Brevi note sul modello di Erdős-Rényi

Giuseppe Benfatto
Università di Roma "Tor Vergata"

a. a. 2007-08

1 Il modello di Erdős-Rényi

Sia G_n la famiglia dei grafi semplici (cioè senza loop e con al più una linea per ogni coppia di vertici) di ordine n. Il modello di Erdős-Rényi è una misura di probabilità P su G_n, dipendente da un parametro $p \in [0,1]$, definita così che, per ogni $g \in G_n$,

$$P(g) = p^l (1-p)^{N-l} \tag{1.1}$$

dove l indica il numero di linee di g e $N = n(n-1)/2$ è il numero di linee del grafo completo K_n. Si noti che P si può pensare come il prodotto di N misure di probabilità, una per ognuna delle linee di K_n; questa misura assegna probabilità p alla presenza della linea e $1-p$ alla sua assenza. È pertanto evidente che P è una misura di probabilità, ma ciò può essere verificato facilmente anche notando che il numero di grafi con l linee è $C_{N,l}$, il numero di combinazioni di N elementi a l a l, per cui $\sum_{g\in G_n} P(g) = \sum_{l=0}^{N} C_{N,l} p^l (1-p)^{N-l} = [1+(1-p)]^N = 1$, grazie alla Formula del binomio di Newton. L'interpretazione di P come misura prodotto rende anche facile dedurre che il valor medio di l/N è eguale a p, indipendentemente da N (quindi anche da n).

Un tipo di problema che è interessante studiare è il seguente. Supponiamo di scegliere p in funzione di n, $p = p(n)$, e sia Q una qualche proprietà dei grafi, per esempio la proprietà di essere connesso, e sia $P_n(Q)$ la probabilità che Q sia verificata. Ci si chiede se esiste una *soglia critica* $p_c(n)$ tale che

$$\lim_{n\to\infty} P_n(Q) = \begin{cases} 0 & \text{se} \quad \lim_{n\to\infty} p(n)/p_c(n) = 0 \\ 1 & \text{se} \quad \lim_{n\to\infty} p(n)/p_c(n) = \infty \end{cases} \tag{1.2}$$

Si noti che, se esiste una tale soglia, ci si aspetta che, fissato n "abbastanza grande", al crescere di p da 0 a 1 si passi in modo "brusco" da una situazione in cui Q non è "quasi mai" verificata ad una in cui lo è "quasi sempre".

1.1 Soglie per la comparsa di sottografici particolari

Sia F un grafo fissato con k vertici e l linee, sia K_n il grafo completo di ordine n e sia $\{F_\alpha\}_{\alpha\in\mathcal{I}_F}$ la famiglia di tutti i sottografi di K_n isomorfi a F (due grafi si dicono isomorfi se hanno lo stesso numero di vertici e di linee e se esiste una corrispondenza biunivoca fra i vertici rispettivi, tale che due vertici dell'uno sono connessi da una linea se e solo se lo sono i vertici corrispondenti dell'altro). Dato $g \in G_n$, indicheremo con $F_\alpha \subset g$ il fatto che F_α è un sottografo di g e useremo la definizione:

$$Y_\alpha(g) = \begin{cases} 1 & \text{se } F_\alpha \subset g \\ 0 & \text{altrimenti} \end{cases} \tag{1.3}$$

Si noti che, se si indica con $\mathbb{E}(f)$ il valor medio di una funzione $f(g)$ rispetto alla misura P, per la (1.1),

$$P(F_\alpha \subset g) = \sum_{g:F_\alpha\subset g} P(g) = \mathbb{E}(Y_\alpha) = p^l \tag{1.4}$$

Indichiamo ora con $N_F(g)$ il numero di sottografi di g isomorfi a F. Per la (1.3)

$$N_F(g) = \sum_{\alpha\in\mathcal{I}_F} Y_\alpha(g) \tag{1.5}$$

D'altra parte, per la (1.4), $\mathbb{E}(Y_\alpha)$ non dipende da α; pertanto $\mathbb{E}(N_F) = p_l|\mathcal{I}_F|$, dove $|\mathcal{I}_F|$ è il numero di sottografi di K_n isomorfi a F. Questo numero è ovviamente eguale al prodotto di $C_{n,k}$, il numero di modi diversi di scegliere k vertici fra gli n vertici di K_n, per il numero k_F di modi diversi di formare un grafico isomorfo a F, usando k vertici fissati.

Si noti che, per calcolare k_F, bisogna considerare tutte le permutazioni dei k vertici fissati ed associare ad ognuna di esse una diversa corrispondenza biunivoca con i vertici di F. Può tuttavia succedere che si ottenga più volte uno stesso sottografico. Si pensi, per esempio, al caso in cui F è un albero con tre vertici; in tal caso, se F_1 è isomorfo a F, il grafico ottenuto da F_1 scambiando i suoi due vertici di grado 1 è ancora isomorfo a F, ma coincide con F_1. In generale è complicato calcolare k_F, ma si può sempre affermare che $k_F \leq k!$.

Si noti anche che

$$C_{n,k} = \frac{n(n-1)\cdots(n-k+1)}{k!} = n^k[1 + O(n^{-1})] \tag{1.6}$$

In conclusione,

$$\mathbb{E}(N_F) = C_{n,k}k_F p^l = \frac{k_F}{k!}(pn^{k/l})^l[1 + O(n^{-1})] \tag{1.7}$$

Di qui segue subito che

$$\lim_{n\to\infty} \mathbb{E}(N_F) = \begin{cases} 0 & \text{se } \lim_{n\to\infty} p(n)n^{k/l} = 0 \\ +\infty & \text{se } \lim_{n\to\infty} p(n)n^{k/l} = \infty \end{cases} \tag{1.8}$$

La (1.8) suggerisce che $n^{-k/l}$ sia una soglia critica per la proprietà di un grafico di contenere sottografi isomorfi a F. Tuttavia, non è facile verificare se ciò sia vero o falso, a meno che non si richieda che F sia un *grafo bilanciato*, cioè un grafo tale che ogni suo sottografo con k' vertici e l' linee soddisfa la condizione

$$l'/k' \leq l/k \tag{1.9}$$

Questa proprietà non è molto restrittiva, in quanto è verificata, per esempio, dagli alberi, dai cicli e dai grafi completi. Dimostriamo pertanto il Teorema seguente.

Teorema 1.1 *Se F è un grafo bilanciato con k vertici e l linee e Q_F è l'insieme dei grafi di ordine n che contengono almeno un sottografo isomorfo a F, allora*

$$\lim_{n\to\infty} P_n(Q_F) = \begin{cases} 0 & \textit{se } \lim_{n\to\infty} p(n)n^{k/l} = 0 \\ 1 & \textit{se } \lim_{n\to\infty} p(n)n^{k/l} = \infty \end{cases} \tag{1.10}$$

Dim. - Il fatto che $\lim_{n\to\infty} P_n(Q_F) = 0$ se $\lim_{n\to\infty} p(n)n^{k/l} = 0$ segue subito dalla (1.8), in quanto

$$P_n(Q_F) = \sum_{m\geq 1} P_n(N_F = m) \leq \sum_{m\geq 1} mP_n(N_F = m) \leq \mathbb{E}(N_F) \tag{1.11}$$

Poniamo ora $\gamma_n = p(n)n^{k/l}$ e supponiamo che $\gamma_n \to \infty$ per $n \to \infty$. Dimostrare che $P_n(Q_F) \to 1$ è equivalente a dimostrare che $P_n(N_F = 0) \to 0$; d'altra parte, posto $\mu = \mathbb{E}(N_F)$,

$$P_n(N_F = 0) \leq P_n(|N_F - \mu| \geq \mu) \tag{1.12}$$

Si osservi ora che, data una qualunque funzione $f(g)$ ed un numero positivo c, se si indica con Q il sottoinsieme di G_n in cui è soddisfatta la condizione $(|f(g)| \geq c)$ e con $Y_Q(g)$ la *funzione indicatrice* di Q, cioè la funzione eguale a 1 se $g \in Q$ e eguale a 0 in caso contrario, allora

$$P_n(Q) = \sum_{g\in G_n} P_n(g)Y_Q(g) \leq \sum_{g\in G_n} P_n(g)Y_Q(g)\frac{f(g)^2}{c^2} \tag{1.13}$$

in quanto, se $Y_Q(g) > 0$, allora $f(g)^2/c^2 \geq 1$. Poiché inoltre $Y_q(g) \leq 1$, si ha infine la cosiddetta *diseguaglianza di Chebyshev*

$$P_n(Q) \leq \sum_{g\in G_n} P_n(g)\frac{f(g)^2}{c^2} = \frac{\mathbb{E}(f^2)}{c^2} \tag{1.14}$$

Pertanto, se si applica la (1.14) al secondo membro della (1.12), si trova

$$P_n(N_F = 0) \leq \frac{1}{\mu^2}\mathbb{E}((N_F - \mu)^2) = \frac{\mathbb{E}(N_F^2)}{\mu^2} - 1 \tag{1.15}$$

avendo anche usato il fatto che $\mathbb{E}((N_F - \mu)^2) = \mathbb{E}(N_F^2) - \mu^2$.

Per la (1.5), se si indica con $V_{\alpha,\beta}$ l'insieme dei vertici in comune a F_α e F_β, si ha

$$\mathbb{E}(N_F^2) = \sum_{\alpha,\beta} \mathbb{E}(Y_\alpha Y_\beta) = \sum_{s=0}^{k} \Pi(s) \tag{1.16}$$

avendo definito

$$\Pi(s) = \sum_{\alpha,\beta:V_{\alpha,\beta}=s} \mathbb{E}(Y_\alpha Y_\beta) \tag{1.17}$$

Se $s = 0$, $\mathbb{E}(Y_\alpha Y_\beta) = \mathbb{E}(Y_\alpha)\mathbb{E}(Y_\beta)$, in quanto le condizioni $(F_\alpha \subset g)$ e $(F_\beta \subset g)$ coinvolgono linee diverse e P_n è costruita come prodotto di misure sulle singole linee di K_n; pertanto

$$\Pi(0) = \sum_{\alpha,\beta:V_{\alpha,\beta}=0} \mathbb{E}(Y_\alpha)\mathbb{E}(Y_\beta) \leq \sum_{\alpha,\beta} \mathbb{E}(Y_\alpha)\mathbb{E}(Y_\beta) = \mu^2 \tag{1.18}$$

e quindi, per la (1.12),

$$P_n(N_F = 0) \leq \frac{1}{\mu^2} \sum_{s=1}^{k} \Pi(s) \tag{1.19}$$

Se $s > 0$, per stimare $\Pi(s)$ si può osservare che, fissato F_α e il sottoinsieme $V_{\alpha,\beta}$ dei suoi s vertici in comune con F_β, viene automaticamente individuato un suo sottografo (formato dalle linee di F_α che uniscono coppie di vertici di $V_{\alpha,\beta}$), che, per la (1.9), ha un numero di linee $t \leq ls/k$; ne segue che F_β deve avere un numero di linee non già contenute in F_α almeno eguale a $l - t$. Pertanto, se si indica con $l_{\alpha,\beta}$ il numero di linee contenute nel grafo $F_\alpha \bigcup F_\beta$, $\mathbb{E}(Y_\alpha Y_\beta) = p^{l_{\alpha,\beta}} \leq p^l p^{l-t}$, in quanto $p < 1$ e $l_{\alpha,\beta} \geq l + l - t$. Rimane da stimare il numero di coppie di grafici isomorfi a F, che hanno s vertici in comune e valore fissato di t. F_α può essere scelto, come nel calcolo del valor medio, in $C_{n,k}k_F$ modi diversi. Dato F_α, ci sono $C_{k,s}$ modi di scegliere i vertici di F_β in comune con F_α, $C_{n-k,k-s}$ modi di scegliere i vertici rimanenti e al più $k!$ modi di ottenere un grafo diverso permutando i vertici. Poiché $C_{n,k}k_F p^l = \mu$, ne segue che

$$\Pi(s) \leq \mu \sum_{t=0}^{ls/k} C_{k,s}C_{n-k,k-s}k!p^{l-t} \tag{1.20}$$

Se si usa la (1.6) e si pone $p = \gamma_n n^{-k/l}$, si vede subito che esiste una costante c_k, dipendente solo da k, tale che

$$\Pi(s) \leq c_k \mu \sum_{t=0}^{ls/k} n^{k-s}\gamma_n^{l-t}n^{-k+kt/l} = c_k \mu \gamma_n^l \sum_{t=0}^{ls/k} \gamma_n^{-t} n^{+kt/l-s} \tag{1.21}$$

D'altra parte, per la (1.7), se n è abbastanza grande, $\gamma_n^l \leq 2k!\mu$; pertanto

$$\Pi(s) \leq 2k! c_k \mu^2 \sum_{t=0}^{ls/k} \gamma_n^{-t} n^{+kt/l-s} \tag{1.22}$$

L'esponente di n in questa stima è ≤ 0 per ogni t (qui si usa in modo decisivo il fatto che il grafo F è bilanciato) e negativo per $t = 0$ (in quanto $s > 0$). Pertanto, se $\gamma_n \to \infty$, $\Pi(s)/\mu^2 \to 0$ e ciò è allora vero anche per $P_n(N_F = 0)$, per la (1.19), il che completa la dimostrazione. ∎

1.2 Distribuzione del grado

Dato un grafo $g \in G_n$, indichiamo con $d_i(g)$ il grado del vertice i. Se si pone $Y_{i,j}(g) = 1$ se la linea (i, j) appartiene a g e $Y_{i,j}(g) = 0$ altrimenti, allora

$$d_i = \sum_{j \neq i} Y_{i,j} \Rightarrow \mathbb{E}(d_i) = \sum_{j \neq i} \mathbb{E}(Y_{i,j}) = p(n-1) \tag{1.23}$$

Inoltre. se si definisce

$$I_{i,k}(g) = \begin{cases} 1 & \text{se } \ d_i(g) = k \\ 0 & \text{altrimenti} \end{cases} \tag{1.24}$$

allora la *distribuzione del grado* è data dalla funzione

$$\rho(k) = P(d_i = k) = \mathbb{E}(I_{i,k}) = \sum_{g: d_i(g) = k} P(g) = C_{n-1,k} p^k (1-p)^{n-1-k} \tag{1.25}$$

Supponiamo ora che

$$p = \frac{\lambda}{n} \quad , \quad \lambda > 0 \tag{1.26}$$

In tal caso $\mathbb{E}(d_i)$ e $\rho(k)$ sono praticamente indipendenti da n. Infatti $\lim_{n\to\infty} \mathbb{E}(d_i) = \lambda$ e, per la (1.6), se $n \to \infty$,

$$\rho(k) = \frac{\lambda^k}{k!} \left(1 - \frac{\lambda}{n}\right)^{n-1-k} [1 + O(n^{-1})] \to \rho^*(k) \equiv \frac{\lambda^k}{k!} e^{-\lambda} \tag{1.27}$$

e si può vedere facilmente che la convergenza è uniforme in k. Si noti che $\rho^*(k)$ è la ben nota *distribuzione di Poisson*.

Vogliamo ora far vedere che la distribuzione del grado è praticamente coincidente con la funzione che rappresenta, dato un grafo qualunque g, la frequenza dei vertici di grado fissato. Indichiamo pertanto con $n_k(g)$ il numero di vertici di grado k presenti in g. Si ha

$$n_k(g) = \sum_{i=1}^{n} I_{i,k}(g) \tag{1.28}$$

Quindi, per la (1.25),

$$\lambda_k \equiv \mathbb{E}(n_k) = n\rho(k) \tag{1.29}$$

Consideriamo ora la funzione

$$\tilde{\rho}(k,g) = \frac{1}{n} n_k(g) \tag{1.30}$$

Pensata come funzione di k, con g fissato, $\tilde{\rho}(k,g)$ rappresenta la *frequenza* dei vertici di grado k. Il suo valor medio, per la (1.29), è eguale alla funzione $\rho(k)$. Vogliamo far vedere che, se si pone $p = \lambda/n$, allora le fluttuazioni di $\tilde{\rho}(k,g)$ intorno a $\rho(k)$ sono "trascurabili per n grande". Più precisamente, dimostreremo il seguente teorema.

Teorema 1.2 *Se $p = \lambda/n$, $\lambda > 0$, allora, $\forall \varepsilon > 0$,*

$$\lim_{n\to\infty} P_n(|\tilde{\rho}(k,g) - \rho(k)| > \varepsilon) = 0 \tag{1.31}$$

Dim. - Usando la diseguaglianza di Chebyshev (1.14) ed il fatto che $\mathbb{E}(\tilde{\rho}(k)) = \rho(k)$, per cui $\mathbb{E}([\tilde{\rho}(k) - \rho(k)]^2) = \mathbb{E}(\tilde{\rho}(k)^2) - \rho(k)^2$, si ha

$$P_n(|\tilde{\rho}(k,g) - \rho(k)| > \varepsilon) \leq \varepsilon^{-2} \left[\frac{\mathbb{E}(n_k^2)}{n^2} - \rho(k)^2 \right] \tag{1.32}$$

D'altra parte, per la (1.28), $\mathbb{E}(n_k^2) = \sum_{i,j} \mathbb{E}(I_{i,k} I_{j,k})$. Inoltre, se $i = j$, poiché $I_{i,k}^2 = I_{i,k}$, $\mathbb{E}(I_{i,k}^2) = \mathbb{E}(I_{i,k}) = \rho(k)$ e, se $i \neq j$, $\mathbb{E}(I_{i,k} I_{j,k}) = \mathbb{E}(I_{1,k} I_{2,k})$, in quanto il modello è invariante rispetto alle permutazioni dei vertici. Ne segue che

$$\mathbb{E}(n_k^2) = n\rho(k) + n(n-1)\mathbb{E}(I_{1,k} I_{2,k}) \tag{1.33}$$

Per calcolare $\mathbb{E}(I_{1,k} I_{2,k})$, si osservi che l'insieme Q in cui $I_{1,k} I_{2,k} = 1$ si può dividere in due insiemi disgiunti, l'insieme Q_0 dei grafi in cui le k linee ancorate al vertice 1 sono tutte diverse dalle k linee ancorate al vertice 2 e l'insieme Q_1 dei grafi in cui i due suddetti insiemi di linee hanno in comune la linea che li congiunge. Un facile conteggio permette inoltre di verificare che

$$\begin{aligned} P_n(Q_0) &= (1-p) \left[C_{n-2,k} p^k (1-p)^{n-2-k} \right]^2 \\ P_n(Q_1) &= p \left[C_{n-2,k-1} p^{k-1} (1-p)^{n-2-(k-1)} \right]^2 \end{aligned} \tag{1.34}$$

Di qui, usando la (1.25), segue che

$$\mathbb{E}(I_{1,k} I_{2,k}) = P_n(Q_0) + P_n(Q_1) = \rho(k)^2 \left[\frac{C_{n-2,k}^2}{(1-p) C_{n-1,k}^2} + \frac{C_{n-2,k-1}^2}{p C_{n-1,k}^2} \right] \tag{1.35}$$

Se si usa la (1.6), si vede allora che

$$\mathbb{E}(I_{1,k} I_{2,k}) = \rho(k)^2 \left[\frac{1}{1-p} + \frac{1}{pn^2} \right] [1 + O(n^{-1})] \tag{1.36}$$

Si ponga ora $p = \lambda/n$; allora la (1.33) e la (1.36) implicano che

$$\frac{1}{n^2}\mathbb{E}(n_k^2) = \frac{\rho(k)}{n} + \rho(k)^2[1 + O(n^{-1})] \tag{1.37}$$

Pertanto il secondo membro della (1.32) va a 0, per $n \to \infty$ (e si può vedere facilmente che la convergenza è uniforme in k). Ciò completa la dimostrazione del teorema. ∎

1.3 Coefficiente di clustering

Dato un grafo $g \in G_n$, il *coefficiente di clustering* $C_i(g)$ *del vertice* i è definito, se $d_i(g) \geq 2$, come il rapporto fra le l_i linee che uniscono fra di loro i d_i vertici connessi a i da una linea ed il numero massimo $n_i = d_i(d_i - 1)/2$ di tali linee (cioè quelle del grafo completo formato a partire dai suddetti vertici). Il *coefficiente di clustering medio* $\bar{C}_i$ *del vertice* i è invece la media di $C_i(g)$ fatta sui grafi in cui $d_i(g) \geq 2$, cioè

$$\bar{C}_i = \mathbb{E}(C_i) = \frac{\sum_{k=2}^{n-1} \rho(k) \frac{1}{n_i} \sum_{l=0}^{n_i} l\ p^l(1-p)^{n_i - l}}{\sum_{k=2}^{n-1} \rho(k)} = p \tag{1.38}$$

dove si è usato il fatto che, per ogni $m > 0$, $\frac{1}{m}\sum_{l=0}^{m} l\ p^l(1-p)^{m-l} = p$, vedi osservazione nel paragrafo che segue la (1.1).

Un'altra grandezza che viene spesso studiata è il *coefficiente di clustering medio del grafo*, cioè la funzione

$$\bar{C}(g) = \frac{1}{n}\sum_{i=1}^{n} C_i(g) \tag{1.39}$$

La (1.38) implica immediatamente che $\mathbb{E}(\bar{C}) = p$

Procedendo come nel §1.2, si potrebbe anche dimostrare che, se $p = \lambda/n$, con $\lambda > 0$, allora $\lim_{n\to\infty} P_n(|\bar{C}(g) - p| > \varepsilon) = 0$, per ogni $\varepsilon > 0$.