i singoli passi 1), 2) e 3).

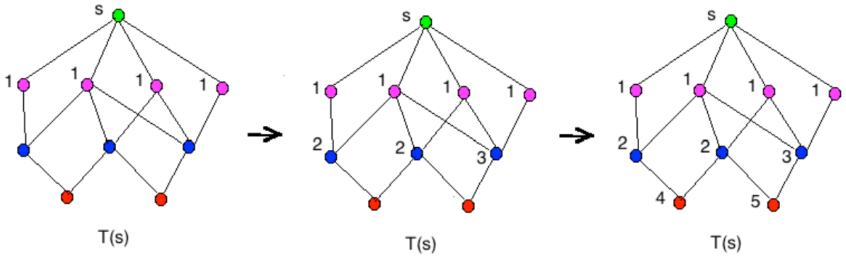
1) Breadth First Search da $s \in V$

Calcoliamo $T(s)$ come insieme di archi e, contemporaneamente, una partizione in livelli di V: $L_0 \leftarrow \{s\}$ e $T(s) \leftarrow \emptyset$. Per $h \geq 0$ e finché $L_h \neq \emptyset$ calcola: $L_{h+1} \leftarrow \{u \in V - \bigcup_{0 \leq i \leq h} L_i : \exists v \in L_h$ tale che $(v, u) \in E\}$ e $T(s) \leftarrow T(s) \cup \{(v, u) \in E : v \in L_h \wedge u \in L_{h+1}\}$.



2) Visita top-down di $T(s)$

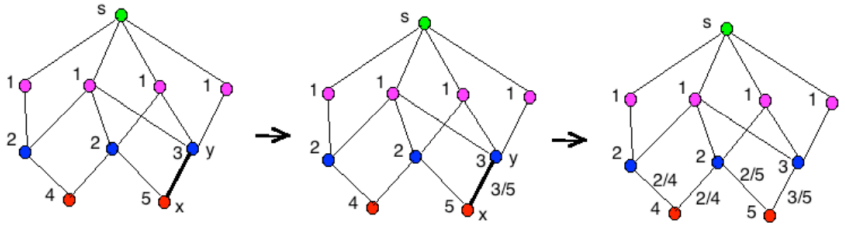
Mediante una visita top-down di $T(s)$, per ogni $v \in V$ calcola σ_{sv}. Il numero di shortest paths dalla radice s a un nodo $v \in L_h$ è pari alla somma dei numeri dei percorsi da s a qualunque "padre" di v. Ossia, dopo aver inizializzato $\sigma_{su} = 1$ per ogni $u \in L_1$, una volta calcolato σ_{su} per ogni $u \in L_{h-1}$, possiamo calcolare $\sigma_{sv} = \sum_{(u,v) \in E : u \in L_{h-1}} \sigma_{su}$, per ogni $v \in L_h$.



3) Visita bottom-up di $T(s)$

Per ogni $(u, v) \in T(s)$, calcola $b_s(u, v) = \sum_{t \in V - \{s\}} b_{st}(u, v)$. Sia d il numero di livelli di $T(s)$. Sia $(y, x) \in T(s)$ con $x \in L_d$ (una foglia): gli unici shortest paths uscenti da s che passano attraverso (y, x) sono gli shortest paths da s a x. E ciò implica anche che $b_{sz}(y, x) = 0$ per ogni $z \neq x$. E dunque, se $x \in L_d$, $b_s(y, x) = \sum_{z \in V} b_{sz}(y, x) = b_{sx}(y, x) = \frac{\sigma_{sx}(y,x)}{\sigma_{sx}} = \frac{\sigma_{sy}}{\sigma_{sx}}$.

$\sigma_{sx}(y, x) = \sigma_{sy}$ perché il numero di shortest paths da s a x che attraversano (y, x) è uguale al numero di shortest paths da s a y!



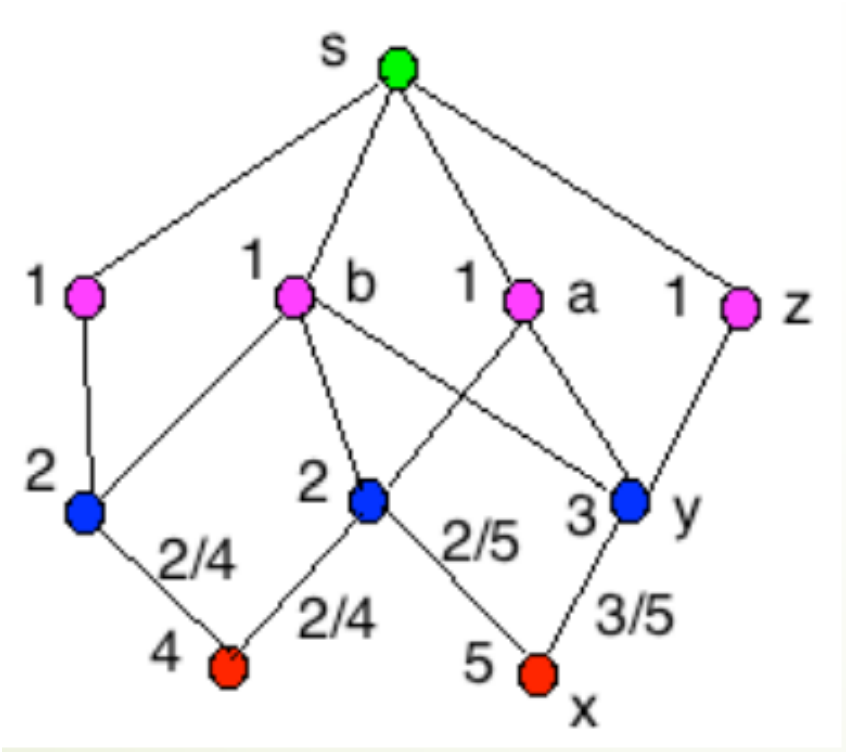
3) Visita bottom-up di $T(s)$ (segue)

Per ogni $(u, v) \in T(s)$, calcola $b_s(u, v) = \sum_{t \in V} b_{st}(u, v)$. Sia $(z, y) \in T(s)$ con $y \notin L_d$ (non foglia). Gli shortest path che passano attraverso (z, y) sono:

- alcuni shortest path da s a y - più precisamente, quelli che passano attraverso z (per lo stesso ragionamento fatto nel caso precedente)
- e, per ogni discendente x di y, alcuni shortest path da s a x - più precisamente, quelli che passano attraverso z e attraverso y.

Perciò, $b_s(z, y)$, ossia la frazione di tutti gli shortest path uscenti da s che passa attraverso l'arco (z, y), è la somma dei seguenti termini:

- $\frac{\sigma_{sz}}{\sigma_{sy}}$, ossia, la frazione degli shortest paths da s a y che passa attraverso (z, y)
- per ogni discendente x di y, una frazione $\frac{\sigma_{sz}}{\sigma_{sy}}$ della frazione di shortest paths da s a x che passano attraverso (y, x) (ossia, $b_s(y, x)$ se x è foglia?), ossia $\frac{\sigma_{sz}}{\sigma_{sy}} \times \frac{\sigma_{sy}}{\sigma_{sx}}$.



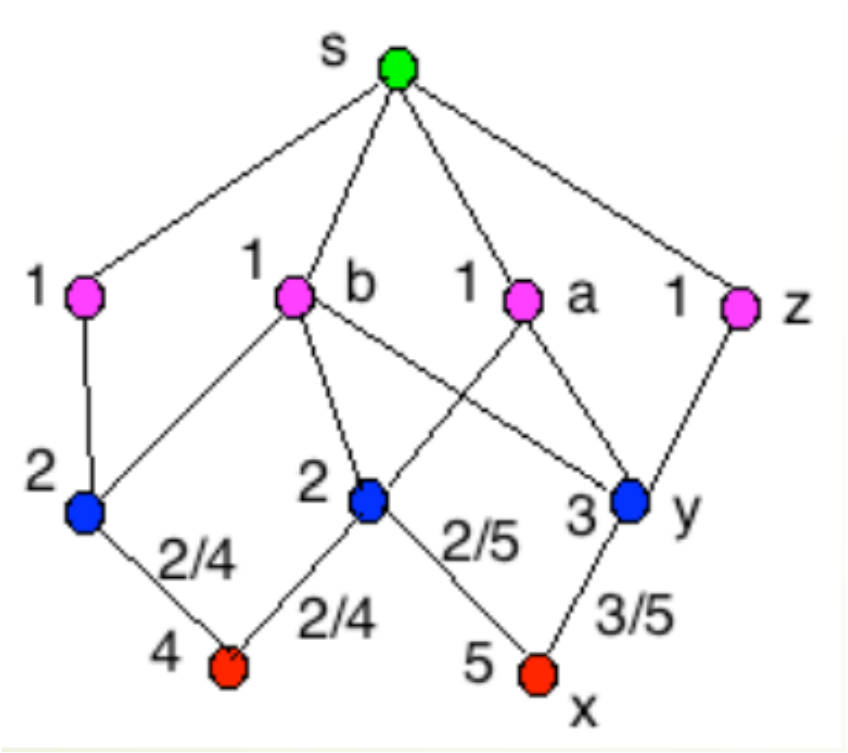
E adesso chiariamo con l'esempio...

3) Visita bottom-up di $T(s)$ (Esempio)

Per ogni $(u, v) \in T(s)$, calcola $b_s(u, v) = \sum_{t \in V} b_{st}(u, v)$. Sia $(z, y) \in T(s)$ con $y \notin L_d$. Attraverso (z, y) passano:

- una frazione pari a $\frac{\sigma_{sz}}{\sigma_{sy}}$ degli shortest paths da s a y
- e, per ogni discendente x di y, una frazione $\frac{\sigma_{sz}}{\sigma_{sy}}$ della frazione di shortest paths da s a x che passano attraverso (y, x).

Esempio: i 3/5 degli shortest paths da s a x passano per (y, x) (valore $b_s(y, x)$ calcolato prima, 3/5). Di questi 3/5, 1/3 passa per (z, y), 1/3 per (a, y) e 1/3 per (b, y) (poiché $\sigma_{sz} = 1, \sigma_{sa} = 1, \sigma_{sb} = 1$ e $\sigma_{sy} = 3$). Perciò, per (z, y) passano: 1/3 degli shortest paths da s a y e $1/3 \times (3/5)$ degli shortest paths da s a x. Dunque, $b_s(z, y) = \frac{1}{3} + \frac{1}{3} \times \frac{3}{5} = \frac{8}{15}$. Perché $b_{st}(z, y) = 0$ per ogni t che non è successore di y (o y stesso).



3) Visita bottom-up di $T(s)$ (Esempio 2)

Per ogni $(u, v) \in T(s)$, calcola $b_s(u, v) = \sum_{t \in V} b_{st}(u, v)$.

Esempio: consideriamo ora l'arco (a, c): (assumendo c sia il nodo con $\sigma_{sc} = 2$)

- una frazione pari a $\frac{\sigma_{sa}}{\sigma_{sc}} = \frac{1}{2}$ degli shortest paths da s a c passano attraverso (a, c).
- e, per ogni discendente di c, una frazione $\frac{\sigma_{sa}}{\sigma_{sc}} = \frac{1}{2}$ della frazione di shortest paths da s a quel discendente che passano attraverso l'arco "figlio" (es. (c, x)).

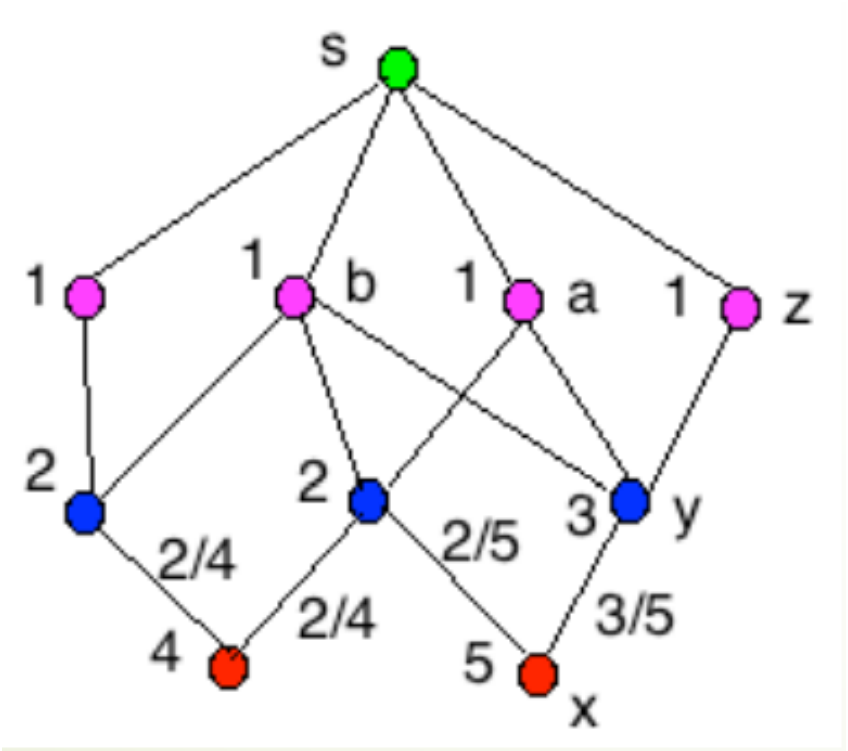
Nell'esempio, i discendenti di c sono x e w (nodi con $\sigma = 5$ e $\sigma = 4$).

- Dei 2/5 di shortest paths da s a x che passano per (c, x) [valore $b_s(c, x)$], 1/2 passano per (a, c) e 1/2 passano per (b, c).
- Dei 2/4 di shortest paths da s a w che passano per (c, w) [valore $b_s(c, w)$], 1/2 passano per (a, c) e 1/2 passano per (b, c).

Perciò, per (a, c) passano:

- 1/2 degli shortest paths da s a c,
- $1/2 \times (2/5)$ degli shortest paths da s a x,
- e $1/2 \times (2/4)$ degli shortest paths da s a w.

Dunque, $b_s(a, c) = \frac{1}{2} + \frac{1}{2} \times \frac{2}{5} + \frac{1}{2} \times \frac{2}{4} = \frac{1}{2} + \frac{1}{5} + \frac{1}{4} = \frac{10+4+5}{20} = \frac{19}{20}$.



3) Visita bottom-up di $T(s)$ (Formula)

Mediante una visita bottom-up di T, per ogni $(u, v) \in T(s)$ calcola $b_s(u, v) = \sum_{t \in V} b_{st}(u, v)$. Più in dettaglio: `for(h = d-1; h >= 0; h-)` do: calcola $b_s(u, v)$ per ogni $(u, v) \in T(s)$ tale che $u \in L_h$ e $v \in L_{h+1}$:

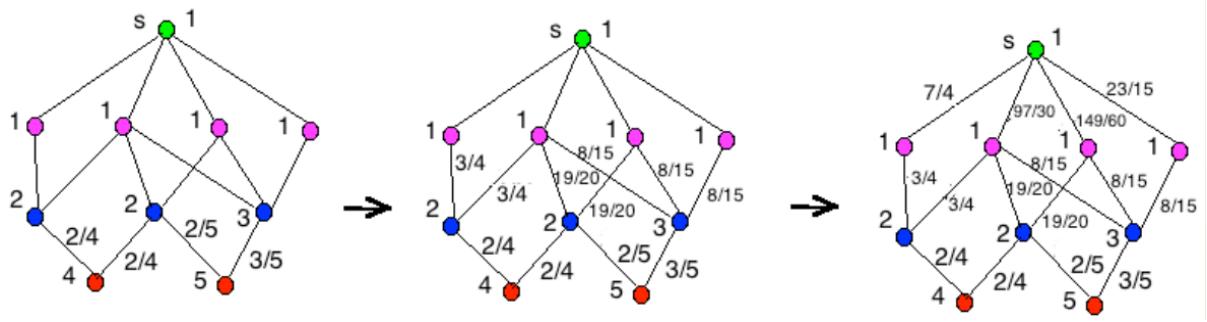
$$b_s(u, v) = \frac{\sigma_{su}}{\sigma_{sv}} + \sum_{(v,x) \in T(s)} \frac{\sigma_{su}}{\sigma_{sv}} b_s(v, x)$$

$$b_s(u, v) = \frac{\sigma_{su}}{\sigma_{sv}} \left(1 + \sum_{(v,x) \in T(s)} b_s(v, x) \right)$$

È una visita bottom-up: dai livelli più in basso a salire!

Osservazione 1. $T(s)$ è un insieme di archi orientati: se $v \in L_h$ e $(v, x) \in T(s)$ allora $x \in L_{h+1}$.

Osservazione 2. Abbiamo assunto $\sigma_{ss} = 1$ (vedi figura).



5 Rilassare il modello

Strong/weak ties, bridge/non bridge: queste distinzioni sembrano un po' troppo nette...

- tipicamente, fra le nostre relazioni, abbiamo amici più stretti di altri
- le nostre relazioni, cioè, si differenziano per "grado di confidenza"
- ma è poco probabile che, in generale, riusciamo a discernere in modo netto fra amici e conoscenti...

Allora, proviamo a rilassare il modello:

- Invece di considerare un grafo $G = (V, S \cup W)$ nel quale gli archi modellano relazioni forti e relazioni deboli.
- Consideriamo un grafo con archi pesati $G = (V, E, w)$, con $w : E \to \mathbb{N}$ nel quale il peso di un arco è tanto maggiore quanto più forte è la relazione che esso modella.

Rilassare il modello (segue)

Analogamente, invece di utilizzare i concetti di bridge e local bridge (che, in una rete reale, sono davvero pochi), consideriamo il concetto di **neighborhood overlap**: dato un arco (u, v),

$$NO(u, v) = \frac{|N(u) \cap N(v)|}{|N(u) \cup N(v) - \{u, v\}|}$$

Un local bridge è un arco (u, v) tale che $NO(u, v) = 0$.

Certo, è ragionevole rilassare il modello, ma nel modello dal quale siamo partiti si individuava una sorta di coincidenza fra archi deboli (proprietà locale) e local bridges (proprietà globali), ogni qualvolta il grafo soddisfa la STCP. Possiamo affermare che qualcosa di analogo valga nel modello rilassato? Ossia, che il neighborhood overlap di un arco è tanto più piccolo quanto minore è il suo peso?

Rilassare il modello (Evidenza)

[Onnela et al., 2007] ci hanno fornito evidenza sperimentale in favore di questa "coincidenza": nel loro esperimento:

- hanno considerato una rete, con archi pesati coerentemente con la "forza" delle corrispondenti relazioni (ad esempio, in una rete di connessioni telefoniche, il peso di un arco potrebbe essere proporzionale alla durata complessiva delle comunicazioni di due utenti)
- poi hanno iniziato a rimuovere gli archi a partire da quelli più pesanti e a seguire, via via, quelli sempre più leggeri
- $\Rightarrow$ e si sono accorti che la rete si disconnetteva molto lentamente.
- Poi hanno ripetuto l'esperimento "al contrario": hanno iniziato a rimuovere gli archi a partire da quelli più leggeri e a seguire, via via, quelli sempre più pesanti
- $\Rightarrow$ e si sono accorti che la rete si disconnetteva molto più velocemente!

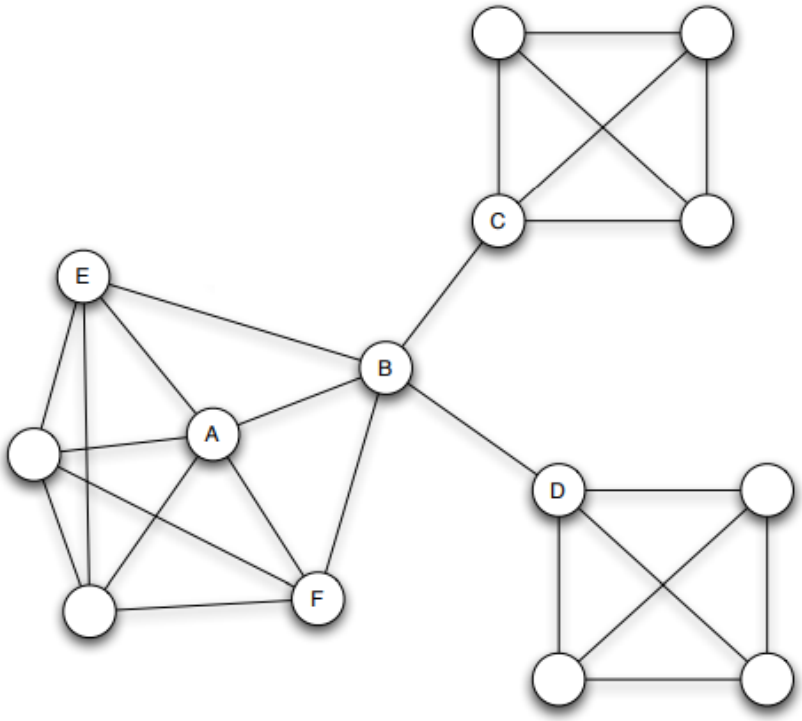
L'esperimento di Onnela mostra che la corrispondenza vale anche nel modello rilassato: poiché è ragionevole assumere che gli archi con neighborhood overlap più basso sono quelli la cui rimozione sconnette la rete, allora il neighborhood overlap di un arco è tanto più piccolo quanto minore è il suo peso.

6 Disomogeneità

La discussione condotta finora ci restituisce l'immagine di una rete costituita da agglomerati di nuclei fortemente connessi collegati da weak ties. In questa immagine gli archi non sono distribuiti uniformemente fra i nodi:

- alcuni nodi sono centrali all'interno dei gruppi coesi cui appartengono (come il nodo A in figura)
- altri nodi hanno posizioni periferiche, ossia, hanno un ruolo di interfaccia con altri gruppi (come il nodo B in figura).

Le posizioni di A e di B sono strutturalmente diverse all'interno della rete.

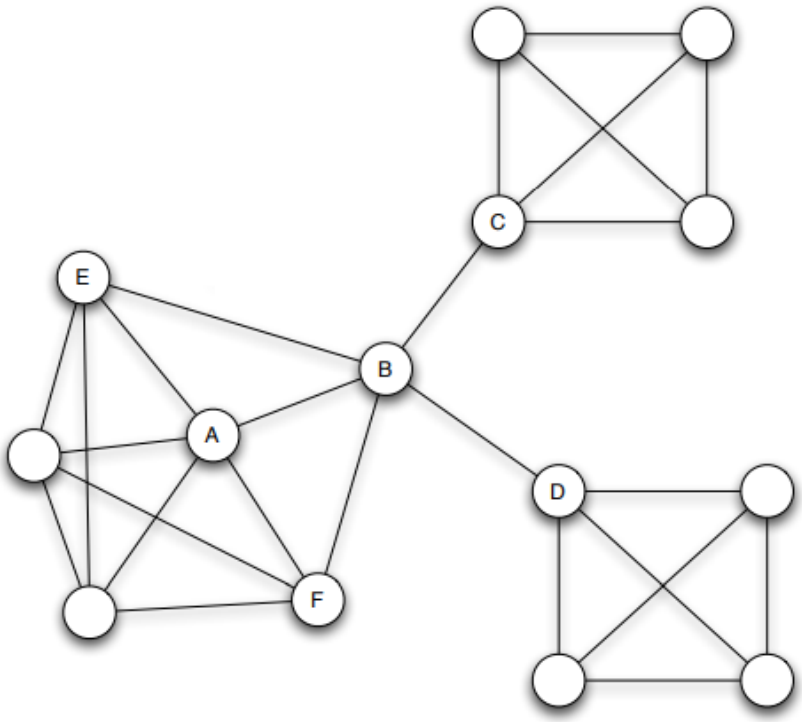


Disomogeneità (segue)

Le posizioni di A e di B sono strutturalmente diverse all'interno della rete. Definiamo l'**embeddedness** di un arco come il numero di vicini che gli estremi di quell'arco hanno in comune: $Emb(u, v) = |N(u) \cap N(v)|$. Se (u, v) è un local bridge allora $Emb(u, v) = 0$.

Tutti gli archi incidenti sul nodo A hanno embeddedness elevata. Possiamo interpretare questa caratteristica come una situazione in cui tutti i vicini di A si fidano di A e viceversa. Le cose per il nodo B sono parecchio diverse. Tuttavia, anche la posizione di B ha i suoi vantaggi: su B incidono numerosi local bridges; $\Rightarrow$ ossia, B si trova al centro di un

buco strutturale e questo, come sappiamo, gli permette di avere accesso a fonti di informazione inaccessibili ai nodi centrati nelle rispettive comunità.



Disomogeneità (Conclusione)

In conclusione: gli individui derivano benefici dalla struttura della rete nella quale sono immersi, benefici che dipendono dalla loro posizione all'interno della rete:

- benefici in termini di affidabilità del tessuto sociale
- benefici in termini di contenuti informativi.

Alejandro Portes definisce **capitale sociale** la "capacità degli attori/agenti di assicurarsi benefici in virtù del loro essere membri di una rete o struttura sociale".