The background is a complex, abstract collage of various mechanical and optical components. It features a dense arrangement of parts, including what appear to be camera lenses, viewfinders, and mechanical housings, rendered in a warm, golden-brown color palette. The components are layered and overlapping, creating a sense of depth and complexity. The overall effect is that of a technical or scientific illustration, possibly related to optics or photography.

Dispense di Analisi di Reti

Emanuele Izzo, 0307385

Corso di laurea magistrale Informatica
Università Tor Vergata, Facoltà di Scienze MM.FF.NN.
12/10/2021

Documento realizzato in L^AT_EX

Indice

1	Modelli generativi di grafi aleatori	2
1.1	Reti	2
1.1.1	Componenti giganti	2
1.2	Il modello di Erdős-Renyi	2
1.3	Il modello Power Law e il fenomeno Rich Get Richer	4
1.3.1	Funzione power law	5
1.3.2	Modello Power Law	5
1.3.3	Fenomeno Rich Get Richer	6
1.3.4	Lunga coda	10
1.4	Grafi Geometrici aleatori e reti wireless	10
1.4.1	Grafi geometrici	10
1.4.2	Grafi geometrici aleatori	10
1.4.3	Reti wireless e reti di sensori	11
1.4.4	Reti wireless e grafi geometrici aleatori	12
1.5	Il fenomeno Small World	20
2	Teoria dei grafi e delle reti sociali	21
2.1	L'esperimento di Granovetter	21
2.2	Bridges e Local Bridges	21
2.3	Legami forti e deboli	22
2.3.1	La chiusura triadica	22
2.4	I clusters (o comunità)	23
2.4.1	Le cut-communities	23
2.4.2	Le web-communities	24
2.5	Partizionare un grafo in due web-communities	25
3	Dinamiche nelle reti	29
4	Istituzioni e comportamento aggregato	30
4.1	Il gioco delle urne	30
4.2	Il teorema di Bayes	31
5	Reti di Informazione: il World Wide Web	36

Capitolo 1

Modelli generativi di grafi aleatori

1.1 Reti

L'analisi di reti si occupa dello studio delle reti nell'accezione del termine più ampia possibile, andandole a studiare da vari punti di vista (prestazioni, struttura, utilizzo, etc.) e utilizzando tecniche prese in prestito da numerose discipline. Genericamente parlando, una **rete** è uno schema di interconnessione fra un insieme di entità connesse, le quali dipendono dal tipo di entità che le compongono (si può parlare infatti di reti sociali, reti informatiche, reti fisiche, etc.). Quando si parla della connessione all'interno di una rete, ci si riferisce sia alle connessioni a livello di struttura, sia alle connessioni a livello di comportamento, ossia il fatto che ogni azione di un individuo ha conseguenze implicite nel risultato di tutti nella rete. Per rappresentare le reti, ovviamente, è possibile utilizzare un grafo dove i nodi sono le entità mentre gli archi sono le connessioni tra esse.

1.1.1 Componenti giganti

Parlando delle reti, è necessario definire cosa è una componente gigante. Una **componente gigante** in un grafo è una componente connessa che contiene una frazione del numero dei nodi (ossia contiene $\Theta(n)$ nodi).

1.2 Il modello di Erdős-Renyi

Il modello di Erdős-Renyi è un modello che permette la generazione di una rete probabilistica. Fissato $n \in \mathbb{N}$, $p \in [0, 1]$, costruiamo il grafo aleatorio $G_{n,p} = (V, E)$ nel seguente modo:

- $V = [n] = \{1, 2, \dots, n\}$.
- $\forall i, j \in [n], (i, j) \in E$ con probabilità p .

Osserviamo innanzi tutto che, fissato il numero n dei nodi, al variare di p si otterranno grafi con caratteristiche molto differenti:

- Quando $p = 0$, $E = \emptyset$ (grafo nullo).
- Quando $p = 1$, $|E| = 2^n$ (grafo completo).

Osservazione: Se $p > \frac{\ln 64}{n}$ allora con alta probabilità (con probabilità proporzionale almeno a $(1 - \frac{b}{n^c})$ per qualche coppia di costanti positive b e c), $G_{p,n}$ contiene una componente connessa costituita da almeno metà dei suoi nodi.

Teorema (1.1): Definiamo con X la variabile aleatoria corrispondente al numero di nodi nella più grande componente connessa di $G_{p,n}$, abbiamo che se $p > \frac{\ln 64}{n}$ allora $P(X \geq \frac{n}{2}) \geq 1 - 2^{-\frac{n}{8}}$.

Lemma (1.1a): Se $|X| < \frac{n}{2}$ allora esiste un insieme $A \subset [n]$ tale che $\frac{n}{4} < |A| < \frac{3n}{4}$ e non esistono archi tra i nodi in A e i nodi in $[n] - A$.

Lemma (1.1a): Siano C_1, C_2, \dots, C_k le componenti connesse di $G_{p,n}$ dove $|C_1| \leq |C_2| \leq \dots \leq |C_k|$. Poiché $X < \frac{n}{2}$ allora $\forall i \in [1, k], |C_i| < \frac{n}{2}$. Scegliamo $h \in [1, k]$ tale che:

- $|C_1| + |C_2| + \dots + |C_{h-1}| < \frac{n}{4}$.
- $|C_1| + |C_2| + \dots + |C_h| > \frac{n}{4}$.

Abbiamo inoltre che $h < k$, infatti poiché $|C_i| < \frac{n}{2}$, se fosse $h = k$, avremmo $|C_1| + |C_2| + \dots + |C_k| < \frac{n}{4} + \frac{n}{2} < n$. Scegliamo pertanto $A = C_1 \cup C_2 \cup \dots \cup C_h$, da cui abbiamo che $A \neq \emptyset$ e $[n] - A \neq \emptyset$. Sappiamo che per costruzione $|A| > \frac{n}{4}$, inoltre possiamo dire che $|A| = |C_1| + |C_2| + \dots + |C_h| < \frac{n}{4} + \frac{n}{2} = \frac{3n}{4}$; inoltre, sempre per costruzione, poiché C_1, C_2, \dots, C_k sono tutte componenti connesse, non esiste alcun arco che connette A ad $[n] - A$. Questo insieme A viene detto buono.

Teorema (1.1): Definiamo con A^* un insieme buono, e definiamo con E l'evento "non esiste alcun arco tra A ed $[n] - A$ ". Calcoliamo ora $P(X < \frac{n}{2})$:

$$\begin{aligned}
 P(X < \frac{n}{2}) &= P(\exists A^* \subset [n]) = \\
 &= P(\cup_{A \subset [n], \frac{n}{2} < |A^*| < \frac{3n}{4}} \emptyset) = \\
 &= \sum_{A \subset [n], \frac{n}{2} < |A^*| < \frac{3n}{4}} P() \leq \\
 &\leq \sum_{A \subset [n], \frac{n}{2} < |A^*| < \frac{3n}{4}} (1 - p)^{\frac{3n^2}{16}} \leq \\
 &\leq 2^n (1 - p)^{\frac{3n^2}{16}} < \\
 &< 2^n (1 - \frac{\ln 64}{n})^{\frac{3n^2}{16}} = \\
 &= 2^n [(1 - \frac{\ln 64}{n})^n]^{\frac{3n}{16}}
 \end{aligned}$$

Ricordando che $\lim_{n \rightarrow \infty} (1 - \frac{1}{n})^{-n} = e$, e che quindi $\lim_{n \rightarrow \infty} (1 - \frac{\ln 64}{n})^{-\frac{n}{\ln 64}} = e$, abbiamo:

$$\begin{aligned}
P(X < \frac{n}{2}) &< P(\exists A^* \subset [n]) = \\
&= 2^n [(1 - \frac{\ln 64}{n})^n]^{\frac{3n}{16}} = \\
&= 2^n [(1 - \frac{\ln 64}{n})^{(-\frac{n}{\ln 64})(-\ln 64)}]^{\frac{3n}{16}} \cong \\
&\cong 2^n [e^{-\ln 64}]^{\frac{3n}{16}} = \\
&= 2^n [64^{-1}]^{\frac{3n}{16}} = 2^n [2^{-\log_2 64}]^{\frac{3n}{16}} = \\
&= 2^n [2^{-6}]^{\frac{3n}{16}} = 2^n 2^{-\frac{18n}{16}} = \\
&= 2^{\frac{16n-18n}{16}} = 2^{-\frac{2n}{16}} = 2^{-\frac{n}{8}}
\end{aligned}$$

Osservazione: Nel grafo $G_{n,p}$ gli archi vengono aggiunti come una sequenza di oggetti, indipendentemente gli uni dagli altri. Perciò, il grado di un nodo è la somma di eventi indipendenti: sia E_{ij} una variabile aleatoria che vale 1 se esiste l'arco (i, j) o 0 altrimenti, il grado del nodo i sarà allora uguale a $\sum_{j \in [n]-1} E_{ij}$. La variabile E_{ij} , per come è definita, segue il teorema del limite centrale:

$$\begin{aligned}
E[E_{ij}] &= 1 \cdot p + 0 \cdot (1 - p) = p = \mu \\
Var[E_{ij}] &= (1 - p)^2 p + (0 - p)^2 (1 - p) = \\
&= (1 - p)^2 p + (1 - p) p^2 = p(1 - p)(1 - p + p) = p(1 - p) = \sigma^2
\end{aligned}$$

Definita la variabile aleatoria $Y_n = \frac{\frac{1}{n} \sum_{j=1}^n E_{ij} - \mu}{\sigma/n}$, allora Y_n presenta una distribuzione normale standard $Y_n \sim N(0, 1)$.

1.3 Il modello Power Law e il fenomeno Rich Get Richer

In una ricerca condotta nel 2000 da Broder sulla struttura del web ha portato alla luce dettagli interessanti; se paragoniamo il web ad un grafo diretto, dove i nodi sono le pagine e gli archi sono gli hyperlink che collegano due pagine (un arco (a, b) indica che esiste un hyperlink nella pagina a che porta alla pagina b), possiamo notare che esso differisce sostanzialmente dal teorema del limite centrale da quanto indicato nel limite centrale: la frazione di pagine web che ha grado entrante k è proporzionale a $1/k^c$ (per una qualche costante c) piuttosto che a $1/k^k$, ossia il numero di nodi del grafo del web che hanno grado entrante elevato è molto maggiore di quello che ci si aspetterebbe assumendo che gli archi si formino indipendentemente gli uni dagli altri.

1.3.1 Funzione power law

Una funzione **Power Law** è una relazione funzionale tra due quantità, dove una varia come la potenza dell'altra (una variazione su una quantità porta ad una variazione quadrica sull'altra). Una Power Law è esprimibile nella forma $f(x) = ax^k + o(x^k)$, dove k viene detto *esponente di scala*; mentre una distribuzione di una power law segue la forma $p(x) = g(x)x^{-k}$, dove $k > 1$ è detto *esponente della power law* e $f(x)$ è una funzione "che varia lentamente"¹, e nel caso in cui $f(x)$ abbia valore costante c allora la distribuzione diventa $p(x) = cx^{-k}$. Di solito le distribuzioni Power Law sono rappresentate su un grafico log-log, cioè un grafico in cui entrambe le variabili sugli assi sono misurate in logaritmi. Trasformando in logaritmi si ha infatti $\log p(x) = \log c - k \log x$, portando la relazione a diventare lineare.

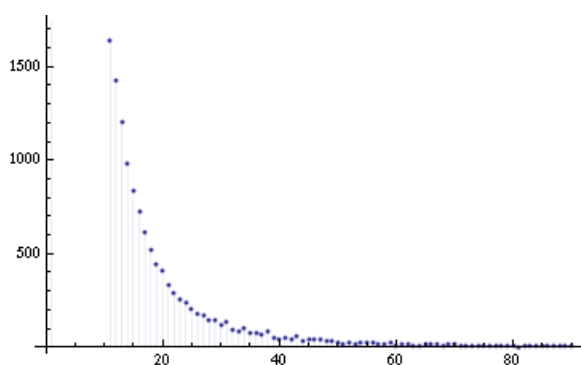


Figura 1.1: Istogramma di un campione estratto da una distribuzione Power Law con esponente 3.

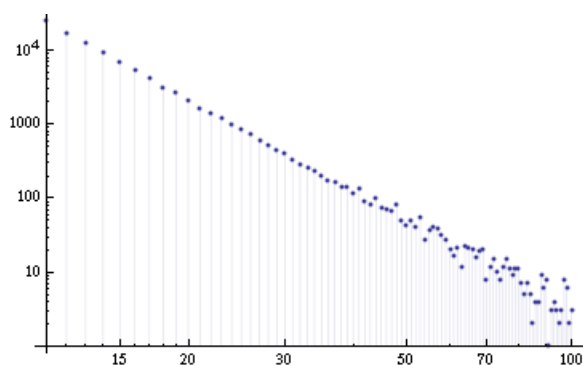


Figura 1.2: Istogramma di un campione estratto da una distribuzione Power Law con esponente 3 in doppia scala logaritmica (log-log plot).

1.3.2 Modello Power Law

Tornando all'esempio delle pagine web, nel caso avessimo il bisogno di collegarci ad una qualsiasi pagina riguardo un certo argomento, andremmo a scegliere una pagina autorevole, e che quindi viene referenziata da molte pagine; se consideriamo il grafo equivalente, l'aggiunta di un arco dipende dagli archi già presenti. Da qui già possiamo notare 3 aspetti importanti:

- Le pagine vengono create una alla volta, ossia in sequenza.
- Quando viene creata una pagina, viene decisa a quale altra pagina debba puntare.
- Ciascun puntatore è un arco diretto, pertanto il grafo risultante darà diretto.

Da qui possiamo definire il modello power law. Fissato $p \in [0, 1]$, costruiamo il grafo diretto aleatorio $G_{n,p} = (V, E)$ in cui i nodi vengono inseriti sequenzialmente in time-step discreti:

- $t_1 \rightarrow$ Viene creato il nodo 1.
- $t_2 \rightarrow$ Viene creato il nodo 2 e l'arco $(2, 1)$.

¹ossia che $\lim_{x \rightarrow \infty} \frac{f(tx)}{f(x)} = 1$ con t costante.

- $\forall t_i > t_2 \rightarrow$ Viene creato il nodo i e un arco uscente da i secondo la seguente regola:
 - Viene scelto uniformemente a caso un nodo $j < i$.
 - Con probabilità p viene creato l'arco (i, j) .
 - Con probabilità $1 - p$ viene creato l'arco (i, k) , dove (j, k) è l'arco uscente da j .

È possibile esprimere la creazione del grafo usando una diversa descrizione degli step. Fissato $p \in [0, 1]$, costruiamo il grafo diretto aleatorio $G_{n,p} = (V, E)$ in cui i nodi vengono inseriti sequenzialmente in time-step discreti:

- $t_1 \rightarrow$ Viene creato il nodo 1.
- $t_2 \rightarrow$ Viene creato il nodo 2 e l'arco $(2, 1)$.
- $\forall t_i > t_2 \rightarrow$ Viene creato il nodo i e un arco uscente da i secondo la seguente regola:
 - Con probabilità p viene scelto uniformemente a caso un nodo $j < i$ e viene creato l'arco (i, j) .
 - Con probabilità $1 - p$ viene scelto uniformemente a caso un nodo $j < i$ e, detto (j, k) l'arco uscente da j , viene creato l'arco (i, k) .

È possibile verificare che i grafi generati secondo questa regola seguono una Power Law. Definiamo per ogni coppia di interi i e j tali che $i > j$ e $i \geq 2$ la variabile aleatoria

$$d_{ij} = \begin{cases} 1 & \text{se } (i, j) \in E \\ 0 & \text{altrimenti} \end{cases}$$

Poniamo con E_i l'evento "viene scelto il nodo i ", allora la probabilità che esista l'arco (i, j) è uguale a:

$$\begin{aligned} P(d_{ij} = 1) &= p \cdot P(E_j) + (1 - p) \cdot P(E_k | d_{kj} = 1) = \\ &= \frac{p}{i-1} + (1 - p) P(\cup_{k < i: (k,j) \in E} E_k) = \\ &= \frac{p}{i-1} + (1 - p) \sum_{1 \leq k < i: (k,j) \in E} P(E_k) = \\ &= \frac{p}{i-1} + (1 - p) \sum_{1 \leq k < i: (k,j) \in E} 1 = \\ &= \frac{p}{i-1} + (1 - p) \sum_{1 \leq k < i} d_{kj} \end{aligned}$$

1.3.3 Fenomeno Rich Get Richer

Intuitivamente possiamo notare che man mano aumenta il numero di archi entranti in un nodo, cresce la probabilità che quel nodo venga selezionato come estremo di un arco uscente da un nodo appena creato. Questo fenomeno prende il nome di **Rich Get Richer**, detto anche *effetto San Matteo*, con il quale si indica un processo per cui le nuove risorse che si rendono disponibili vengono ripartite fra i partecipanti in proporzione a quanto hanno già. Un dettaglio importante di questo fenomeno è che è pesantemente

suscettibile alle condizioni iniziali, e che quindi diverse situazioni di partenza produrranno grafi estremamente diversi tra loro.

Teorema (1.2): La probabilità di scegliere il nodo j a cui connettere il nodo i è uniforme con probabilità p e proporzionale al grado di j con probabilità $1 - p$.

$$P((i, j) \in E) = P(d_{ij} = 1) = \frac{p}{i-1} + (1-p) \sum_{1 \leq k < i} d_{kj}$$

In tale modello, la frazione del numero di nodi con k è con alta probabilità proporzionale a $k^{-\frac{1}{1-p}-1}$.

Osservazione: In primis ha senso calcolare $P((i, j) \in E)$ per $i > 1$ essendo che 1 non ha archi uscenti. Inoltre, ogni nodo $i > 1$ ha esattamente un arco uscente, e pertanto si deve avere che $P(\exists j < i : (i, j) \in E) = 1$; questo si può dimostrare per induzione, infatti:

- Per $i = 2$ abbiamo che $\sum_{1 \leq j < 2} P((2, j) \in E) = P((2, 1) \in E) = 1$ per costruzione.
- Assumendo $\sum_{1 \leq j < a} P((2, j) \in E)$ per ogni $a \leq i - 1$ si ha che

$$\begin{aligned} \sum_{1 \leq j < i} P((i, j) \in E) &= \sum_{1 \leq j < i} \left[\frac{p}{i-1} + \frac{1-p}{i-1} \sum_{1 \leq k < i} d_{kj} \right] = \\ &= \sum_{1 \leq j < i} \frac{p}{i-1} + \sum_{1 \leq j < i} \left(\frac{1-p}{i-1} \sum_{1 \leq k < i} d_{kj} \right) = \\ &= \frac{(i-1)p}{i-1} + \sum_{1 \leq k < i} \left(\frac{1-p}{i-1} \sum_{1 \leq j < i} d_{kj} \right) = \\ &= p + \sum_{1 \leq k < i} \frac{1-p}{i-1} = p + \frac{(i-1)(1-p)}{i-1} = p + 1 - p = 1 \end{aligned}$$

Teorema (1.2): Quello che si vuole dimostrare è che il modello descritto sopra segue una Power Law. Sia $D_j(t)$ la variabile aleatoria che esprime il numero di archi entranti nel nodo j al passo t di generazione del grafo; naturalmente, la variabile $D_j(t)$ è definita per $t \geq j$ e per ogni $j \geq 1$. Al passo $t = j$ il grado entrante di j è 0, ossia $D_j(j) = 0$. Al passo $t + 1$ il grado di j può essere invariato rispetto al passo t oppure può essere aumentato di un'unità, ossia $D_j(t + 1) = D_j(t) + 1$ se e solo se è stato creato l'arco $(t + 1, j)$. Quindi

$$P(D_j(t + 1) - D_j(t) = 1) = P(d_{t+1,j} = 1) = \frac{p}{t} + \frac{1-p}{t} \sum_{1 \leq h < t+1} d_{hj} = \frac{p}{t} + \frac{1-p}{t} D_j(t)$$

Per ogni $j \geq 1$ definiamo una funzione discreta deterministica $X_j(t)$ che "assomiglia" a $D_j(t)$:

$$\begin{cases} X_j(j) = 0 \\ X_j(t+1) - X_j(t) = \frac{p}{t} + \frac{1-p}{t} X_j(t) \end{cases}$$

In questo modo possiamo approssimare il comportamento di $X_j(j)$ con una funzione $x_j(j)$ definita su un dominio continuo per $t \geq j$:

$$\begin{cases} x_j(j) = 0 \\ x'_j(t) = \frac{p}{t} + \frac{1-p}{t} x_j(t) \end{cases}$$

Quello che abbiamo davanti è un problema di Cauchy, il quale possiamo risolvere per ottenere il valore di $x_j(t)$:

$$\begin{aligned} x'_j(t) &= \frac{p}{t} + \frac{1-p}{t} x_j(t) \\ x'_j(t) &= \frac{1}{t} [p + (1-p)x_j(t)] \\ \frac{x'_j(t)}{p + (1-p)x_j(t)} &= \frac{1}{t} \\ \int \frac{x'_j(t)}{p + (1-p)x_j(t)} dt &= \int \frac{1}{t} dt \\ \int \frac{(1-p)x'_j(t)}{p + (1-p)x_j(t)} dt &= (1-p) \int \frac{1}{t} dt \end{aligned}$$

Se poniamo $f(t) = p + (1-p)x_j(t)$ e $f'(t) = (p + (1-p)x_j(t))' = (1-p)x'_j(t)$, l'integrale di sinistra diventa $\int \frac{f'(t)}{f(t)} dt = \ln f(t)$, pertanto:

$$\begin{aligned} \int \frac{(1-p)x'_j(t)}{p + (1-p)x_j(t)} dt &= (1-p) \int \frac{1}{t} dt \\ \ln[p + (1-p)x_j(t)] + c &= (1-p) \ln t + c \\ \ln[p + (1-p)x_j(t)] &= \ln t^{1-p} + c \\ e^{\ln[p + (1-p)x_j(t)]} &= e^{\ln t^{1-p} + c} \\ p + (1-p)x_j(t) &= t^{1-p} e^c \end{aligned}$$

Se poniamo $x_j(j) = 0$ otteniamo

$$\begin{aligned} p + (1-p)x_j(j) &= j^{1-p} e^c \\ p &= j^{1-p} e^c \\ e^c &= \frac{p}{j^{1-p}} \end{aligned}$$

Ossia $x_j(t) = \frac{p}{1-p} \left[\left(\frac{t}{j} \right)^{1-p} - 1 \right]$. Definiamo a questo punto l'insieme $A_t(k) := \{j \leq t : x_j(t) \geq k\}$, nel quale per definizione $j \in A_t(k)$ se e solo se $j \leq t$ e $x_j(t) \geq k$.

Calcoliamo $x_j(t) \geq k$:

$$\begin{aligned}
x_j(t) &\geq k \\
\frac{p}{1-p} \left[\left(\frac{t}{j} \right)^{1-p} - 1 \right] &\geq k \\
\left(\frac{t}{j} \right)^{1-p} &\geq \frac{1-p}{p} k + 1 \\
\frac{t}{j} &\geq \left(\frac{1-p}{p} k + 1 \right)^{1-p} \\
j &\leq t \left(\frac{1-p}{p} k + 1 \right)^{1-p}
\end{aligned}$$

Arrivati a questo punto possiamo notare una cosa interessante:

$$\begin{cases} \frac{1-p}{p} k + 1 \geq 1 \\ \frac{1}{1-p} < 0 \end{cases} \rightarrow \left(\frac{1-p}{p} k + 1 \right)^{1-p} < 1 \rightarrow t \left(\frac{1-p}{p} k + 1 \right)^{1-p} \leq t$$

Dato ciò possiamo dire che $j \in A_t(k)$ se e solo se $j \leq t \left(\frac{1-p}{p} k + 1 \right)^{1-p}$, e quindi $A_t(k) := \{j \leq t \left(\frac{1-p}{p} k + 1 \right)^{1-p}\}$ con $|A_t(k)| := t \left(\frac{1-p}{p} k + 1 \right)^{1-p}$. Arrivati qui calcoliamo, dati k e t , quale frazione dei nodi che al passo t presentano k archi entranti, ossia $\frac{1}{t} |A_t(k) - A_t(k+1)|$. Poiché $A_t(k+1) \subseteq A_t(k)$, e ponendo $F(k) = \left(\frac{1-p}{p} k + 1 \right)^{1-p}$, otteniamo:

$$\begin{aligned}
&\frac{1}{t} |A_t(k) - A_t(k+1)| = \\
&= \frac{1}{t} (|A_t(k)| - |A_t(k+1)|) \\
&= \frac{1}{t} \left[t \left(\frac{1-p}{p} k + 1 \right)^{1-p} - t \left(\frac{1-p}{p} (k+1) + 1 \right)^{1-p} \right] = \\
&= \left(\frac{1-p}{p} k + 1 \right)^{1-p} - \left(\frac{1-p}{p} (k+1) + 1 \right)^{1-p} = \\
&= F(k) - F(k+1) = - \left[- \frac{F(k) - F(k+1)}{1} \right] = \\
&= - \left[\frac{F(k+1) - F(k)}{1} \right] \cong - \frac{d}{dk} F(k) = \\
&= \frac{d}{dk} \left[- \left(\frac{1-p}{p} k + 1 \right)^{1-p} \right] = \\
&= - \frac{1-p}{p} \left(- \frac{1}{1-p} \right) \left(\frac{1-p}{p} k + 1 \right)^{1-p} = \\
&= \frac{1}{p} \left(\frac{1-p}{p} k + 1 \right)^{-\frac{1}{1-p}-1}
\end{aligned}$$

Pertanto, fissati t e k , la frazione dei nodi che al passo t presentano k archi entranti è $\frac{1}{p} \left(\frac{1-p}{p} k + 1 \right)^{-\frac{1}{1-p}-1}$, la quale è una Power Law con esponente $-\frac{1}{1-p} - 1$.

1.3.4 Lunga coda

Poniamo di esaminare un negozio on-line che si occupa di vendere un'enorme quantità di libri. Quello che ci chiediamo è se questo sito guadagna di più vendendo tante copie di pochi libri molto popolari o vendendo poche copie di tanti libri poco popolari; in altre parole ci stiamo interessando di questioni di popolarità sotto un altro punto di vista, ossia se complessivamente il volume di affari derivante dai nodi poco popolari (con grado basso) equivale al volume di affari dei nodi molto popolari (con grado alto). Per esaminare ciò possiamo realizzare un grafo dove ogni punto indica il numero di libri con uguale probabilità (ossia il numero di nodi con uguale grado); per fare ciò dobbiamo presupporre che non esistono due gruppi di nodi di ugual numero con diverso grado. Notiamo che se $n = d(k)$ è la funzione che esprime il numero di nodi di grado k , allora il grafico realizzato rappresenta grosso modo la funzione $k = d^{-1}(n)$ (ammesso che la funzione sia invertibile). È possibile dimostrare che la funzione $d(k)$ è una Power Law, ossia per qualche costante $c > 0$, $d(k) \cong k^{-c}$, e dunque che, se $n = d(k) \cong k^{-c}$, abbiamo che $k = d^{-1}(n) \cong n^{-\frac{1}{c}}$. Possiamo notare che questa funzione decresce molto lentamente, ossia forma una **lunga coda**. Se fissiamo un j abbastanza grande, l'area della regione compresa fra la coda del grafico e l'asse delle ascisse è tutt'altro che trascurabile.

1.4 Grafi Geometrici aleatori e reti wireless

1.4.1 Grafi geometrici

Sia $d \geq 1$ e $r > 0$ due valori fissati, viene definita **grafo geometrico** un grafo $G_{d,r} := (V, E)$ definito in tale modo:

- $V \subseteq \mathbb{R}^d$.
- $E := \{(x, y) : x \in V \wedge y \in V \wedge \sqrt{\sum_{k=1}^d (x_k - y_k)^2} \leq r\}$

Generalmente il grafo $G_{d,r}$ viene normalizzato rispetto ad r , riportando i punti in scala $1 : r$ cosicché due punti sono adiacenti se e solo se la loro distanza è minore di 1. In questo caso, ossia quando $r = 1$, il grafo $G_d := (V, E)$ prende il nome di **Unit Disk Graph**, ed è definito in tale modo:

- $V := \{x' = \frac{x}{r} : x \in \mathbb{R}^d\}$.
- $E := \{(x', y') : x' \in V \wedge y' \in V \wedge \sqrt{\sum_{k=1}^d (x'_k - y'_k)^2} \leq 1\}$

1.4.2 Grafi geometrici aleatori

Fissiamo $n \in \mathbb{N}$ e $r > 0$, scegliamo uniformemente a caso n punti nel quadrato $[0, 1] \times [0, 1]$, e costruiamo un grafo geometrico $G_{n,r}$ corrispondente. Possiamo notare subito che, essendo il quadrato di lato pari ad 1, e pertanto la diagonale è pari a $\sqrt{2}$, è sufficiente scegliere $r \leq \sqrt{2}$; infatti, con $r = \sqrt{2}$ otteniamo un grafo completo ed è dunque inutile scegliere per r un valore maggiore di $\sqrt{2}$. Di contro, scegliendo r molto vicino a 0 otteniamo, più o meno, un grafo privo di archi. Naturalmente, la aleatorietà del grafo $G_{n,r}$ dipende dalla scelta dei punti nel quadrato unitario. Poniamo ora di scegliere r in funzione di n , ossia $r = r(n)$; il problema del quale ci occupiamo è scegliere il più piccolo

valore di $r(n)$ che permette di ottenere un grafo connesso. Naturalmente, essendo $G_{n,r(n)}$ un evento aleatorio, vogliamo studiarne la connessione in ambito probabilistico, ossia ci interessa che $G_{n,r(n)}$ sia connesso *con buona probabilità*.

1.4.3 Reti wireless e reti di sensori

Le **reti wireless** ad-hoc sono costituite da nodi che dispongono di ricetrasmittitori wireless mediante i quali realizzano comunicazioni peer-to-peer. In una rete wireless ad hoc, tipicamente, avvengono comunicazioni multi-hop, ossia, due nodi i cui trasmettitori non sono in grado di connettersi l'uno con l'altro (a causa dell'elevata distanza dei nodi, o della presenza di ostacoli, o per qualunque altra causa) possono comunicare mediante una serie di nodi intermedi che si inoltrano l'uno con l'altro il messaggio da comunicare finché esso raggiunge il nodo destinazione. Le **reti di sensori** sono una particolare classe di reti wireless ad-hoc in cui i nodi, ciascuno contenente un sensore, sono dotati di una batteria di capacità particolarmente limitata, di un ricetrasmittitore wireless, e di un semplice processore. I sensori raccolgono i dati di interesse (ad esempio, temperatura, pressione, composizione del suolo, ecc...), e li trasmettono (eventualmente compressi e/o aggregati con quelli dei nodi vicini) agli altri nodi. In questo modo, ogni nodo della rete acquisisce una visione globale della zona monitorata, che può essere accessibile ad un utente esterno collegato alla rete attraverso uno o più nodi gateway. Il ricetrasmittitore wireless può essere configurato in modo da trasmettere entro un certo **raggio di trasmissione** (se il trasmettitore di un dispositivo x è configurato per trasmettere entro un raggio r_x , quel dispositivo potrà inviare messaggi solo ai dispositivi che distano meno di r_x da esso). È possibile pertanto rappresentare tale rete mediante un grafo diretto $\vec{G} := (V, \vec{E})$ dove $\vec{E} := \{(x, y) : y \text{ è a distanza } d \leq r_x \text{ da } x\}$. Il grafo diretto che rappresenta la rete è chiamato **grafo di comunicazione**, e utilizza una comunicazione *multi-hop*.

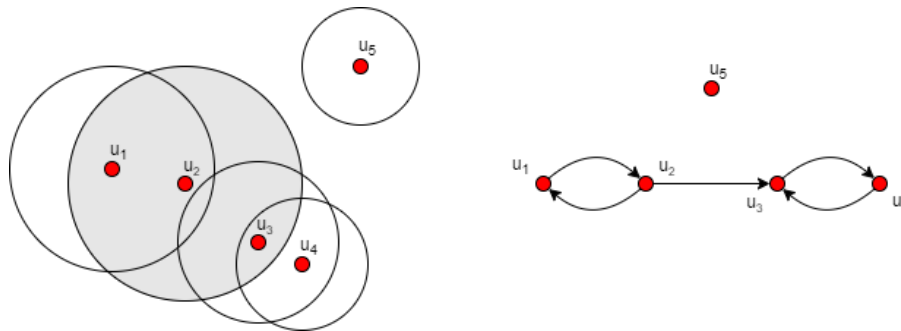


Figura 1.3: Un esempio di rete wireless e corrispettivo grafo di comunicazione.

Se vogliamo che ciascuno nodo possa comunicare con qualunque altro nodo è necessario che il grafo di comunicazione sia *fortemente connesso*; fare questo è facile: se configuriamo il trasmettitore di ciascun nodo ad un raggio di trasmissione pari alla distanza di quel nodo dal nodo ad esso più distante (ossia, detto V l'insieme dei nodi e indicata con $d(u, v)$ la distanza tra i nodi u e v , per ogni $u \in V$ poniamo $r_u = \max\{d(u, v) : v \in V - \{u\}\}$) allora il grafo di comunicazione è un grafo completo in cui ogni nodo può inviare messaggi direttamente al destinatario, senza ricorrere alla comunicazione multi-hop. Il problema è che questa soluzione presenta alti costi di mantenimento e di comunicazione; pertanto, è necessario assegnare a ciascun nodo un raggio di trasmissione il più piccolo possibile.

1.4.4 Reti wireless e grafi geometrici aleatori

Assumiamo che tutti i nodi abbiano lo stesso raggio di trasmissione r , e assumiamo inoltre che gli n nodi siano distribuiti uniformemente a caso in una regione limitata di piano (senza perdita di generalità, tale regione è il quadrato $Q = [0, 1] \times [0, 1]$) e che il raggio di trasmissione, uguale per tutti i nodi, sarà una funzione di n . La nostra rete è allora modellata da un grafo geometrico aleatorio che in particolare è un grafo non orientato (perché $d(u, v) \leq r(n)$ se e solo se $d(v, u) \leq r(n)$). Il problema che ci accingiamo a studiare è pertanto il seguente: dati n punti distribuiti uniformemente a caso nel quadrato $Q = [0, 1] \times [0, 1]$, calcolare il valore minimo di $r(n)$ affinché $G_{n,r(n)}$ sia connesso.

Teorema (1.3): Detto $r^*(n)$ il minimo valore per $r(n)$ che garantisce, con probabilità ragionevole, che $G_{n,r(n)}$ è connesso, allora $r^*(n) \in \Theta(\sqrt{\frac{\ln n}{n}})$.

È possibile dimostrare questo teorema dividendolo in due teoremi separati:

Teorema (1.4): Esiste una costante $\gamma > 0$ tale che se $r(n) \geq \gamma \sqrt{\frac{\ln n}{n}}$, allora $G_{n,r(n)}$ è connesso con alta probabilità.

Teorema (1.5): Per ogni costante $c > 0$, se $r(n) = \sqrt{\frac{\ln n + c}{\pi n}}$, allora $\lim_{n \rightarrow \infty} P(G_{n,r(n)} \text{ non è connesso}) > 0$.

Per dimostrare questi teoremi inoltre useremo il seguente lemma:

Lemma (1.3a): Per ogni $x \in \mathbb{R} : 1 - x \leq e^{-x}$. Inoltre, se $x \neq 0$, allora $1 - x < e^{-x}$.

Lemma (1.3a): Definiamo la funzione $f(x) = 1 - x - e^{-x}$, e calcoliamo la derivata prima di $f(x)$, ossia $f'(x) = e^{-x} - 1$. Studiamo il segno di $f'(x)$:

$$f'(x) \geq 0 \rightarrow e^{-x} - 1 \geq 0 \rightarrow e^{-x} \geq e^0 \rightarrow x \leq 0$$

Abbiamo che $f'(x) \geq 0$ per $x \leq 0$, allora $f(x)$ ha un punto di massimo relativo in $x = 0$; inoltre, essendo l'unico punto in cui la derivata si annulla, $x = 0$ è anche un punto di massimo assoluto. Poiché $f(0) = 1 - 0 - e^{-0} = 0$, questo implica che $f(x) \leq f(0) = 0$ per ogni $x \in \mathbb{R}$, ossia $1 - x \leq e^{-x}$ per ogni $x \in \mathbb{R}$, inoltre $f(x) \leq f(0) = 0$ per ogni $x \neq 0$.

Teorema (1.4): Sia $k(n) > 0$ un intero dipendente da n , e partizioniamo il quadrato $Q = [0, 1] \times [0, 1]$ in $k^2(n)$ celle, ciascuna di lato $\frac{1}{k}$. Poniamo $r(n)$ pari alla

lunghezza della diagonale di una coppia di celle adiacenti (due celle sono adiacenti se hanno un lato in comune), ossia $r(n) = \sqrt{(\frac{2}{k(n)})^2 + (\frac{1}{k(n)})^2} = \frac{\sqrt{5}}{k(n)}$. In questo modo, ciascun nodo in una qualsiasi cella è collegato da un arco a tutti i nodi (eventualmente) contenuti in tutte le celle adiacenti.

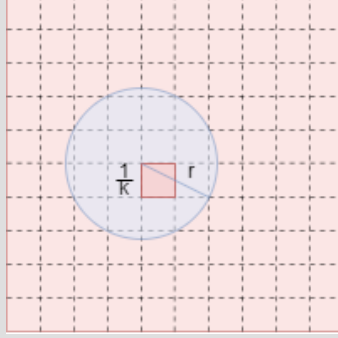


Figura 1.4: Rappresentazione grafica dello spazio partizionato e della grandezza del raggio.

Dimostriamo ora che è possibile scegliere $k(n)$ in modo tale che, con alta probabilità, nessuna cella è vuota. Invece di calcolare direttamente la probabilità di questo evento, calcoliamo la probabilità dell'evento complementare, ossia che esiste almeno una cella vuota. Sia C una cella, l'evento " $C = \emptyset$ " coincide con l'intersezione degli eventi " $1 \notin C \wedge 2 \notin C \wedge \dots \wedge n \notin C$ ", ossia " $\cap_{1 \leq i \leq n} (i \notin C)$ ", quindi possiamo esprimere la probabilità dell'evento come $P(C = \emptyset) = P(\cap_{1 \leq i \leq n} (i \notin C))$. Poiché i nodi sono posizionati in Q indipendentemente gli uni dagli altri, possiamo esprimere la probabilità dell'evento come $P(\cap_{1 \leq i \leq n} (i \notin C)) = \prod_{1 \leq i \leq n} P(i \notin C)$. Sia i un nodo, la probabilità che il nodo i sia scelto all'interno della cella C è pari al rapporto fra l'area di C e l'area del quadrato $Q = [0, 1] \times [0, 1]$; essendo l'area di Q uguale ad 1, abbiamo che $P(i \in C) = \frac{\text{area di } C}{\text{area di } Q} = \frac{1}{k^2(n)}$. Conseguentemente abbiamo che $P(i \notin C) = 1 - \frac{1}{k^2(n)}$, da cui abbiamo che:

$$P(C = \emptyset) = P(\cap_{1 \leq i \leq n} (i \notin C)) = \prod_{1 \leq i \leq n} P(i \notin C) = (1 - \frac{1}{k^2(n)})^n$$

Tramite la quale otteniamo che:

$$P(\exists C : C = \emptyset) = P(\cup_{C \in Q} (C = \emptyset)) \leq \sum_{C \in Q} P(C = \emptyset) = k^2(n)(1 - \frac{1}{k^2(n)})^n$$

Sappiamo che $r(n) = \frac{\sqrt{5}}{k(n)} \rightarrow k(n) = \frac{\sqrt{5}}{r(n)}$; sostituendo pertanto $\frac{\sqrt{5}}{r(n)}$ a $k(n)$ otteniamo che $P(\exists C : C = \emptyset) \leq \frac{5}{r^2(n)}(1 - \frac{r^2(n)}{5})^n$. Infine, ponendo $r(n) = c\sqrt{\frac{\ln n}{n}}$ otteniamo $P(\exists C : C = \emptyset) \leq \frac{5n}{c^2 \ln n}(1 - \frac{c^2 \ln n}{5n})^n$. Poiché $\frac{c^2 \ln n}{5n} \neq 0$ per ogni $n \in \mathbb{N}$, abbiamo

$(1 - \frac{\gamma \ln n}{5n}) < e^{-\frac{\gamma \ln n}{5n}}$, e quindi:

$$\begin{aligned}
P(\exists C : C = \emptyset) &< \frac{5n}{\gamma^2 \ln n} e^{-n \frac{\gamma^2 \ln n}{5n}} = \\
&= \frac{5n}{\gamma^2 \ln n} e^{-\frac{\gamma^2 \ln n}{5}} < \\
&< \frac{5n}{\gamma^2} e^{-\frac{\gamma^2 \ln n}{5}} = \\
&= \frac{5n}{\gamma^2} n^{-\frac{\gamma^2}{5}} = \\
&= \frac{5}{\gamma^2} n^{1-\frac{\gamma^2}{5}}
\end{aligned}$$

Osserviamo che $1 - \frac{c^2}{5} < 0$ per $c > \sqrt{5}$. In conclusione, scegliendo $r(n) = \Gamma \sqrt{\frac{\ln n}{n}}$ con $\gamma > \sqrt{5}$, e ponendo $b = \frac{5}{\gamma^2}$ e $c = \frac{\gamma^2}{5} - 1$, abbiamo che $P(\exists C : C = \emptyset) < \frac{b}{n^c}$ con $c > 0$; ossia abbiamo dimostrato che, se $r(n) = \gamma \sqrt{\frac{\ln n}{n}}$ con $\gamma > 0$, abbiamo che $P(G_{n,r(n)} \text{ è connesso}) = 1 - P(\exists C : C = \emptyset) > 1 - \frac{b}{n^c}$.

Teorema (1.5): Per cominciare, introduciamo i seguenti eventi:

- $\mathcal{E}_{\geq 1} := "G_{n,r(n)} \text{ contiene almeno un nodo isolato}"$.
- $\mathcal{E}_{i_1, i_2, \dots, i_h} := "tutti i nodi i_1, i_2, \dots, i_h sono isolati in $G_{n,r(n)}$, con $i_1, i_2, \dots, i_h \in [n]"$.$
- $\mathcal{E}_{i!} := "i \text{ è l'unico nodo isolato in } G_{n,r(n)}, \text{ con } i \in [n]"$.

Esprimiamo ora la probabilità che $G_{n,r(n)}$ non sia connesso usando gli eventi appena definiti. Ovviamente, se $G_{n,r(n)}$ contiene almeno un nodo isolato, allora $G_{n,r(n)}$ non è connesso: dunque $P(G_{n,r(n)} \text{ non è connesso}) \geq P(\mathcal{E}_{\geq 1})$; se i è l'unico nodo isolato in $G_{n,r(n)}$, con $i \in [n]$, allora $G_{n,r(n)}$ contiene almeno un nodo isolato: dunque $P(\mathcal{E}_{\geq 1}) \geq P(\cup_{i \in [n]} \mathcal{E}_{i!})$; poiché $\mathcal{E}_{1!}, \mathcal{E}_{2!}, \dots, \mathcal{E}_{n!}$ sono tutti eventi disgiunti, abbiamo che $P(\cup_{i \in [n]} \mathcal{E}_{i!}) = \sum_{i \in [n]} P(\mathcal{E}_{i!})$. In conclusione, abbiamo che:

$$P(G_{n,r(n)} \text{ non è connesso}) \geq P(\mathcal{E}_{\geq 1}) \geq P(\cup_{i \in [n]} \mathcal{E}_{i!}) = \sum_{i \in [n]} P(\mathcal{E}_{i!})$$

Poiché calcolare direttamente $P(\mathcal{E}_{i!})$ non è semplice, è più facile cercare di minorarla; a questo scopo possiamo osservare che i è l'unico nodo isolato in $G_{n,r(n)}$ se e solo se:

- i è un nodo isolato in $G_{n,r(n)}$.
- Comunque scegliamo un altro nodo j , i e j non sono entrambi isolati in $G_{n,r(n)}$.

Dunque:

$$\begin{aligned}\mathcal{E}_{i!} &= \mathcal{E}_i \cap_{j \in [n] - \{i\}} \mathcal{E}_{ij}^C = \mathcal{E}_i - \cup_{j \in [n] - \{i\}} \mathcal{E}_{ij} \\ P(\mathcal{E}_{i!}) &= P(\mathcal{E}_i - \cup_{j \in [n] - \{i\}} \mathcal{E}_{ij}) \geq P(\mathcal{E}_i) - P(\cup_{j \in [n] - \{i\}} \mathcal{E}_{ij}) \geq P(\mathcal{E}_i) - \sum_{j \in [n] - \{i\}} P(\mathcal{E}_{ij})\end{aligned}$$

Non ci resta a questo punto che trovare una minorazione per $P(\mathcal{E}_i)$ e una maggiorazione per $P(\mathcal{E}_{ij})$. Prima di procedere, indichiamo:

- Per $i \in [n]$, $t_i \in Q$ è il punto del quadrato nel quale è posizionato il nodo i .
- Per un punto $t \in Q$, $C_r(t)$ è il cerchio di centro t e raggio r .

A questo punto possiamo procedere:

- Partiamo dal minorare $P(\mathcal{E}_i)$: l'evento \mathcal{E}_i si verifica se e solo se, una volta fissato t_i , nessun nodo $j \neq i$ è posizionato in $C_r(t_i)$. Fissato t_i e fissato $j \neq i$, $P(t_j \notin C_r(t_i)) = \frac{\text{area di } (Q - C_r(t_i))}{\text{area di } Q} \geq 1 - \pi r^2$ (Il maggiore o uguale viene usato perché $C_r(t_i)$ potrebbe non essere completamente contenuto in Q , ossia se t_i è vicino al bordo di Q); allora, fissato t_i , $P(\forall j \neq i : t_j \notin C_r(t_i)) \geq (1 - \pi r^2)^{n-1}$. Il punto t_i , nel quale posizionare i , è scelto uniformemente a caso in Q , il quale è un insieme continuo, e la funzione di densità corrispondente alla scelta uniformemente a caso di un punto in Q è $f(t) = \frac{1}{\text{area di } Q} = 1$. Pertanto:

$$P(\mathcal{E}_i) \geq \int_{t_i \in Q} f(t_i)(1 - \pi r^2)^{n-1} dt_i = \int_{t_i \in Q} (1 - \pi r^2)^{n-1} dt_i = (1 - \pi r^2)^{n-1}$$

- Cerchiamo ora una maggiorazione per $P(\mathcal{E}_{ij})$: l'evento \mathcal{E}_{ij} si verifica se e solo se, una volta fissato t_i , j è posizionato in un nodo $t_j \notin C_r(t_i)$ e nessun nodo $h \in [n] - \{i, j\}$ è posizionato in $C_r(t_i) \cup C_r(t_j)$. Possiamo esprimere questo evento come unione di due eventi mutuamente esclusivi (ossia disgiunti):

- $\mathcal{E}_{ij}^1 := t_j \notin C_{2r}(t_i) \wedge \forall h \in [i, j], t_h \notin C_r(t_i) \cup C_r(t_j)$, ossia t_j è nella regione gialla.
- $\mathcal{E}_{ij}^2 := t_j \in C_{2r}(t_i) - C_r(t_i) \wedge \forall h \in [i, j], t_h \notin C_r(t_i) \cup C_r(t_j)$, ossia t_j è nell'anello azzurro.

Pertanto $P(\mathcal{E}_{ij}) = P(\mathcal{E}_{ij}^1 \cup \mathcal{E}_{ij}^2) = P(\mathcal{E}_{ij}^1) + P(\mathcal{E}_{ij}^2)$.

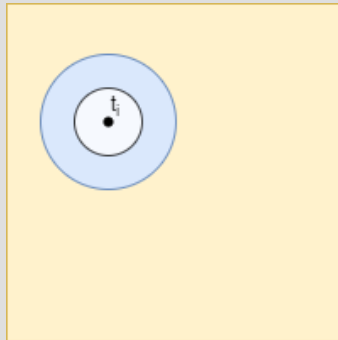


Figura 1.5: Rappresentazione grafica delle aree prese in considerazione dato un punto t_i .

Calcoliamo $P(\mathcal{E}_{ij}^1)$: fissiamo i punti t_i e t_j nella zona gialla, e fissiamo poi $h \in [n] - \{i, j\}$, la probabilità di scegliere t_h nella regione rimanente è pari al rapporto dell'area della regione rimanente con l'area del quadrato, ossia $1 - 2\pi r^2$.

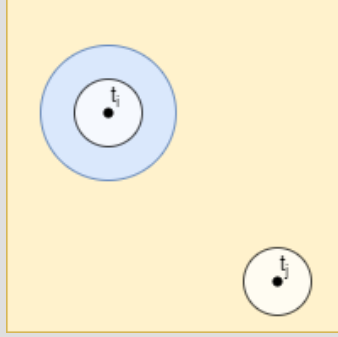


Figura 1.6: Rappresentazione grafica delle aree prese in considerazione dato un punto t_i e un punto t_j .

Pertanto, fissato t_i e fissato t_j nella zona gialla, la probabilità che per ogni $h \in [n] - \{i, j\} : t_h \notin C_r(t_i) \cup C_r(t_j)$ è $(1 - 2\pi r^2)^n - 2$. Fissato t_i la probabilità che, scegliendo t_j nella zona gialla, per ogni $h \in [n] - \{i, j\} : t_h \notin C_r(t_i) \cup C_r(t_j)$ è $\int_{t_j \in Q - C_{2r}(t_i)} f(t_i)(1 - 2\pi r^2)^{n-2} dt_j$. Infine:

$$\begin{aligned} P(\mathcal{E}_{ij}^1) &= \int_{t_i \in Q} f(t_i) \int_{t_j \in Q - C_{2r}(t_i)} f(t_i)(1 - 2\pi r^2)^{n-2} dt_j dt_i = \\ &= \int_{t_i \in Q} \int_{t_j \in Q - C_{2r}(t_i)} (1 - 2\pi r^2)^{n-2} dt_j dt_i = \\ &= \int_{t_i \in Q} (1 - 4\pi r^2)(1 - 2\pi r^2)^{n-2} dt_i = (1 - 4\pi r^2)(1 - 2\pi r^2)^{n-2} \end{aligned}$$

Maggioriamo adesso $P(\mathcal{E}_{ij}^2)$: fissiamo il punto t_i , fissiamo il punto t_j nella zona celeste, e fissiamo poi $h \in [n] - \{i, j\}$, la probabilità di scegliere t_h nella regione rimanente è pari al rapporto dell'area della regione rimanente con l'area del quadrato, e questa volta dipende dalla posizione di t_j nella zona celeste.

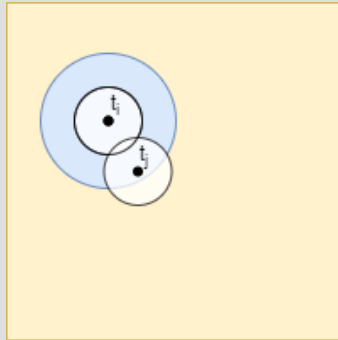


Figura 1.7: Rappresentazione grafica delle aree prese in considerazione dato un punto t_i e un punto t_j .

La probabilità di scegliere t_h nella regione rimanente è massima quando è massima l'intersezione di $C_r(t_i)$ con $C_r(t_j)$, ossia quando j è sulla circonferenza che delimita $C_r(t_i)$. In questo caso l'area dell'unione di $C_r(t_i)$ con $C_r(t_j)$ è $2\pi r^2$ meno l'area della lente formata da $C_r(t_i)$ e $C_r(t_j)$. Nel caso massimo la lente presenta un angolo pari a $\theta = \frac{2\pi}{3}$, pertanto la sua area è pari a $r^2(\theta - \sin \theta) = r^2(\frac{2\pi}{3} - \frac{\sqrt{3}}{2}) = 2r^2(\frac{\pi}{3} - \frac{\sqrt{3}}{4})$.

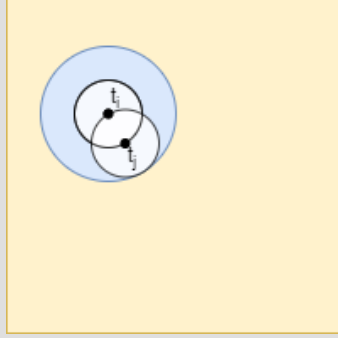


Figura 1.8: Rappresentazione grafica delle aree prese in considerazione dato un punto t_i e un punto t_j .

Da qui abbiamo che l'area dell'intersezione dei due cerchi è pari a:

$$2\pi r^2 - 2r^2(\frac{\pi}{3} - \frac{\sqrt{3}}{2}) = \frac{4\pi}{3}r^2 + \frac{\sqrt{3}\pi}{2\pi}r^2 = \pi r^2(\frac{4}{3} + \frac{\sqrt{3}}{2\pi}) > \frac{8}{5}\pi r^2$$

Pertanto, la probabilità di scegliere t_h nella regione rimanente è minore di $1 - \frac{8}{5}\pi r^2$, e quindi la probabilità di scegliere tutti gli $n - 2$ punti t_h nella regione rimanente è minore di $(1 - \frac{8}{5}\pi r^2)^{n-2}$. Fissato t_i , la probabilità che, scegliendo t_j nella zona celeste, per ogni $h \in [n] - \{i, j\}$, $t_h \notin C_r(t_i) \cup C_r(t_j)$ è minore di $\int_{t_j \in C_{2r}(t_i) - C_r(t_i)} f(t_j)(1 - \frac{8}{5}\pi r^2)^{n-2} dt_j$. Infine:

$$\begin{aligned} P(\mathcal{E}_{ij}^2) &= \int_{t_i \in Q} f(t_i) \int_{t_j \in C_{2r}(t_i) - C_r(t_i)} f(t_j)(1 - \frac{8}{5}\pi r^2)^{n-2} dt_j dt_i = \\ &= \int_{t_i \in Q} \int_{t_j \in C_{2r}(t_i) - C_r(t_i)} (1 - \frac{8}{5}\pi r^2)^{n-2} dt_j dt_i = \\ &= \int_{t_i \in Q} (4\pi r^2 - \pi r^2)(1 - \frac{8}{5}\pi r^2)^{n-2} dt_i = 3\pi r^2(1 - \frac{8}{5}\pi r^2)^{n-2} \end{aligned}$$

Da qui abbiamo che:

$$\begin{aligned} P(\mathcal{E}_{ij}) &= P(\mathcal{E}_{ij}^1) + P(\mathcal{E}_{ij}^2) < \\ &< (1 - 4\pi r^2)(1 - 2\pi r^2)^{n-2} + 3\pi r^2(1 - \frac{8}{5}\pi r^2)^{n-2} < \\ &< (1 - 2\pi r^2)^{n-1} + 3\pi r^2(1 - \frac{8}{5}\pi r^2)^{n-2} < \\ &< (1 - 2\pi r^2)^{n-2} + 3\pi r^2(1 - \frac{8}{5}\pi r^2)^{n-2} \end{aligned}$$

Notare che nel calcolare $P(\mathcal{E}_{ij})$ abbiamo omesso il caso che t_i o t_j sono vicini al bordo; in realtà, complicando ellgermente la dimostrazione, è possibili arrivare agli stessi risultati considerando anche gli effetti al bordo tramite le stesse tecniche, e pertanto è possibile ignorare tali casi. Sostituiamo ora $r = \sqrt{\frac{\ln n + c}{n\pi}}$:

$$\begin{aligned}
P(\mathcal{E}_{ij}) &< (1 - 2\pi \frac{\ln n + c}{n})^{n-2} + 3\pi \frac{\ln n + c}{n} (1 - \frac{8}{5}\pi \frac{\ln n + c}{n})^{n-2} < \\
&< e^{-2\frac{\ln n + c}{n}(n-2)} + 3\frac{\ln n + c}{n} e^{-\frac{8}{5}\frac{\ln n + c}{n}(n-2)} = \\
&= e^{-2\frac{\ln n + c}{n}(n-2)} + 3\frac{\ln n + c}{n^{\frac{3}{5}}} n^{-\frac{2}{5}} e^{-\frac{8}{5}\frac{\ln n + c}{n}(n-2)} < \\
&< n^{-2\frac{n-2}{n}} e^{-2\frac{c(n-2)}{n}} + 3\frac{\ln n + c}{n^{\frac{3}{5}}} n^{-2\frac{n-2}{n}} e^{-\frac{8}{5}\frac{c(n-2)}{n}} = \\
&= n^{-2\frac{n-2}{n}} e^{-2\frac{c(n-2)}{n}} (1 + 3\frac{\ln n + c}{n^{\frac{3}{5}}} e^{\frac{2}{5}\frac{c(n-2)}{n}})
\end{aligned}$$

Riassumiamo:

$$\begin{aligned}
P(G \text{ non è connesso}) &\geq \sum_{i \in [n]} P(\mathcal{E}_i) \geq \sum_{i \in [n]} (P(\mathcal{E}_i) - \sum_{j \in [n] - \{i\}} P(\mathcal{E}_{ij})) > \\
&> n(1 - \pi r^2)^{n-1} - n(n-1) n^{-2\frac{n-2}{n}} e^{-2\frac{c(n-2)}{n}} (1 + 3\frac{\ln n + c}{n^{\frac{3}{5}}} e^{\frac{2}{5}\frac{c(n-2)}{n}})
\end{aligned}$$

Osserviamo che:

$$\lim_{n \rightarrow \infty} n(n-1) n^{-2\frac{n-2}{n}} e^{-2\frac{c(n-2)}{n}} (1 + 3\frac{\ln n + c}{n^{\frac{3}{5}}} e^{\frac{2}{5}\frac{c(n-2)}{n}}) = e^{-2c}$$

Ossia, per ogni $\epsilon > 0$, esiste un $n_\epsilon > 0$ tale che per ogni $n \geq n_\epsilon$ abbiamo che:

$$n(n-1) n^{-2\frac{n-2}{n}} e^{-2\frac{c(n-2)}{n}} (1 + 3\frac{\ln n + c}{n^{\frac{3}{5}}} e^{\frac{2}{5}\frac{c(n-2)}{n}}) = (1 + \epsilon) e^{-2c}$$

Allora, per ogni $\epsilon > 0$, esiste un $n_\epsilon > 0$ tale che per ogni $n \geq n_\epsilon$ abbiamo che $P(G \text{ non è connesso}) > n(1 - \pi r^2)^{n-1} - (1 + \epsilon) e^{-2c}$. Per dimostrare che $\lim_{n \rightarrow \infty} P(G \text{ non è connesso}) > 0$ è sufficiente dimostrare che, da un certo n in poi:

$$\begin{aligned}
n(1 - \pi r^2)^{n-1} - (1 + \epsilon) &> 0 \\
n(1 - \pi r^2)^{n-1} &> (1 + \epsilon) e^{-2c}
\end{aligned}$$

Calcoliamo il logaritmo del membro sinistro della disuguaglianza:

$$\ln(n - (1 - \pi r^2)^{n-1}) = \ln n + \ln(1 - \pi r^2)^{n-1} = \ln n + (n-1) \ln(1 - \pi r^2)$$

Ricordiamo che, per $x < 1$, $\ln(1 - x) = -\sum_{k=1}^{\infty} \frac{(x)^k}{k}$, da cui:

$$\ln n + (n-1) \ln(1 - \pi r^2) = \ln n - (n-1) \sum_{k=1}^{\infty} \frac{(\pi r^2)^k}{k}$$

E poichè $r = \sqrt{\frac{\ln n + c}{n\pi}}$, ossia $\pi r^2 = \frac{\ln n + c}{n}$, allora:

$$\ln n - (n-1) \sum_{k=1}^{\infty} \frac{(\pi r^2)^k}{k} = \ln n - (n-1) \sum_{k=1}^{\infty} \frac{(\ln n + c)^k}{kn^k}$$

A questo punto poniamo $\delta(n) = \sum_{k=3}^{\infty} \frac{(\ln n + c)^k}{kn^k}$ così che:

$$\begin{aligned} \ln n - (n-1) \sum_{k=1}^{\infty} \frac{(\ln n + c)^k}{kn^k} &= \ln n - (n-1) \left(\sum_{k=1}^2 \frac{(\ln n + c)^k}{kn^k} + \delta(n) \right) = \\ &= \ln n - (n-1) \left(\frac{\ln n + c}{n} + \frac{(\ln n + c)^2}{2n^2} + \delta(n) \right) \end{aligned}$$

A questo punto, non ci resta che maneggiare $\delta(n)$:

$$\begin{aligned} \delta(n) &= \sum_{k=3}^{\infty} \frac{(\ln n + c)^k}{kn^k} \leq \frac{1}{3} \sum_{k=3}^{\infty} \frac{(\ln n + c)^k}{n^k} \leq \\ &\leq \frac{1}{3} \int_2^{\infty} \left(\frac{\ln n + c}{n} \right)^x dx = \\ &= \lim_{k \rightarrow \infty} \frac{1}{3} \int_2^k \left(\frac{\ln n + c}{n} \right)^x dx = \\ &= \lim_{k \rightarrow \infty} \frac{1}{3} \int_2^k e^{x \ln \frac{\ln n + c}{n}} dx = \\ &= \lim_{k \rightarrow \infty} \left[\frac{1}{3} \frac{1}{\ln \left(\frac{\ln n + c}{n} \right)} e^{x \ln \left(\frac{\ln n + c}{n} \right)} \right]_2^k = \\ &= \lim_{k \rightarrow \infty} \left[\frac{1}{3} \frac{1}{\ln \left(\frac{\ln n + c}{n} \right)} \left(\frac{\ln n + c}{n} \right)^x \right]_2^k = \\ &= \frac{1}{3} \frac{1}{\ln \left(\frac{\ln n + c}{n} \right)} \frac{(\ln n + c)^2}{n^2} \end{aligned}$$

E poichè $\lim_{n \rightarrow \infty} \frac{1}{\ln \left(\frac{\ln n + c}{n} \right)} = 0$, allora $\frac{1}{\ln \left(\frac{\ln n + c}{n} \right)} < 1$ per n sufficientemente grande, abbiamo in conclusione che $\delta(n) < \frac{1}{3} \frac{(\ln n + c)^2}{n^2}$, e quindi:

$$\begin{aligned} \ln n - (n-1) \sum_{k=1}^{\infty} \frac{(\ln n + c)^k}{kn^k} &= \ln n - (n-1) \left(\frac{\ln n + c}{n} + \frac{(\ln n + c)^2}{2n^2} + \delta(n) \right) > \\ &> \ln n - (n-1) \left(\frac{\ln n + c}{n} + \frac{(\ln n + c)^2}{2n^2} + \frac{1}{3} \frac{(\ln n + c)^2}{n^2} \right) = \\ &= \ln n - (n-1) \left(\frac{\ln n + c}{n} + \frac{5(\ln n + c)^2}{6n^2} \right) = \\ &= \ln n - \frac{n-1}{n} (\ln n + c) - \frac{5(n-1)(\ln n + c)^2}{6n^2} > \\ &> \ln n - (\ln n + c) - \frac{5(n-1)(\ln n + c)^2}{6n^2} = \\ &= -c - \frac{5(n-1)(\ln n + c)^2}{6n^2} \end{aligned}$$

Poiché $\lim_{n \rightarrow \infty} \frac{5(n-1)(\ln n + c)^2}{6n^2} = 0$, allora per ogni $\omega > 0$ esiste un $n_\omega \geq n_\epsilon$ tale che per ogni $n \geq n_\omega$ abbiamo che $\frac{5(n-1)(\ln n + c)^2}{6n^2} < \omega$; allora, per n sufficientemente grande, $\ln(n(1 - \pi r^2)^{n-1}) > -c - \omega$, e dunque $n(1 - \pi r^2)^{n-1} > e^{-c-\omega}$, ossia:

$$n(1 - \pi r^2)^{n-1} - (1 + \epsilon)e^{-2c} > e^{-c-\omega} - (1 + \epsilon)e^{-2c}.$$

Scegliamo $\omega < c - \ln(1 + \epsilon)$, allora:

$$n(1 - \pi r^2)^{n-1} > e^{-c-\omega} > e^{-c-c+\ln(1+\epsilon)} = e^{-2c}e^{\ln(1+\epsilon)} = (1 + \epsilon)e^{-2c}$$

Pertanto, per n sufficientemente grande:

$$P(G \text{ non è connesso}) > n(1 - \pi r^2)^{n-1} - (1 + \epsilon)e^{-2c} > (1 + \epsilon)e^{-2c} - (1 + \epsilon)e^{-2c} > 0$$

1.5 Il fenomeno Small World

Capitolo 2

Teoria dei grafi e delle reti sociali

2.1 L'esperimento di Granovetter

Negli anni '60 il sociologo statunitense Mark Granovetter, nel preparare la sua tesi di dottorato, intervistò un gruppo di individui che avevano recentemente cambiato lavoro. Granovetter fece loro una serie di domande volte a capire in che modo erano venuti a conoscenza della possibilità di ottenere l'impiego che avevano ottenuto (come avevano saputo dell'esistenza dell'azienda che poi li aveva assunti, come avevano saputo che quell'azienda cercava personale con le loro competenze, ecc.). Il risultato fu che molti degli intervistati avevano avuto le informazioni che li avevano condotti alla loro occupazione attuale attraverso una comunicazione da parte di qualcuno che conoscevano, ossia tramite una sorta di passa parola, cosa che può ben essere prevedibile. Il fatto abbastanza inaspettato che emerse dalle sue interviste fu che spesso l'informazione era arrivata da qualcuno che conoscevano superficialmente e non dagli amici più stretti. Sorprendentemente, perché uno si aspetta che i nostri più cari amici sono quelli che più si darebbero da fare per aiutarci nella ricerca di un lavoro. Quello che l'esperimento di Granovetter sembra suggerire è che non è tanto la forza del legame quella che ha aiutato a trovare lavoro quanto il tipo di informazione che una relazione è capace di veicolare. In effetti, pensandoci bene, i nostri più cari amici frequentano, grosso modo, gli stessi posti che frequentiamo noi (se siamo così tanto amici vuol dire che abbiamo tante cose in comune e che ci piace trascorrere il tempo insieme), e così noi e i nostri grandi amici abbiamo accesso alle stesse fonti di informazione. Invece i nostri conoscenti (alla lontana) frequentano posti diversi da quelli che frequentiamo noi, cosicché hanno accesso a fonti di informazione diverse provenienti da "zone" diverse della rete, alle quali non avremmo accesso senza il loro aiuto, così ampliano il nostro "raggio d'azione".

2.2 Bridges e Local Bridges

Poniamo di avere un grafo $G := (V, E)$ che modella una rete sociale, dove ogni nodo è una persona e ogni arco rappresenta una conoscenza/amicizia/legame tra due persone. Gli archi che "aumentano" le nostre informazioni sono quelli che ci collegano a regioni della rete inaccessibili ai nodi vicini; in paragone, le conoscenze più lontane dalla nostra cerchia di amicizie possono fornirci fonti di informazioni diverse da quelle che abbiamo. Questi archi vengono detti **bridge** e sono archi di maggiore "valore informativo" la cui rimozione disconnette la rete. Sappiamo però che una rete con una buona probabilità contiene componenti giganti, e che quindi la presenza di bridge è poco probabile; è possibile

rilassare la definizione di bridge definendo come **local bridge** un arco i cui nodi non hanno vicini in comune (ossia $(u, v) \in E$ è un local bridge se $N(u) \cap N(v) = \emptyset$).

2.3 Legami forti e deboli

L'esperimento di Granovetter ci fornisce anche un altro tipo di indicazione oltre all'individuazione del potere "informativo" di bridge e local bridge: spesso l'informazione era arrivata da qualcuno che conosciuto solo superficialmente e non dagli amici più stretti. Tramite ciò possiamo riconoscere all'interno di una rete sociale due insieme di relazioni:

- **Relazioni forti**, ossia quelle che ci collegano agli amici stretti.
- **Relazioni deboli**, ossia quelle che ci collegano ai semplici conoscenti.

Conseguentemente, possiamo modellare una rete di questo tipo mediante un grafo $G := (V, E)$ in cui gli archi sono partizionati in due sottoinsiemi:

- **archi forti** (o *strong ties*) indicato con S .
- **archi deboli** (o *weak ties*) indicato con W .

Pertanto, $G := (V, E) = (V, S \cup W)$ con $S \cap W = \emptyset$. Come ha mostrato l'esperimento di Granovetter, gli archi deboli veicolano informazioni alle quali non avremmo accesso tramite gli archi forti: e questa è la forza degli archi deboli

2.3.1 La chiusura triadica

I nostri amici più stretti tendono a frequentare i nostri stessi posti, e pertanto hanno una buona possibilità di incontrarsi ed entrare in contatto. Poniamo di avere tre individui (nodi) a, b e c , tali che esiste un rapporto di amicizia tra a e b e tra a e c ($(a, b), (a, c) \in S$), allora è probabile che prima o poi b e c entreranno in contatto ($(b, c) \in S \cap W$), ossia si genererà una **chiusura triadica**; per descrivere queste reti necessitiamo di grafi dinamici, ossia di grafi che evolvono nel tempo. Poniamo di avere un grafo dinamico $G := (V, S \cup W)$, il quale si è evoluto per un certo periodo di tempo raggiungendo una configurazione stabile; un nodo $v \in V$ soddisfa la *proprietà della chiusura triadica forte (STCP)* se $\forall (v, u) \in S, \forall (v, w) \in S, (u, w) \in S \cup W$. Il grafo G soddisfa la STCP se tutti i suoi nodi soddisfano la STCP.

Teorema (2.1): Sia $G := (V, S \cup W)$, se un nodo $x \in V$ soddisfa la STCP e se esistono nodi distinti $u, v \in V$ tali che $(x, u) \in S$ e $(x, v) \in S \cup W$ è un local bridge, allora $(u, v) \in W$.

Teorema (2.1): Poiché (x, v) è un local bridge allora $N(x) \cap N(v) = \emptyset$. Se fosse $(x, v) \in S$, allora poiché $(x, u) \in S$ e v soddisfa la STCP, dovrebbe essere $(x, v) \in S \cup W$, ossia $u \in N(x) \cap N(v) \neq \emptyset$, ossia una contraddizione.

Osservazione: Questo teorema mostra come una proprietà globale (essere un local bridge) si riflette in una proprietà locale (essere un weak tie).

2.4 I clusters (o comunità)

Se $(a, b), (a, c) \in S$ allora è probabile che prima o poi si formi l'arco $(b, c) \in S \cup W$, ossia è probabile che si generi una chiusura triadica. Essendo il grafo dinamico, si svilupperanno nel tempo varie chiusure triadiche, formando un gruppo di nodi fortemente coeso e con un elevato grado di interconnessione fra i nodi che lo compongono che prende il nome di **cluster** o **comunità**. Per misurare il grado di coesione di un unodo v all'interno di un gruppo di nodi viene definito il **coefficiente di clustering** come il rapporto fra il numero di relazioni tra i vicini di v rispetto a tutte le coppie possibili dei vicini di v :

$$\begin{aligned} \forall v \in V, c(v) &= \frac{|\{(x, y) \in E : x \in N(v) \wedge y \in N(v)\}|}{\binom{|N(v)|}{2}} = \\ &= \frac{|\{(x, y) \in E : x \in N(v) \wedge y \in N(v)\}|}{\frac{|N(v)|[|N(v)|-1]}{2}} \end{aligned}$$

Il coefficiente di clustering è anche un **indice di centralità** del nodo v : infatti, un nodo con coefficiente di clustering basso si troverà in una posizione periferica del gruppo, mentre un nodo con coefficiente di clustering alto si troverà al centro del gruppo. Sia $C \subseteq V$, definiamo per ogni nodo v in C il **coefficiente di clustering relativo** come il rapporto fra il numero di relazioni tra i vicini di v presenti in C rispetto a tutte le coppie possibili dei vicini di v in C :

$$\begin{aligned} \forall v \in V, c_C(v) &= \frac{|\{(x, y) \in E : x \in N(v) \cap C \wedge y \in N(v) \cap C\}|}{\binom{|N(v) \cap C|}{2}} = \\ &= \frac{|\{(x, y) \in E : x \in N(v) \cap C \wedge y \in N(v) \cap C\}|}{\frac{|N(v) \cap C| [|N(v) \cap C| - 1]}{2}} \end{aligned}$$

Se tutti i nodi contenuti in C hanno coefficiente di clustering relativo a C elevato, possiamo ben pensare che C sia una comunità.

2.4.1 Le cut-communities

Una **cut-community** per un grafo $G := (V, E)$ è un sottoinsieme proprio e non vuoto C dei nodi G che minimizza gli archi del taglio, ossia gli archi che collegano nodi in C a nodi in $V - C$. Formalmente, dato un grafo $G := (V, E)$, $C \subset V$ è una cut-community se $C \neq \emptyset$ e

$$|\{(u, v) : u \in C \wedge v \in V - C\}| = \min_{C' \subset V : C' \neq \emptyset} (|\{(u, v) : u \in C' \wedge v \in V - C'\}|)$$

Dati un grafo $G := (V, E)$ e una coppia di nodi $s, t \in V$, un taglio minimo rispetto alla coppia (s, t) è un sottoinsieme proprio e non vuoto $C \subset V$ dei nodi di G che contiene s e non contiene t , e che minimizza gli archi del taglio. Calcolare il taglio minimo rispetto ad una coppia di nodi è facile; ad esempio si può utilizzare l'algoritmo di Fordo-Fulkerson che calcola il taglio minimo che separa due dati nodi, ossia: dati due nodi s e t , l'algoritmo calcola un sottoinsieme C di nodi tale che $s \in C$, $t \notin C$ e

$$|\{(u, v) : u \in C \wedge v \in V - C\}| = \min_{C' \subset V : s \in C' \wedge t \notin C'} (|\{(u, v) : u \in C' \wedge v \in V - C'\}|)$$

Quindi per calcolare una cut-community di un grafo possiamo procedere così: per ogni coppia di nodi distinti $s, t \in V$ calcoliamo l'insieme $C_{s,t}$ tale che $s \in C_{s,t}$, $t \in V - C_{s,t}$ e che minimizza il taglio; la cut-community cercata è il sottoinsieme $C_{x,y}$ tale che

$$|\{(u, v) : u \in C_{x,y} \wedge v \in V - C_{x,y}\}| = \min_{s, t \in V : s \neq t} (|\{(u, v) : u \in C'_{s,t} \wedge v \in V - C'_{s,t}\}|)$$

Il problema dell'algoritmo di Ford-Fulkerson e di altri algoritmi simili è che non permettono di "controllare" i due insiemi che costituiscono il taglio. Poniamo di avere un grafo $G := (V, E)$ composto dall'unione di due clique, $A := \{a_0, a_1, a_2, a_3, a_4\}$ e $B := \{b_0, b_1, b_2, b_3, b_4\}$; l'insieme degli archi E è definito come

$$E = \{(a_i, a_j) : a_i, a_j \in A\} \cup \{(b_i, b_j) : b_i, b_j \in B\} \cup \\ \cup \{(a_i, b_j) : a_i \in A, b_j \in B, i \equiv_5 (j - 1, j, j + 1)\}$$

In questo modo, ogni nodo ha grado 7, e usando un qualsiasi algoritmo per il calcolo del taglio minimo otterremmo che i tagli minimi sono tutti e soli quelli che isolano i singoli nodi, poiché tagliano 7 archi ciascuno.

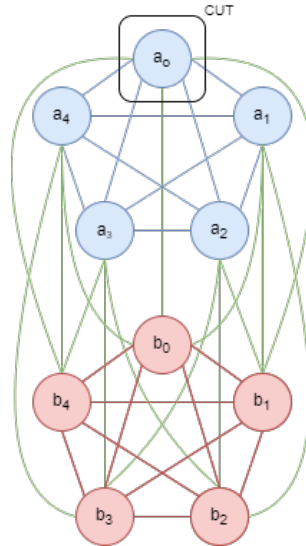


Figura 2.1: Grafo dell'esempio citato sopra: le clique sono colorate una di blu e una di rosso, mentre gli archi di collegamento sono in verde. Inoltre è riportato un esempio di cut minimo.

2.4.2 Le web-communities

Una **web-community** è un sottoinsieme dei nodi in un grafo ciascuno dei quali ha più vicini fra i nodi del sottoinsieme che fra quelli esterno ad esso. Formalmente, dato un grafo $G := (V, E)$, $C \subset V$ è una **(strong) web community** se $C \neq \emptyset$ e

$$\forall u \in C, |N(u) \cap C| > |N(u) - C| = |N(u) \cap (V - C)|$$

O, equivalentemente, poiché $|N(u) \cap C| + |N(u) - C| = |N(u)|$, $\frac{|N(u) \cap C|}{|N(u)|} > \frac{1}{2}$. Analogamente, dato un grafo $G := (V, E)$, $C \subset V$ è una **(weak) web community** se $C \neq \emptyset$ e

$$\forall u \in C, |N(u) \cap C| \geq |N(u) - C|$$

O, equivalentemente, $\frac{|N(u) \cap C|}{|N(u)|} \geq \frac{1}{2}$. Questa definizione è semplice e intuitivamente ragionevole, e può essere generalizzata richiedendo $\frac{|N(u) \cap C|}{|N(u)|} \geq \alpha$ con $\alpha \in [0, 1]$. Le definizioni di cut-community e web-community non sono del tutto scorrelate, infatti una cut-community che contiene almeno 2 nodi è anche una web-community. Ma abbiamo già osservato che nel calcolare una cut-community è difficile controllarne la cardinalità, poiché possono restituire comunità contenenti un solo nodo e pertanto poco significative; inoltre, non è detto che una cut-community di un solo nodo sia anche una web-community.

Teorema (2.2): Sia $G := (V, E)$ un grafo, se $C \subset V$ è una cut-community per G tale che $|C| > 1$, allora C è una weak web-community per G .

Teorema (2.2): Supponiamo per assurdo che C non sia una weak web-community, allora esiste un nodo $u \in C$ tale che $|N(u) \cap C| < |N(u) - C|$. Poiché $|C| > 1$, allora esiste in C un nodo v distinto da u , ossia $C - \{u\} \neq \emptyset$, inoltre

$$\begin{aligned} & |\{(x, y) \in E : x \in C - u \wedge y \in (V - C) \cup u\}| = \\ & = |\{(x, y) \in E : x \in C \wedge y \in V - C\}| - |\{(u, w) : w \in N(u) - C\}| + \\ & \quad + |\{(u, w) : w \in N(u) \cap C\}| = \\ & = |\{(x, y) \in E : x \in C \wedge y \in V - C\}| - |N(u) - C| + |N(u) \cap C| < \\ & < |\{(x, y) \in E : x \in C \wedge y \in V - C\}| \end{aligned}$$

E dunque $C' = C - \{u\}$ è un sottoinsieme proprio e non vuoto di V e il numero di archi del taglio indotto da C' è minore del numero di archi dal taglio indotto da C , così contraddicendo l'ipotesi che C è una cut-community.

Osservazione: Se C è una cut-community, allora anche $V - C$ è una cut-community perchè $V - C \neq V, V - C \neq \emptyset$ e il taglio indotto da $V - C$ è lo stesso di quello indotto da C ; perciò, un algoritmo che calcola un taglio minimo individua una partizione di un grafo in due comunità. Si può aggiungere che, se C è una cut-community con $|C| > 1$ e $|V - C| > 1$, allora $\langle C, V - C \rangle$ è una partizione del grafo in due web-communities.

2.5 Partizionare un grafo in due web-communities

Il problema **Strong Web Communities Partition (SWCP)** è definito così: dato un grafo $G := (V, E)$, decidere se esiste un sottoinsieme C di V tale che C e $C - V$ sono due strong web-communities.

Teorema (2.3): SWCP è NP-completo.

Lemma (2.3a): Se $G := (V, E)$ è partizionabile in due strong web-communities e esistono $u, v, w \in V$ tali che $N(u) = \{v, w\}$, allora, per ogni $C \subset V$ tale che C e $V - C$ sono due strong web-communities, $u, v, w \in C$ oppure $u, v, w \in V - C$

Lemma (2.3a): Sia $C \subset V$ tale che C e $V - C$ sono due strong web-communities (senza perdita di generalità assumiamo che $u \in C$); se fosse $v, w \in V - C$ allora $|N(u) \cap C| = 0 < |N(u) \cap (V - C)| = 2$, se fosse $v \in C$ e $w \in V - C$ allora $|N(u) \cap C| = |N(u) \cap (V - C)| = 1$. In tutti e due i casi verrebbe contraddetta l'ipotesi che C e $V - C$ sono due strong web-communities.

Teorema (2.3): Per dimostrare la completezza di SWCP riduciamo ad esso 3SAT. Siano $\{x_1, x_2, \dots, x_n\}$ variabili booleane, e sia $f = c_1 \wedge c_2 \wedge \dots \wedge c_m$ con:

- $\forall i \in [1, m], c_i = l_{i1} \vee l_{i2} \vee l_{i3}$.
- $\forall i \in [1, m], \forall j \in \{1, 2, 3\}, l_{ij} = x$ oppure $l_{ij} = \neg x$, dove $x \in X$.

Costruiamo un grafo $G := (V, E)$ costruito in questo modo:

- V contiene due nodi "specializzati" T ed F .
- Un gadget per ogni variabile x_i , formato dai nodi t_i, f_i, u_i, v_i, x_i e $\neg x_i$, più vari nodi senza nome collegati ai nodi x_i e $\neg x_i$ quante sono le clausole che contengono queste variabili più uno. Se T ed F si trovano nella stessa comunità allora tutti i nodi in figura sono in quella comunità; se T ed F sono in due comunità, per esempio $T \in C$ e $F \in V - C$, poiché t_i ed f_i hanno grado 2 allora T, t_i ed u_i devono essere contenuti in C mentre F, f_i e v_i devono essere contenuti in V_C . Affinché questo sia possibile è necessario che esattamente uno dei nodi x_i e $\neg x_i$ sia contenuto in C ed esattamente uno dei nodi x_i e $\neg x_i$ sia contenuto in $V - C$ (naturalmente, ciascun nodo senza nome deve essere contenuto nella stessa comunità che contiene il padre).
- Un gadget per ogni clausola, formato dai nodi c_j, l_{j1}, l_{j2} ed l_{j3} ; il nodo c_j è collegato alle variabili x_i ed $\neg x_i$ se queste sono contenute nella clausola corrispettiva. Se T ed F si trovano nella stessa comunità allora tutti i nodi in figura sono in quella comunità; se T ed F sono in due comunità, per esempio $T \in C$ e $F \in V - C$, poiché l_{j1}, l_{j2} ed l_{j3} hanno grado 2 allora T, l_{j1}, l_{j2} ed l_{j3} devono essere contenuti in C . Affinché questo sia possibile è necessario che almeno uno dei nodi x_i e $\neg x_i$ collegati a c_j sia contenuto in C , altrimenti c_j avrebbe tanti vicini in C quanti in V_C .

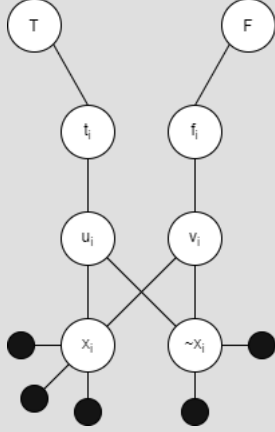


Figura 2.2: Un esempio di gadget per la variabile x_i . Nel nostro esempio x_i è contenuto in due clausole, mentre $\neg x_i$ in una sola.

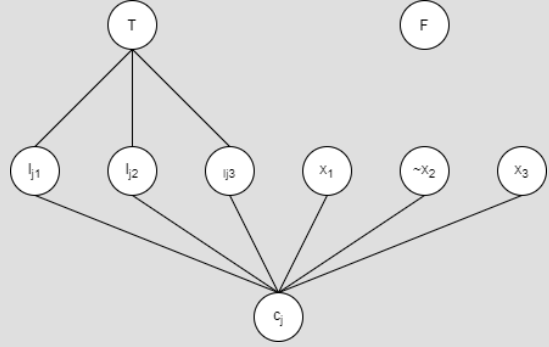


Figura 2.3: Un esempio di gadget per la clausola c_j .

Di seguito è riportato un esempio di grafo corrispondente alla funzione $f(x_1, x_2, x_3) = c_1 \wedge c_2$ con $c_1 = x_1 \vee \neg x_2 \vee x_3$ e $c_2 = \neg x_1 \vee \neg x_2 \vee \neg x_3$.

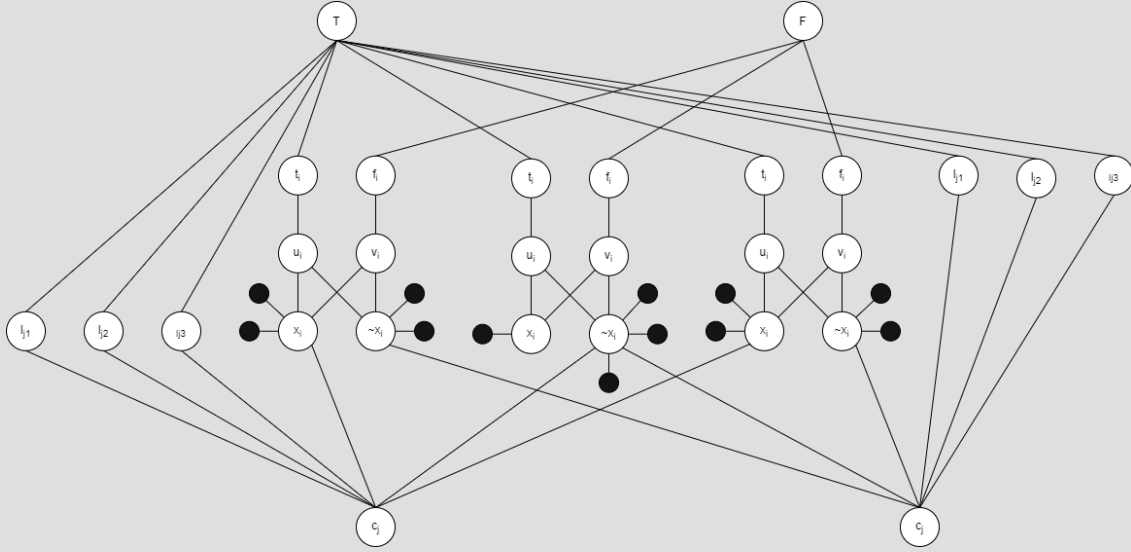


Figura 2.4: Un esempio di gadget per la clausola $f(x_1, x_2, x_3)$.

G è partizionabile in due strong web-communities solo se T ed F non sono nella stessa comunità (altrimenti, se T ed F sono nella stessa comunità, tutti i nodi in G sono in quella comunità). Affinché $T \in C$ ed $F \in V - C$:

- Per ogni variabile x_i in X , esattamente uno dei nodi x_i e $\neg x_i$ deve essere contenuto in C ed esattamente uno dei nodi x_i e $\neg x_i$ deve essere contenuto in $V - C$, e quindi ogni partizione di G in due strong web-communities corrisponde ad un'assegnazione di verità a per X .
- Per ogni clausola c_j , il nodo corrispondente deve appartenere in C (che contiene T), e perché questo sia possibile è necessario che almeno uno dei nodi

nel gadget variabile collegato a c_j sia contenuto in C , ossia uno dei nodi corrispondenti a una letterale nella clausola c_j deve essere contenuto in C .

Non resta che concludere la prova: mostrare che G è partizionabile in due strong web-communities se e soltanto se f è soddisfacibile.

- Se G è partizionabile in due strong web-communities C e $V - C$, allora T ed F non sono nella stessa comunità: sia $T \in C$ ed $F \in V - C$, allora, per ogni variabile x_i in X , esattamente uno dei nodi x_i e $\neg x_i$ deve essere contenuto in C ed esattamente uno dei nodi x_i e $\neg x_i$ deve essere contenuto in $V - C$; possiamo porre $\forall x_i \in X, a(x_i) = \text{vero} : x_i \in C$ e $\forall x_i \in X, a(x_i) = \text{falso} : x_i \in V - C$. Inoltre, per ogni clausola c_j , il nodo corrispettivo deve appartenere a C , e perché questo sia possibile è necessario che almeno uno dei nodi nei gadget variabile collegato a c_j sia contenuto in C , ossia uno dei nodi corrispondenti a una letterale nella clausola c_j deve essere contenuto in C : e quindi se $l_{jh} = x_i$ allora $x_i \in C$ ($\neg x_i \in V - C$) e $a(x_i) = \text{vero}$, mentre se $l_{jh} = \neg x_i$ allora $\neg x_i \in C$ ($x_i \in V - C$) e $a(x_i) = \text{falso}$. Quindi a è un'assegnazione di verità che soddisfa ogni clausola in f , ossia f è soddisfacibile.
- Se f è soddisfacibile, allora esiste un'assegnazione di verità a per X che soddisfa ogni clausola c_j in f . Costruiamo C inserendo i seguenti nodi:
 - Il nodo T .
 - Per ogni $j \in [m]$ i nodi c_j, l_{j1}, l_{j2} ed l_{j3} .
 - Per ogni $i \in [n]$ i nodi x_i, u_i, t_i e i nodi senza nome adiacenti a x_i tali che $a(x_i) = \text{vero}$.
 - Per ogni $i \in [n]$ i nodi $\neg x_i, u_i, t_i$ e i nodi senza nome adiacenti a $\neg x_i$ tali che $a(x_i) = \text{falso}$.

Possiamo subito notare che C e $V - C$ formano due comunità (ogni nodo che abbiamo inserito in C ha un numero maggiore di vicini in C che non in $V - C$ e viceversa). Vale la pena osservare il ruolo dei nodi senza nome: se x_i è stato collocato in $V - C$ e x_i compare in k clausole di f , poiché i k nodi c_j tali che $x_i \in c_j$ sono in C , allora k vicini di x_i sono in C ; allora, se non ci fossero i nodi senza nome x_i avrebbe k vicini in C e 1 vicino, u_i , in $V - C$, e quindi $V - C$ non sarebbe una comunità (caso analogo per $\neg x_i \in V - C$).

Capitolo 3

Dinamiche nelle reti

Capitolo 4

Istituzioni e comportamento aggregato

4.1 Il gioco delle urne

Poniamo di avere uno scommettitore, locato in una stanza chiusa, ed un gruppo di giocatori che attendono all'esterno di questa stanza. Lo scommettitore ha a disposizione 3 palle blu e 3 palline rosse, le quali vengono divise come segue:

- Un'urna conterrà 2 palline blu e 1 pallina rossa (chiameremo quest'urna MB , ossia "a maggioranza blu").
- L'altra conterrà 2 palline rosse e 1 pallina blu (chiameremo quest'urna MR , ossia "a maggioranza rossa").

A questo punto lo scommettitore sceglie casualmente una delle due urne uniformemente (la probabilità quindi di scelta dell'urna è $P(MB) = P(MR) = \frac{1}{2}$). A questo punto è possibile definire le regole del gioco:

- Un giocatore alla volta entrerà nella stanza, estrarrà una pallina (la quale può essere blu o rossa) e la reinerà nell'urna.
- Il giocatore non deve comunicare per alcuna motivazione il colore della pallina estratta (pena la squalifica di tutti quanti i giocatori).
- Il giocatore deve comunicare allo scommettitore e a tutti gli altri giocatori quale urna secondo lui ha davanti (ossia se l'urna è a maggioranza blu o a maggioranza rossa).

Al termine del gioco, solo chi ha indovinato quale urna è stata usata vince. Definite le regole del gioco, cerchiamo ora di capire in che modo risponderanno i vari giocatori:

- *Giocatore 1)* Prima di estrarre la pallina, l'unica informazione in suo possesso è che le urne MB ed MR sono equiprobabili. Dopo aver estratto una pallina, la probabilità che l'urna sia di un certo tipo verrà modificata: se estraesse una pallina blu (o rossa) il giocatore penserà che è più probabile che l'urna sia MB (o MR).
- *Giocatore 2)* Prima di estrarre la pallina, il giocatore può dedurre che il giocatore 1 ha estratto una pallina blu dato che ha risposto MB . A questo punto possono avvenire due casi:

- Se estraesse una pallina rossa si troverebbe nella stessa situazione in cui si trovava il giocatore 1 prima dell'estrazione (caso di equiprobabilità), e per tale motivo risponderebbe MR (in accordo con la pallina estratta); per tale motivo possiamo ignorare questo caso.
- Se estraesse una pallina blu saprebbe che sono state estratte due palline blu, rinforzando quindi l'ipotesi che l'urna sia MB : per tale motivo anche lui risponderà MB .
- *Giocatore 3)* Prima di estrarre la pallina, il giocatore può dedurre che il giocatore 1 e il giocatore 2 hanno estratto due palline blu in base alle loro risposte. Da qui abbiamo due casi interessanti:
 - Se estraesse una pallina blu saprebbe che finora sono state estratte tre palline blu, il che confermerebbe che l'urna che ha davanti è MB : anche lui pertanto risponderà MB .
 - Se estraesse una pallina rossa, la probabilità di avere davanti l'urna MR rimane più bassa rispetto a quella di avere davanti MB : per tale motivo, anche in questo caso risponderà MB .
- *Giocatore 4)* Prima di estrarre la pallina, il giocatore può trarre le stesse considerazioni del giocatore 3, ossia che il giocatore 1 e il giocatore 2 hanno estratto due palline blu, mentre non riesce a dedurre niente sul giocatore 3, se non che avrebbe risposto MB qualunque fosse stata la sua estrazione; avendo pertanto le stesse informazioni che aveva il giocatore 3, non potrà far altro che comportarsi esattamente come il giocatore 3, ossia risponderà MB .
- Tutti i giocatori dopo il giocatore quattro si comporteranno esattamente come il giocatore 4.

Possiamo notare che si è verificato un fenomeno a cascata, chiamata esattamente **cascata informatica**: infatti, qualunque saranno le estrazioni, ogni giocatore risponderà MB .

4.2 Il teorema di Bayes

Per capire se l'intuizione che ha guidato i giocatori sia fondata razionalmente, dobbiamo calcolare la probabilità di vittoria di ciascun giocatore, e per fare ciò necessitiamo del teorema di Bayes.

Teorema (4.1): Siano A, B due eventi, $P(A|B)P(B) = P(B|A)P(A)$.

Lemma (4.1a): Siano A, B due eventi, $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$.

Lemma (4.1b): Siano A, B due eventi, poiché

$$\begin{aligned} P(B) &= P((B \cap A) \cup (B \cap A^C)) = \\ &= P(B \cap A) + P(B \cap A^C) = \\ &= P(B|A)P(A) + P(B|A^C)P(A^C) \end{aligned}$$

abbiamo che
$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^C)P(A^C)}.$$

Possiamo usare il lemma (4.1b) per analizzare il gioco delle urne tramite i seguenti eventi. Definiamo in primis i seguenti eventi:

- MB = "l'urna è a maggioranza blu".
- MR = "l'urna è a maggioranza rossa".
- b = "viene estratta una pallina blu".
- r = "viene estratta una pallina rossa".

per capire se l'intuizione che ha guidato i giocatori sia fondata razionalmente dobbiamo considerare le seguenti probabilità:

- $P(MB) = P(MR) = \frac{1}{2}$.
- $P(b|MB) = \frac{2}{3}$ e $P(r|MB) = \frac{1}{3}$.
- $P(b|MR) = \frac{1}{3}$ e $P(r|MR) = \frac{2}{3}$.
- $MB^C = MR$ e $MR^C = MB$.

Da qui possiamo partire dall'analisi vera e propria, utilizzando le informazioni discusse prima:

- *Giocatore 1)* Poniamo che estrae una pallina blu:

$$\begin{aligned} P(MB|b) &= \frac{P(b|MB)P(MB)}{P(b|MB)P(MB) + P(b|MR)P(MR)} = \frac{\frac{2}{3} \cdot \frac{1}{2}}{\frac{2}{3} \cdot \frac{1}{2} + \frac{1}{3} \cdot \frac{1}{2}} = \frac{2}{3} \\ P(MR|b) &= \frac{P(b|MR)P(MR)}{P(b|MB)P(MB) + P(b|MR)P(MR)} = \frac{\frac{1}{3} \cdot \frac{1}{2}}{\frac{2}{3} \cdot \frac{1}{2} + \frac{1}{3} \cdot \frac{1}{2}} = \frac{1}{3} \end{aligned}$$

Dato che la probabilità di avere davanti l'urna MB è maggiore, il giocatore 1 risponderà MB .

- *Giocatore 2)* Lui sa che il giocatore 1 ha estratto una pallina blu. Se eastre una pallina rossa:

$$\begin{aligned} P(MB|br) &= \frac{P(br|MB)P(MB)}{P(br|MB)P(MB) + P(br|MR)P(MR)} = \frac{\frac{2}{3} \cdot \frac{1}{3} \cdot \frac{1}{2}}{\frac{2}{3} \cdot \frac{1}{3} \cdot \frac{1}{2} + \frac{1}{3} \cdot \frac{2}{3} \cdot \frac{1}{2}} = \frac{1}{2} \\ P(MR|br) &= \frac{P(br|MR)P(MR)}{P(br|MB)P(MB) + P(br|MR)P(MR)} = \frac{\frac{1}{3} \cdot \frac{2}{3} \cdot \frac{1}{2}}{\frac{2}{3} \cdot \frac{1}{3} \cdot \frac{1}{2} + \frac{1}{3} \cdot \frac{2}{3} \cdot \frac{1}{2}} = \frac{1}{2} \end{aligned}$$

Le probabilità di avere una delle due urne davanti è uguale, pertanto il giocatore 2 si baserebbe sulla pallina estratta per rispondere, pertanto la sua risposta sarebbe MR . Se invece estrae una pallina blu:

$$P(MB|bb) = \frac{P(bb|MB)P(MB)}{P(bb|MB)P(MB) + P(bb|MR)P(MR)} = \frac{\frac{2}{3} \cdot \frac{2}{3} \cdot \frac{1}{2}}{\frac{2}{3} \cdot \frac{2}{3} \cdot \frac{1}{2} + \frac{1}{3} \cdot \frac{1}{3} \cdot \frac{1}{2}} = \frac{4}{5}$$

$$P(MR|bb) = \frac{P(bb|MR)P(MR)}{P(bb|MB)P(MB) + P(bb|MR)P(MR)} = \frac{\frac{1}{3} \cdot \frac{1}{3} \cdot \frac{1}{2}}{\frac{2}{3} \cdot \frac{2}{3} \cdot \frac{1}{2} + \frac{1}{3} \cdot \frac{1}{3} \cdot \frac{1}{2}} = \frac{1}{5}$$

La probabilità di avere davanti l'urna MB è maggiore, e quindi il giocatore 1 risponderà MB .

- *Giocatore 3)* Lui sa che il giocatore 1 e il giocatore 2 hanno estratto due palline blue. Se eastrae una pallina rossa:

$$P(MB|bbr) = \frac{P(bbr|MB)P(MB)}{P(bbr|MB)P(MB) + P(bbr|MR)P(MR)} =$$

$$= \frac{\frac{2}{3} \cdot \frac{2}{3} \cdot \frac{1}{3} \cdot \frac{1}{2}}{\frac{2}{3} \cdot \frac{2}{3} \cdot \frac{1}{3} \cdot \frac{1}{2} + \frac{1}{3} \cdot \frac{1}{3} \cdot \frac{2}{3} \cdot \frac{1}{2}} = \frac{2}{3}$$

$$P(MR|bbr) = \frac{P(bbr|MR)P(MR)}{P(bbr|MB)P(MB) + P(bbr|MR)P(MR)} =$$

$$= \frac{\frac{1}{3} \cdot \frac{1}{3} \cdot \frac{2}{3} \cdot \frac{1}{2}}{\frac{2}{3} \cdot \frac{2}{3} \cdot \frac{1}{3} \cdot \frac{1}{2} + \frac{1}{3} \cdot \frac{1}{3} \cdot \frac{2}{3} \cdot \frac{1}{2}} = \frac{1}{3}$$

Poiché la probabilità di avere davanti l'urna MB è maggiore, il giocatore 1 risponderà MB . Se invece estrae una pallina blu:

$$P(MB|bbb) = \frac{P(bbb|MB)P(MB)}{P(bbb|MB)P(MB) + P(bbb|MR)P(MR)} =$$

$$= \frac{\frac{2}{3} \cdot \frac{2}{3} \cdot \frac{2}{3} \cdot \frac{1}{2}}{\frac{2}{3} \cdot \frac{2}{3} \cdot \frac{2}{3} \cdot \frac{1}{2} + \frac{1}{3} \cdot \frac{1}{3} \cdot \frac{1}{3} \cdot \frac{1}{2}} = \frac{8}{9}$$

$$P(MR|bbb) = \frac{P(bbb|MR)P(MR)}{P(bbb|MB)P(MB) + P(bbb|MR)P(MR)} =$$

$$= \frac{\frac{1}{3} \cdot \frac{1}{3} \cdot \frac{1}{3} \cdot \frac{1}{2}}{\frac{2}{3} \cdot \frac{2}{3} \cdot \frac{2}{3} \cdot \frac{1}{2} + \frac{1}{3} \cdot \frac{1}{3} \cdot \frac{1}{3} \cdot \frac{1}{2}} = \frac{1}{9}$$

Anche in questo caso la probabilità di avere davanti l'urna MB è maggiore, e pertanto il giocatore 1 risponderà MB . Da qui notiamo che qualunque pallina estragga il giocatore 3, le probabilità di avere davanti l'urna MB rimangono alte (almeno superiori al 50%).

- *Giocatore 4)* Lui sa che il giocatore 1 e il giocatore 2 hanno estratto due palline blu,

mentre non sa dire niente sul giocatore 3. Se eastre una pallina rossa:

$$\begin{aligned}
P(MB|bbbr) &= \frac{P(bbbr|MB)P(MB)}{P(bbbr|MB)P(MB) + P(bbbr|MR)P(MR)} = \frac{8}{9} \\
P(MB|bbrr) &= \frac{P(bbrr|MB)P(MB)}{P(bbrr|MB)P(MB) + P(bbrr|MR)P(MR)} = \frac{1}{2} \\
P(MR|bbbr) &= \frac{P(bbbr|MR)P(MR)}{P(bbbr|MB)P(MB) + P(bbbr|MR)P(MR)} = \frac{1}{9} \\
P(MR|bbrr) &= \frac{P(bbrr|MR)P(MR)}{P(bbrr|MB)P(MB) + P(bbrr|MR)P(MR)} = \frac{1}{2}
\end{aligned}$$

Se invece estraee una pallina blu:

$$\begin{aligned}
P(MB|bbbb) &= \frac{P(bbbb|MB)P(MB)}{P(bbbb|MB)P(MB) + P(bbbb|MR)P(MR)} = \frac{16}{17} \\
P(MB|bbrb) &= \frac{P(bbrb|MB)P(MB)}{P(bbrb|MB)P(MB) + P(bbrb|MR)P(MR)} = \frac{8}{9} \\
P(MR|bbbb) &= \frac{P(bbbb|MR)P(MR)}{P(bbbb|MB)P(MB) + P(bbbb|MR)P(MR)} = \frac{1}{17} \\
P(MR|bbrb) &= \frac{P(bbrb|MR)P(MR)}{P(bbrb|MB)P(MB) + P(bbrb|MR)P(MR)} = \frac{1}{9}
\end{aligned}$$

Qui possiamo notare un caso interessante: seppur esista un caso in cui le due urne sono equiprobabili, la probabilità di avere davanti l'urna MB rimane comunque alta (almeno del 50%).

- *Giocatore 5*) Lui sa che il giocatore 1 e il giocatore 2 hanno estratto due palline blu, mentre non sa dire niente sul giocatore 3 e sul giocatore 4. Se eastre una pallina rossa:

$$\begin{aligned}
P(MB|bbbbrr) &= \frac{8}{9} \\
P(MB|bbbrr) &= P(MB|bbrbr) = \frac{2}{3} \\
P(MB|bbrrrr) &= \frac{1}{2} \\
P(MR|bbbbrr) &= \frac{1}{9} \\
P(MR|bbbrr) &= P(MR|bbrbr) = \frac{1}{3} \\
P(MR|bbrrrr) &= \frac{1}{2}
\end{aligned}$$

Se invece estrae una pallina blu:

$$P(MB|bbbbbb) = \frac{32}{33}$$

$$P(MB|bbbrb) = P(MB|bbrbb) = \frac{8}{9}$$

$$P(MB|bbrrb) = \frac{2}{3}$$

$$P(MR|bbbbbb) = \frac{1}{33}$$

$$P(MR|bbbrb) = P(MR|bbrbb) = \frac{1}{9}$$

$$P(MR|bbrrb) = \frac{1}{3}$$

Anche qui possiamo notare un caso interessante: seppur esista un caso in cui le due urne sono equiprobabili, la probabilità di avere davanti l'urna MB rimane comunque alta (almeno del 50%).

Una cosa che si evince dall'analisi appena fatta è che nel caso il giocatore 1 e il giocatore 2 abbiano estratto due palline blu dall'urna MR , rispondendo però MB , tutti i giocatori dopo si ritrovano a dare la stessa risposta sbagliata.

Capitolo 5

Reti di Informazione: il World Wide Web