Grafi Erdos-Renyi

Lo studio delle reti mira a costruire modelli che riproducono le proprietà delle reti reali. A tal fine introduciamo un modello per la generazione di grafi casuali.

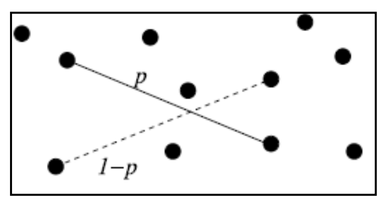
Dal punto di vista del modello, una rete è un oggetto relativamente semplice, composto da soli nodi e collegamenti. La difficoltà sta nel decidere dove posizionare i collegamenti tra i nodi in modo da riprodurre la complessità di un sistema reale. A questo proposito la filosofia alla base di una rete casuale è semplice: l'obiettivo è realizzato al meglio disponendo i collegamenti in modo casuale tra i nodi.

Considereremo il modello per la generazione di grafi casuali che prende il nome da Paul Erdős e Alfréd Rényi, che per primi lo hanno introdotto nel 1959.

Il modello produce un grafo non orientato i cui nodi sono collegati in maniera casuale.

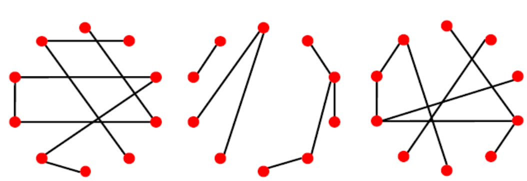
Definizione. Il grafo Erdos-Renyi $G_{n,p}$ con parametri *n* e *p*

- ha *n* vertici e
- per ogni coppia di vertici esiste un collegamento con probabilità *p* e non esiste con probabilità *1- p*.



Si noti che *n* e *p* non determinano univocamente il grafo che è il risultato di un processo casuale. E' quindi possibile avere molte realizzazioni diverse dati gli stessi valori di *n* e *p*

I grafi nel seguente esempio sono riportate tre diverse realizzazioni ottenute usando gli stessi parametri: 10 nodi ed una probabilità p=1/6

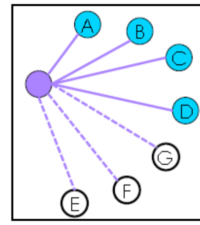


n = 10
p= 1/6

Grado di un nodo. Ogni nodo in $G_{n,p}$ ha (n - 1) *potenziali* vicini. Vogliamo valutare la probabilità che un nodo abbia *k* vicini, per *k=0,…n-1.*

Dalla definizione di $G_{n,p}$, per ogni possibile edge si lancia una moneta truccata e

- con probabilità *p* l'edge viene inserito
- con probabilità *1-p* l'edge non viene inserito.



Esempio. Sia *n=8,* vogliamo valutare la probabilità che il nodo viola ha grado 4. Indichiamo con A,B,…,G i suoi 7 potenziali vicini e supponiamo che essi siano marcati neri o blu: i nodi blu sono quelli con cui il nodo viola condivide un edge, quelli neri sono nodi con cui il nodo viola non è unito da un edge. In quanti modi diversi riusciamo a colorare di blu 4 dei 7 nodi? Mostriamo che sono

$$\binom{7}{4}=7!/(3!4!).$$

Come prima cosa notiamo che possiamo scegliere il primo nodo blu tra tutti i 7 possibili, il secondo tra i rimanenti 6, il terzo tra i rimanenti 5, il quarto tra 4. Quindi abbiamo 7x6x5x4 possibili scelte di nodi blu (i nodi non scelti rimangono neri). Notiamo poi che tutte le scelte con gli stessi 4 nodi blu in ordine diverso sono equivalenti. Per ogni gruppo di 4 esistono 4! permutazioni ; ciò implica che il numero di possibilità *diverse* è

$$(7x6x5x4)/4!= ((7x6x5x4x3x2x1)/(3x2x1))/ 4!=7!/(3!\ 4!)$$

Fissati i 4 nodi A,B,C,D, valutiamo la probabilità che il nodo viola V ha come vicini i nodi blu **A, B, C, D** e non ha come vicini E,F,G.

Indichiamo con $Pr(\exists\, V-X)$ la probabilità che esista l'edge tra V e X e con $Pr(\nexists\, V-X)$ la probabilità che l'edge tra V e X non esiste nel grafo. Poichè l'inserimento di ogni edge è indipendente da quello di ogni altro edge, otteniamo che la probabilità cercata è

$$\Pr(\exists\, V-A)Pr(\exists\, V-B)Pr(\exists\, V-C)Pr(\exists\, V-D)\,Pr(\nexists\, V-E)\,Pr(\nexists\, V-F)Pr(\nexists\, V-G)$$
$$=\ p^4(1-p)^3$$

Sapendo che vi sono $\binom{7}{4}$ possibili scelte di nodi blu abbiamo che la probabilità di avere esattamente 4 vicini è

$$\binom{7}{4}p^4(1-p)^3.$$

Calcoliamo ora in generale, la probabilità che un nodo in $G_{n,p}$ abbia k vicini. Sappiamo che ogni nodo ha (n - 1) potenziali vicini e che Il numero di scelte di k vicini tra tali n-1 possibili sono

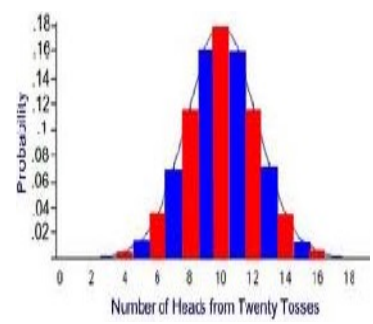
$$\binom{n-1}{k} = \frac{(n-1)!}{k!(n-1-k)!} = \frac{numero\ di\ permutazioni\ (n-1)\ oggetti}{(numero\ permutazioni\ k\ oggetti)} \frac{(n-1-k)!}{(n-1-k)!}$$

Poiché ogni edge ad uno degli altri nodi esiste con probabilità p

(e non esiste con probabilità 1-p) otteniamo che la probabilità che

un nodo abbia grado k è

$$p_k = \binom{n-1}{k} p^k (1-p)^{n-1-k} \qquad \textbf{(1)}$$



In accordo alla (1), abbiamo che :

- la distribuzione dei gradi risulta binomiale.

- Il *grado medio* è

$$\begin{aligned}\overline{k} &= \sum_{k=0}^{n-1} k\ p_k = \sum_{k=1}^{n-1} k \binom{n-1}{k} p^k(1-p)^{n-1-k} \\ &= \sum_{k=1}^{n-1}(n-1)\binom{n-2}{k-1}p^k(1-p)^{n-1-k} \\ &= (n-1)\, p \sum_{i=0}^{n-2}\binom{n-2}{i}p^i(1-p)^{n-2-i} \\ &= (n-1)\, p\ [p+(1-p)]^{n-2} \\ &= (n-1)p\end{aligned}$$

La varianza* è $\sigma^2 = p(1-p)(n-1)$

Notiamo che il grado sarà sempre *più prossimo alla media* in reti sempre più grandi, infatti

$$\frac{\sigma}{\overline{k}} = \left[\frac{1-p}{p}\frac{1}{(n-1)}\right]^{1/2} \approx \frac{1}{(n-1)^{1/2}}$$

----

*Fornisce una misura di quanto i valori assunti da una variabile casuale si discostino quadraticamente rispettivamente dal valore medio

Evoluzione dei grafi Erdos-Renyi: Cosa succede quando varia p?

Consideriamo un ricevimento in cui inizialmente poche persone si conoscono tra loro, poi iniziano a presentarsi l'un l'altra. Man mano sempre più persone sono unite da un legame di conoscenza.

Si verifica un processo dinamico: a partire da N nodi isolati, i collegamenti vengono aggiunti gradualmente attraverso incontri casuali tra gli ospiti. Ciò corrisponde ad un aumento graduale di *p*, con conseguenze notevoli sulla topologia della rete

Per quantificare questo processo, studiamo tra le altre cose come la dimensione della componente connessa all'interno della rete varia con *p* (e quindi con il grado medio).

Consideriamo come prima cosa i casi estremi:

- per p = 0 il grado medio è pari a 0, quindi tutti i nodi sono isolati;
- per p = 1 si ha grado medio pari N-1, e la rete è un grafo completo

Connettività. Quanto deve essere grande *p* per non avere nodi isolati?

Sia $p=d/(n-1)$ (quindi il grado medio è x).

Valutiamo la probabilità che un nodo v sia isolato. Abbiamo,

$$Pr(v \text{ ha grado } 0)=(1-p)^{n-1}=(1-d/(n-1)\ )^{n-1}\approx (1/e^{d}),$$

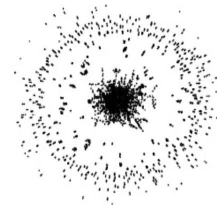
(notiamo che la funzione $f(x)=(1-d/x)^x$ è crescente e tende rapidamente al valore limite e^{-d})

Se indichiamo con $v_1, v_2, \ldots, v_n$ i nodi del grafo, abbiamo che la probabilità che *almeno* un nodo sia isolato è

$$P(v_1 \text{ isolato OR } v_2 \text{ isolato OR} \ldots \text{OR } v_n \text{ isolato}) \approx \sum_{i=1}^{n} P(v_i \text{ isolato}) \approx n(1/e^{d})$$

- Se poniamo $d=\ln n$ otteniamo $(n/e^{d})=1$, quindi per *p* minore o uguale a $(\ln n)/(n-1)$ avremo sicuramente dei nodi isolati
- Se poniamo $d=2\ln n$ otteniamo $(n/e^{d})=1/n \rightarrow 0$, per $n \rightarrow \infty$; quindi in reti grandi, se *p* è almeno $(2 \ln n)/(n-1)$ non vi saranno nodi isolati.

Giant component. Due domande fondamentali riguardano la dimensione della componente gigante ed il valore di p che ne provoca l'esistenza.



La differenza qualitativa fondamentale dipende dalla *dimensione* della componente gigante come frazione del numero totale n di nodi, in particolare dal *tendere a 0 per* $n \to \infty$, *oppure tendere a qualche valore finito tra 0 e 1.*

Fissata la probabilità *p*, vogliamo stabilire come varia la dimensione della componente gigante come funzione di *n*.

Sia *X* la variabile casuale che conta il numero di vertici della più grande componente connessa. Dimostriamo il seguente teorema.

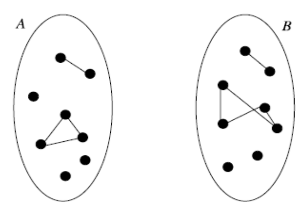
Teorema (Erdos-Rényi, 1961) Supponendo *p> (ln 64/n)* il grafo G_{np} ha una componente gigante con alta probabilità, cioè

$$\Pr[X > n/2] \geq 1 - 2^{-n/8}.$$

Ricordiamo che *ln 64 ≈4,158*

Per dimostrare il teorema useremo il seguente risultato.

Fatto. Se *X <n/2* (cioè, la componente più grande ha dimensione al più n/2) allora esiste una partizione dei vertici in due insiemi *A, B* tale che *n/4 < IAI, IBI <3n/4* e nessun edge esiste tra *A* e *B*.



Dimostrazione. Siano n_1, n_2, …, n_k le dimensioni delle componenti connesse del grafo. Osserviamo che n_1+ n_2+ …+ n_k=n.

Per ipotesi *X<n/2*, cioè, ogni componente ha dimensione *<n/2*. Pertanto n_i*<n/2* per tutti *i = 1, ... , k*. Mostriamo che possiamo trovare un indice *m <k* tale che

$$\frac{n}{4} \leq \sum_{i=1}^{m} n_i \leq \frac{3n}{4} \qquad \frac{n}{4} \leq \sum_{i=m+1}^{k} n_i \leq \frac{3n}{4}.$$

Selezioniamo *m* in modo che

$$n_1 + \cdots + n_{m-1} < \frac{n}{4} \qquad \frac{n}{4} \leq n_1 + \cdots + n_{m-1} + n_m$$

Poichè sappiamo che n_m<*n/2* avremo

$$(n_1 + \ldots + n_{m-1}) + n_m < n/4+n/2=3n/4.$$

Otteniamo quindi

$$\frac{n}{4} \leq n_1 + \cdots + n_{m-1} + n_m \leq \frac{3n}{4}$$

Pertanto, possiamo selezionare *A* come l'unione delle componenti connesse di dimensioni n_1, n_2, …, n_m e *B* come l'unione delle componenti connesse rimanenti.

Dim. del Teorema. Chiamiamo 'good'' una coppia *(A, B)* di insiemi che soddisfano *n/4* $\leq$ *|A|* e *|B|* $\leq$ *3n/4.* Poichè sappiamo che se X≤n/2 allora esiste una coppia good (dal Fatto prec.), avremo che la probabilità che X≤n/2 è maggiorizzata dalla somma, su tutte le coppie good, della probabilità che non vi siano edge tra i nodi dei due insiemi che formano la coppia. Inoltre, il numero di coppie (A,B) good non può superare 2^n (cioè il numero di possibili scelte di A come sottoinsieme dei nodi). Avremo quindi

$$
\begin{aligned}
& \Pr[X \leq n/2] \\
\leq \; & 2^n \sup_{(A,B) \text{ is good}} \Pr[\exists(\text{no edges between } A \text{ and } B)] \\
\leq \; & 2^n \sup_{n/4 \leq m \leq 3n/4} (1-p)^{m(n-m)} \\
\leq \; & 2^n (1-p)^{3n^2/16} \\
\leq \; & 2^n \left[\left(1 - \frac{\ln 64}{n}\right)^n\right]^{3n/16} \\
\sim \; & 2^n (e^{-\ln 64})^{3n/16} \\
= \; & 2^n (2^{-6})^{3n/16} \\
= \; & 2^{-n/8}
\end{aligned}
$$

(Numero coppie (A,B) good) moltiplicato per la (prob. max che non vi siano edges tra A e B)

Se |A|=m, possono esservi tra A e B |A||B|=m(n-m) edges, Ognuno di questi *non* compare nel grafo con prob. 1-p

Il max di $(1-p)^{m(n-m)}$ si ha per m=3n/4

Ponendo p=ln 64/n

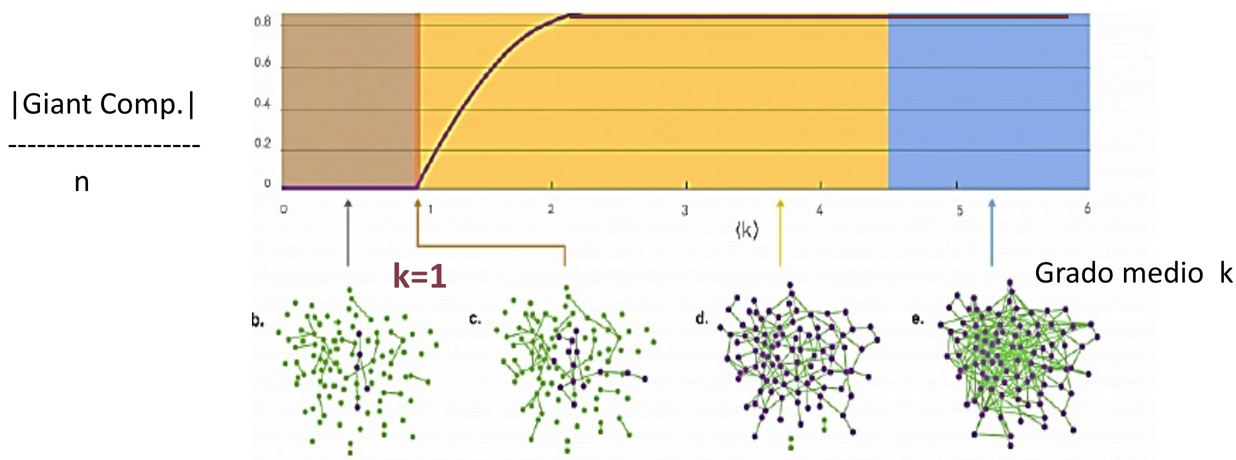
Noto: $(1-x/n)^n \approx e^{-x}$

È possibile dimostrare il seguente **risultato più forte sull'emergere di una componente gigante.**

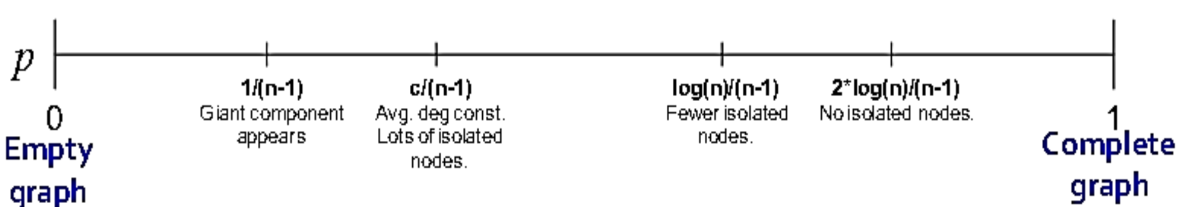
Teorema Siano k = p(n-1) il grado medio e $\varepsilon>0$

- Se $k = 1-\varepsilon$ allora tutte le componenti sono di dimensione O(log n)
- Se $k = 1+\varepsilon$ allora 1 componente ha dimensione Ω(n), e tutte altre sono di dimensione O(log n)

La figura seguente mostra l'andamento del rapporto tra la dimensione della componente gigante ed il numero n di nodi della rete al variare del grado medio k, si noti la transizione tra k=1 e k=2



La seguente figura ricapitola le proprietà dei grafi di Erdos-Renyi in termini di grado medio e giant component al variare del parametro p

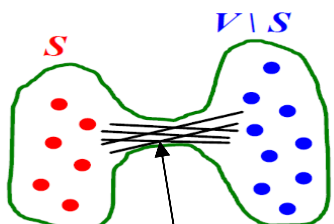


Struttura di G_{np} al variare di p

Espansione.

Definizione. Dato un grafo G=(V,E) ed un insieme S di nodi di G, definiamo *Eout(S)* come l' insieme degli edges uscenti da S (cioè aventi esattamente un estremo in S). L'espansione di G è definita come la minima espansione di un sottoinsieme dei nodi di G, cioè

$$a=\min_{S\subseteq V}\frac{|Eout(S)|}{\min(|S|,|V-S|)})=\min_{S\subseteq V,|S|\leq|V|/2}\frac{|Eout(S)|}{|S|}.$$



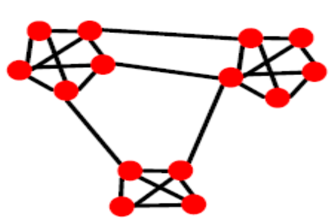
Espansione di S

L' espansione a di un grafo di G misura la robustezza di G. Infatti ci dice che è necessario eliminare almeno as edges per separare un sottoinsieme di s nodi dal resto del grafo.

La seguente figura mostra tre grafi con diversi gradi di espansione.

Espansione piccola: 1/5

Espansione alta



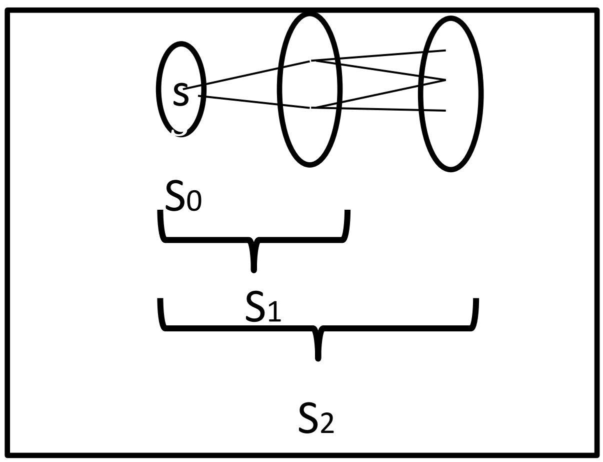
Reti sociali: Situazione mista; Comunità hanno espansione elevata.

Diametro.

Si può mostrare che vale la seguente proprietà

Teorema. *In un grafo con n nodi, grado massimo d ed espansione a, per ogni coppia di nodi s e t esiste un cammino di lunghezza O((d/a)log n))*

Dimostrazione. Consideriamo una BFS da s. Indichiamo con Sj l'insieme dei nodi trovati nei primi j livelli della BFS.



Il livello (j+1)-mo è ottenuto seguendo gli edges uscenti da Sj. Aggiungendo i nodi di questo livello ad Sj otteniamo Sj+1.
Se Sj contiene meno di n/2 nodi, l'espansione di G implica che ci sono almeno a|Sj| edges che escono da Sj.
Alcuni edges possono portare allo stesso nodo, ma sapendo che il grado massimo è d, il numero di nuovi nodi è almeno $(a/d)\,|S_j|$, *quindi*

$$|S_{j+1}| \geq (a/d)\,|S_j|$$

Quindi

$$\begin{aligned}|Sr| \geq (1+(a/d))\,|Sr\text{-}1| &\geq (1+(a/d))^2\,|Sr\text{-}2| \\ &\geq (1+(a/d))^3\,|Sr\text{-}3| \\ &\;\;\ldots \\ &\geq (1+(a/d))^r\,|S0| \\ &= (1+(a/d))^r\end{aligned}$$

Notando che $a \leq d$, scegliamo $r= (d/a)\log n$. Otteniamo

$|Sr| \geq (1+(a/d))^r=(1+(a/d))^{(d/a)\log n}$

Ricordiamo che la funzione $(1+1/x)^x$ *tra 1 ed infinito è crescente, vale 2 in 1 ed il limite per x tendente ad infinito è e. Otteniamo quindi* $(1+(a/d))^{(d/a)} \geq 2$.

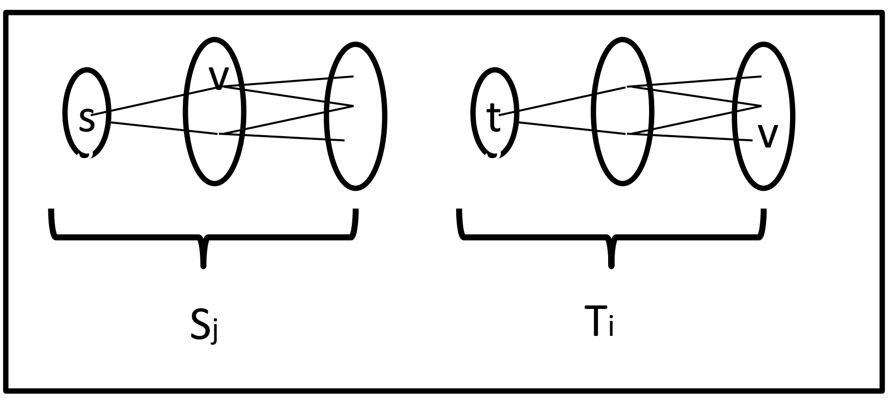
Da cui

$$|Sr| \geq (1+(a/d))^{(d/a)\log n} \geq 2^{\log n} = n$$

Sapendo che $|S(d/a)\log n| \geq n$, otteniamo che per qualche $j \leq (d/a)\log n$

Sj contiene più di n/2 nodi (e |Sj-1|<n/2, cfr. la definizione di espansione).

- Se t appartiene ad Sj, abbiamo trovato il cammino cercato.
- Se t non appartiene ad Sj, allora procediamo come segue:
  - Effettuiamo una seconda BFS a partire dal nodo t.



  - Come prima sappiamo che per un qualche livello $i \leq (d/a)\log n$ l'insieme Ti dei nodi trovati ha cardinalità maggiore di n/2.

A questo punto sappiamo che |Sj| + |Ti| >n.

Ne consegue che gli insiemi devono avere almeno un elemento v in comune.

La concatenazione delle path s---v e t---v ci assicura l'esistenza di un cammino s---t di lunghezza al più

$$i+j \leq 2\,(d/a)\log n$$

In maniera simile si può mostrare che

Teorema. *In un grafo con n nodi ed espansione a, il diametro è O((log n)/log a)*

IDEA: *Si dimostra che al crescere di i, l'insieme S_i cresce almeno in maniera geometrica in funzione di i (cioè come a^i).*

➔ Per r ≈*(log n)/log a,* Sr contiene più della metà dei nodi

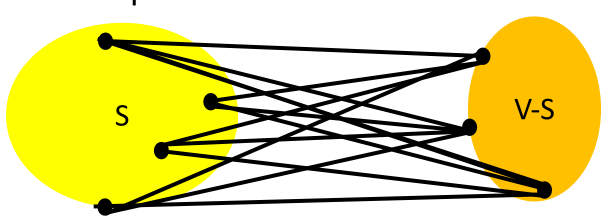
➔ Per ogni altro nodo t si ha ***Sr ∩ Tr ≠ Ø***

➔ La distanza tra s e t è al più 2r

Espansione e Diametro di $G_{n,p}$.

Teorema: In $G_{n,p}$ ha un espansione pari almeno a pn/2

Dimostrazione: Consideriamo $G_{n,p}$. Per ogni insieme S di s nodi gli edges presenti nel grafo sono scelti tra gli s(n-s) edges aventi un estremo in S ed un estremo in V-S. Ogni edge viene inserito con probabilità p e viene scartato con probabilità 1-p

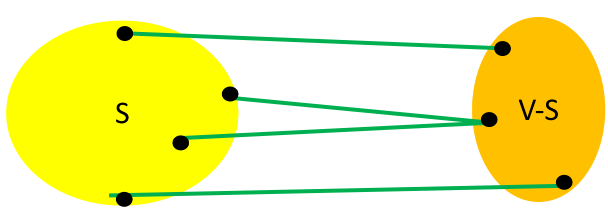


Tutti gli edges possibili tra S e V-S

Ricordiamo che ogni edge viene inserito con probabilità p e viene scartato con probabilità 1-p.

Ne consegue che, in media il numero di edges inseriti nel grafo è

$$p\, s(n-s)\,.$$



Una possibile realizzazione con p=1/3

Di conseguenza, l' espansione attesa è min $\frac{ps(n-s)}{s}$ = min p(n-s)≥ $\frac{\boldsymbol{pn}}{\mathbf{2}}$

Teorema: *In $G_{n,p}$ se p>c (log n)/n per una costante c sufficientemente grande allora il diametro è O(log n)*

Dimostrazione. Sia *p>c (log n)/n. Per c>2 l'espansione è*

$$a = p\frac{n}{2} > c\,\frac{log\ n}{n}\,\frac{n}{2} > \log n.$$

Dal teorema precedente sappiamo che *il diametro è O((log n)/log a)), con*

$$\frac{\log n}{\log a} = \frac{\log n}{\log \log n} < \log n.$$

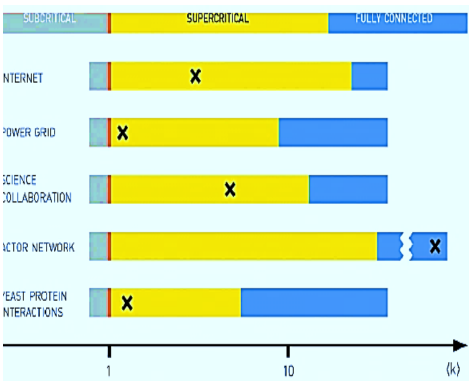
In generale è possibile vedere che anche per il diametro esiste un valore di p oltre il quale la rete diventa di tipo small-world.

Per *n* grande:

- Se $p<n^{-5/6}$, la probabilità che la rete abbia un diametro al più 6 tende a 0.
- Se $p>n^{-5/6}$, la probabilità che la rete abbia un diametro al più 6 tende a 1

Reti reali vs grafi casuali. Consideriamone i parametri principali.
Il modello Erdos-Renyi permette di spiegare

- **Giant component.** La maggior parte delle reti reali hanno grado medio tale da implicare una componente gigante, *ciò è in accordo con le osservazioni*



Supercritical: grado medio superiore alla soglia critica 1
Le reti reali sono in massima parte supercritical (e non completamente connesse)

- **Lunghezza media cammini.** C'è concordanza tra il valore atteso O(log n) determinato per i grafi casuali ed i valori riscontrati nelle reti reali.
  l modello ER è un buon predittore del diametro e della lunghezza media dei cammini rispetto alle reti reali
  Il modello dà reti di diametro piccolo. Cattura molto bene la proprietà "small-world" osservata in molte reti reali

Il modello Erdos-Renyi non permette di spiegare

- **Distribuzione dei gradi.** La distribuzione dei gradi differisce da quello delle reti reali che hanno una distribuzione dei gradi meno omogenea (si pensi per esempio al numero di collegamenti/popolarità nelle reti sociali online).
  La maggior parte delle reti reali mostra una distribuzione dei gradi con una lunga coda

- **Coefficiente Clustering**. A differenza delle reti reali, i grafi casuali non hanno struttura locale. il coefficiente di clustering risulta pari a

$$p=k/(n-1)$$

  Coefficiente di clustering troppo piccolo e troppo vicino alla densità degli edge
  La maggior parte delle reti reali è altamente clusterizzata con coefficienti di clustering molto maggiori della densità degli edge