# La ricerca sulle reti complesse in Microsoft

Filippo Palombi

ENEA

27 Nov. 2017 – C.R. Frascati

Jennifer Tour Chayes

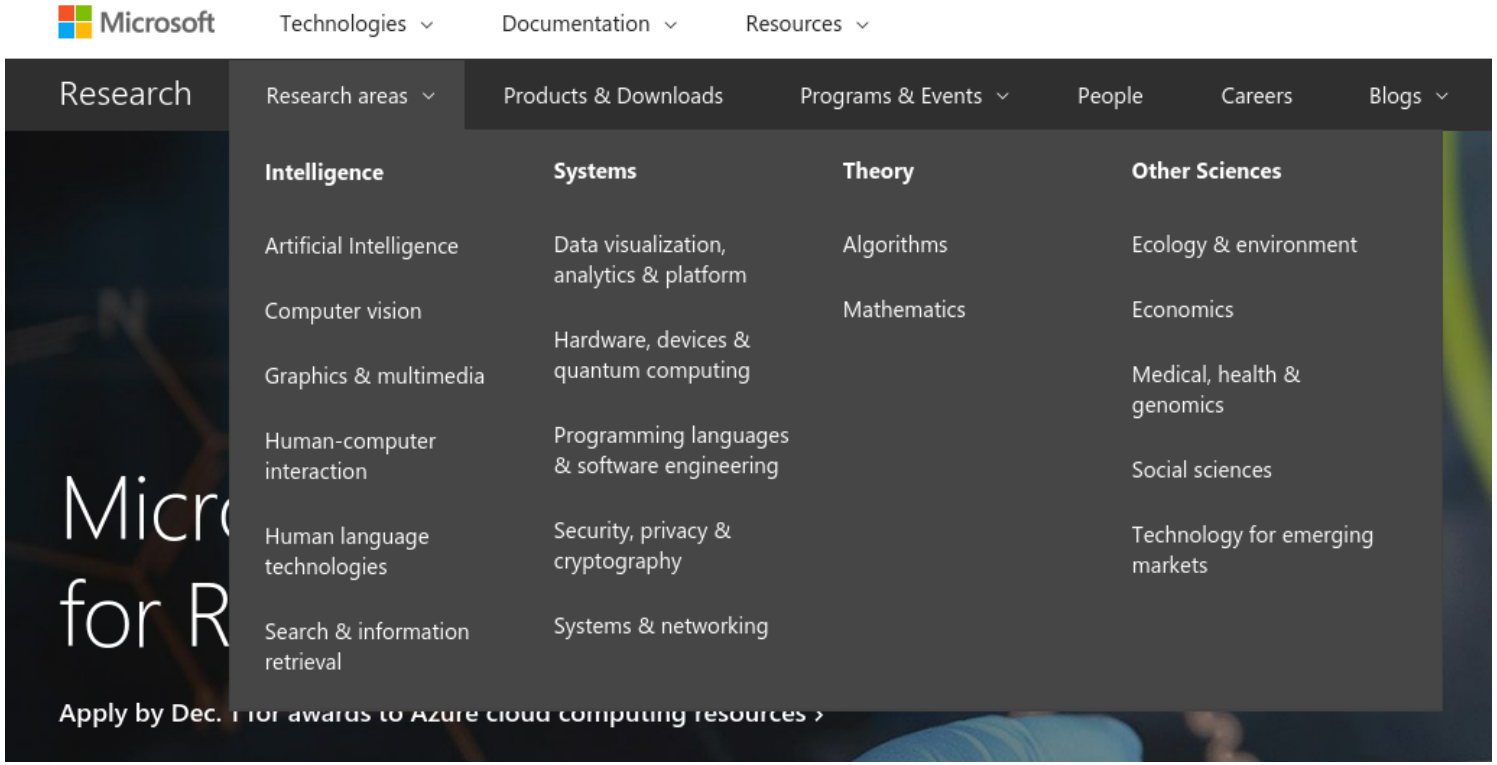
(Managing Director and Distinguished Scientist at Microsoft Research New England)

ha tenuto una Conference Keynote a ISC 2017 (Francoforte).

In questo seminario discutiamo le sue slides (...ed altro...)

## Chi è Jennifer Tour Chayes?

# Aree di ricerca



https://en.wikipedia.org/wiki/Jennifer_Tour_Chayes
The Theory Group analyzes fundamental questions in theoretical computer science using techniques from statistical physics and discrete mathematics. Chayes opened Microsoft Research New England in July 2008 with Christian Borgs.

https://www.microsoft.com/en-us/research/people/jchayes/
Chayes lives with her husband, Borgs, who also happens to be her principal scientific collaborator. **In her spare time she enjoys overworking**.

# Network Science: From the Online World to Cancer Genomics

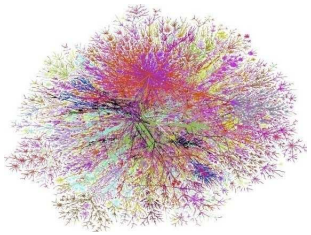
Jennifer Tour Chayes
Microsoft Research New England
Microsoft Research New York City

ISC High Performance Conference
Frankfurt, Germany
June 19, 2017

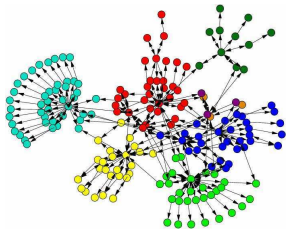
# Outline

- "Observed" Networks
- Mathematical and Algorithmic Problems on Networks
- A Couple of Specific Applications:
  - Collaborative Filtering on Bipartite Networks (e.g., Netflix or Amazon)
  - Reconstruction of Genomic and Proteomic Networks

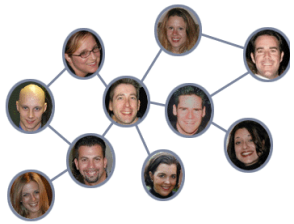
# Motivation: The Age of Networks



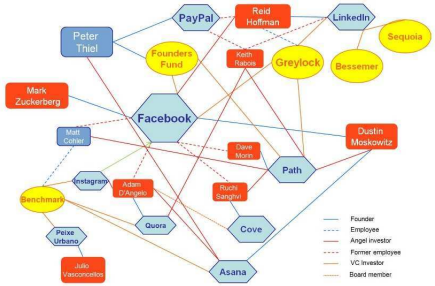
Internet



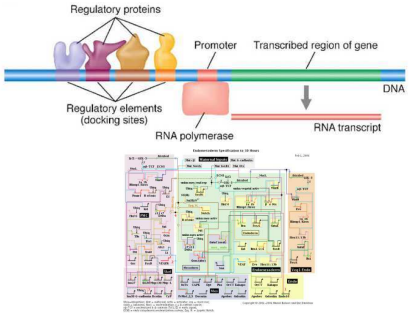
WWW



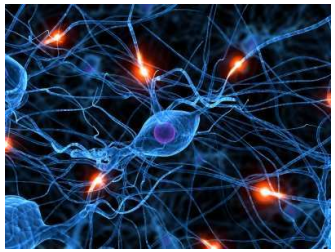
Social network



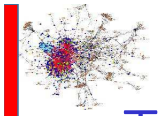

Economic network




Gene regulatory network



Neural network



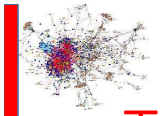
# Types of Networks

Modeling Networks: We model networks as graphs

$$G = (V, E)$$

Types of Networks:

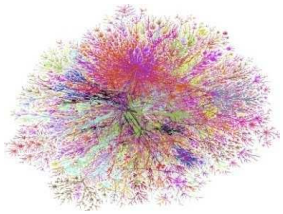
- Technological
- Social
- Economic
- Biological

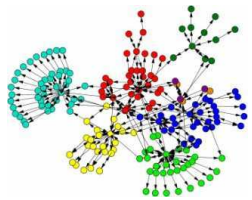


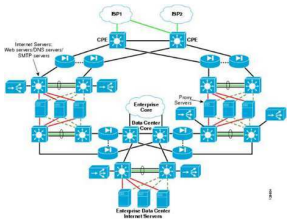
# Technological Networks

- AS (Autonomous System) Internet
  - $V$ autonomous systems (AOL, MSN, MIT.edu, etc.)
  - $E$ connections
- WWW
  - $V$ webpages
  - $E$ hyperlinks (directed)
- Cloud (data center) networks

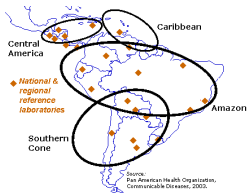
....

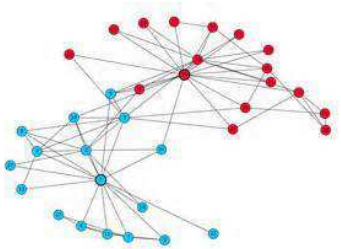


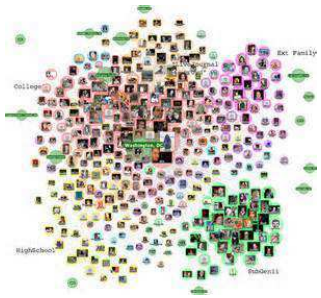


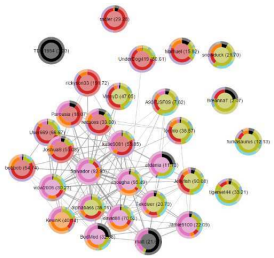


# Social Networks







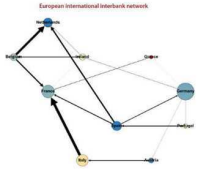
- Offline
  - Epidemiological networks
  - Offline activity-based social networks,
    e.g., karate club
- Online
  - Facebook, LinkedIn
  - Twitter (directed)
  - Email
  - Instant Messaging (IM)
  - Mobile phone

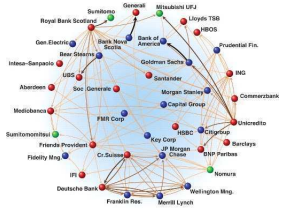
  ....
  - Crisis communication via online networks

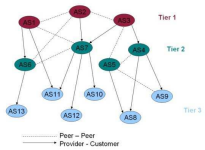
  ...

# Economic Networks

- Offline
  - Lending networks, e.g. Eurozone, banks
  - Tripartite networks of firms, VCs & “influentials”
- Online
  - Peering agreement networks
  - Banner advertiser & publisher networks

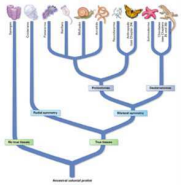


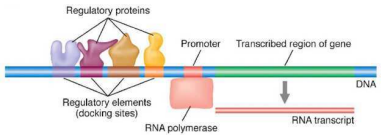


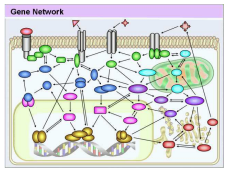

# Biological Networks

- Phylogenetic networks

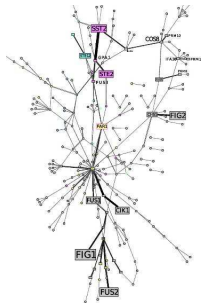


- Molecular biology – genomic & proteomic – networks

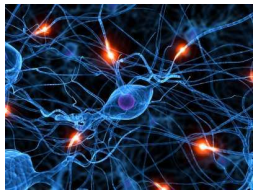


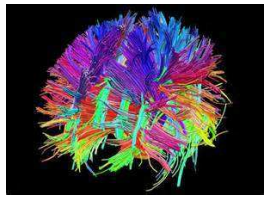




Protein Interactome, e.g., pheromone pathway in yeast



- Neurological networks





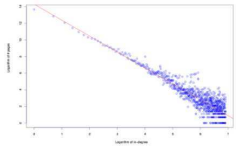
Connectome

# Mathematical and Algorithmic Problems on Networks

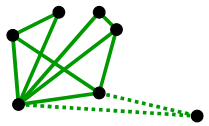
- Modeling networks
- Sampling from and machine learning of networks
- Processes on networks
- Algorithms on networks
- Network reconstruction algorithms

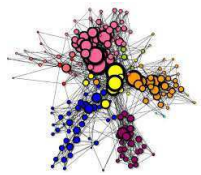
# 1. Modeling networks

- Observations of tech and social networks
  - Small diameter
    ~ 6 degrees of separation: 1929
    Frigyes Karinthy's short story, "Chains"
  - Power-law degree distribution



- Two types of models:
  - Growth models, e.g., preferential attachment (Barabási-Albert 1999)
    - At each time step, a new vertex is created and attaches to m old vertices, with probability proportional to the degree of the old vertex
  - Latent feature models, e.g., stochastic block models (Holland, Laskey, Leinhardt '83) or graphons (Hoff, Raftery, Handcock '02)
    - Each vertex has a latent unobserved feature
    - Vertices connect with a probability depending on their features





# Modelli di rete

- **Modelli statici**

  Grafi di Erdős-Rényi
  dati $n$ nodi, li si connettono a coppie con prob. $p$)

  Modelli a blocchi stocastici
  dati $n$ nodi, li si dividono in gruppi $C_k$ e si connettono i nodi nei gruppi $(C_i, C_k)$ con prob. $p_{ik}$

- **Modelli dinamici (o di crescita)**

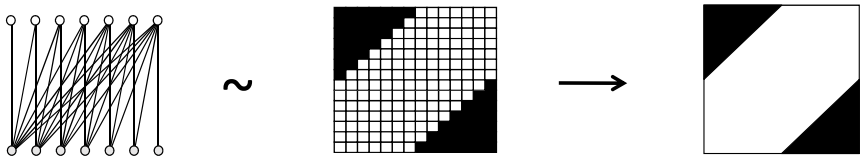
  Reti di Barabási-Albert
  preferential attachment (rich gets richer)

  Reti con parametri di fitness
  non sempre i nodi più vecchi sono quelli di maggior successo
  (cfr. Google vs. Altavista)... si introducono parametri di fitness (fit gets richer)

  Reti basate sulla teoria dei giochi
  prima di aggiungere un nuovo sito la rete attuale trova un equilibrio
  sulla base di parametri di fitness in competizione

## 2. Sampling from and machine learning on networks

- The WWW is very large (order of trillions of static sites) and growing.
- How do we sample from it, e.g., to calculate pagerank?
- To deal with this, we developed a theory of graph limits (graphons) and testing for dense graphs: Borgs, Chayes, Lovász, Sós, Vesztergombi '06 – '12



Results for sparse graphs:

Limits of sparse graphs with power-law tails
(Borgs, Chayes, Cohn, Zhao '14 - 15)

Limits of sparse graphs with multiscale power-law tails
(Borgs, Chayes, Cohn, Holden '15 - 16)

# 2. Sampling from and machine learning on networks (cont.)

**Question**: How can you take a single snapshot of FB or LinkedIn today, and learn the network to predict

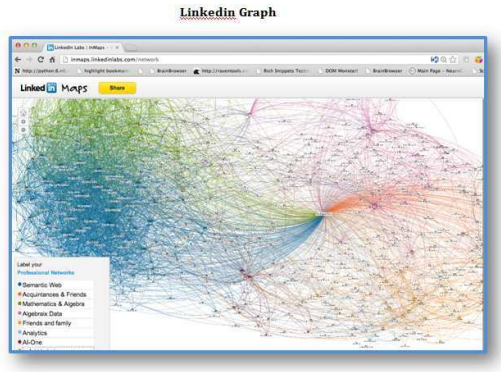
- how the network will look when it's twice as big?
- network parameters,* e.g., to do A/B testing or collaborative filtering on the network.

**Answer**: Get an estimate of the graph limit (non-parametric network model) corresponding to a snapshot of FB or LinkedIn today, & use this estimate to predict how the network will look at a later time or larger size.

**Theorem**: This gives consistent estimation for massive networks like FB & LinkedIn. (Borgs, Cohen, Chayes, Ganguly '15)

Also new (provable) graphon algorithms for A/B testing (Airoldi, Borgs, Chayes, Lee '17) and collaborative filtering (Borgs, Chayes, Lee, Shah '17).

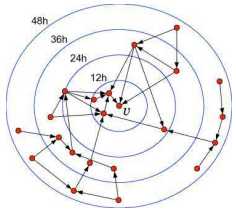
Specific Application

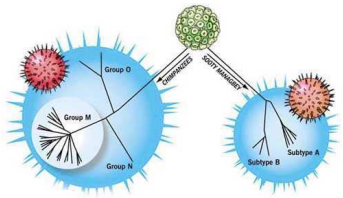

*Or nonparametric network model

# 3. Processes on networks

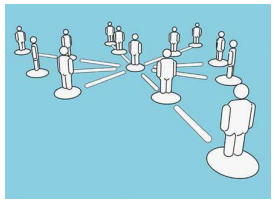
- Flow of information
  - J. Kleinberg *et al.*




- Spread of epidemics
  - Berger, Borgs, Chayes, Saberi ’05, ’10

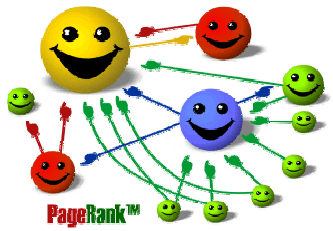



- Viral marketing
  - Kempe, J. Kleinberg, E. Tardos



# 4. Algorithms on networks: A simple example

- Early search engines used semantics (i.e., content and language) to find the most relevant webpages
- Later search engines (e.g., Google, Bing) used the structure of the web graph (i.e., graph theory and algorithms) to find the most relevant webpages
- Pagerank: Do a random walk on the web graph, following the hyperlinks, restarting every say 7 steps. The relative weight of webpage in the stationary distribution of this walk is its Pagerank (Brin and Page, '96)
- Later ranking algorithms detect and avoid anomalies to downgrade rankings of web-spammers (Andersen, Borgs, Chayes, Hopcroft, Mirrokni, Teng '07)



# L'algoritmo PageRank (PR) - 1996

Larry Page & Sergey Brin

- La parola Google è la versione "storpiata" di Googol
- 1 Googol = $10^{100}$ è una quantità introdotta dal matematico Edward Kasner nel 1940 nel suo libro "Mathematics and Imagination"
- La parola Googol era stata inventata dal nipote di Kasner nel 1920 all'età di 9 anni.
- La scelta del nome Google è associata all'eccellente scalabilità dell'algoritmo PR

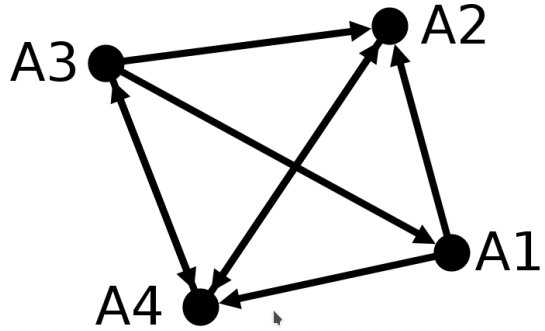
---

- Supponiamo di avere un grafo $G = (V, E)$ direzionale con nodi $V = \{A_1, A_2, \ldots, A_N\}$
- Assegnamo a ciascun nodo $A_k$ un rango $0 \leq PR(A_k) \leq 1$
- Definiamo i ranghi implicitamente attraverso le equazioni

$$PR(A_k) = \frac{1-d}{N} + d \sum_{A \in M(A_k)} \frac{PR(A)}{L(A)}, \qquad k = 1 \ldots, N$$

- $0 \leq d \leq 1$ è una costante decisa a priori
- $M(A_k)$ è l'insieme dei nodi $A$ che puntano ad $A_k$
- $L(A)$ è il numero di archi uscenti da $A$

# Esempio




- $A1 \rightarrow A2, A4$ $\quad L(A1) = 2$ $\quad M(A1) = \{A3\}$
- $A2 \rightarrow A4$ $\quad L(A2) = 1$ $\quad M(A2) = \{A1, A3, A4\}$
- $A3 \rightarrow A1, A2, A4$ $\quad L(A3) = 3$ $\quad M(A3) = \{A4\}$
- $A4 \rightarrow A2, A3$ $\quad L(A4) = 2$ $\quad M(A4) = \{A1, A2, A3\}$

$$PR(A_1) = \frac{1-d}{N} + d\frac{PR(A3)}{3}$$
$$PR(A_2) = \frac{1-d}{N} + d\left[\frac{PR(A1)}{2} + \frac{PR(A3)}{3} + \frac{PR(A4)}{2}\right]$$
$$PR(A_3) = \frac{1-d}{N} + d\frac{PR(A4)}{2}$$
$$PR(A_4) = \frac{1-d}{N} + d\left[\frac{PR(A1)}{2} + PR(A2) + \frac{PR(A3)}{3}\right]$$

- Fissiamo d = 0.85
- Soluzione:

$PR(A1) = 0.095673$ $\quad PR(A2) = 0.304150$

$PR(A3) = 0.205316$ $\quad PR(A4) = 0.394861$

- N.B.: $PR(A1) + PR(A2) + PR(A3) + PR(A4) = 1$

## Proprietà

$$\sum_{A \in V} PR(A) = 1$$

---

## Dimostrazione

Partiamo dalla formula generale

$$PR(A) = \frac{1-d}{N} + d \sum_{B \in M(A)} \frac{PR(B)}{L(B)}$$

Sommiamo ambo i membri su tutti i nodi della rete:

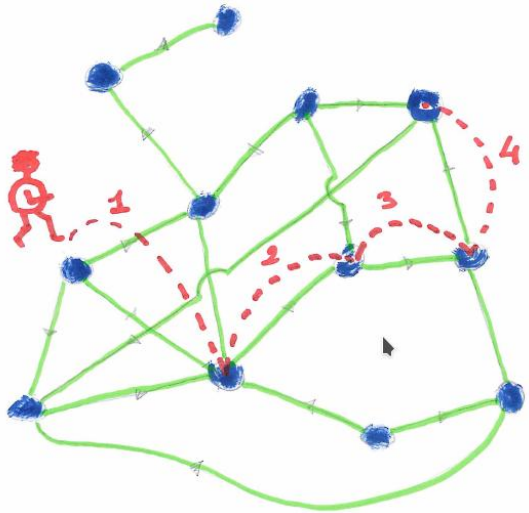
$$\sum_{A \in V} PR(A) = (1-d) + d \sum_{A \in V} \sum_{B \in M(A)} \frac{PR(B)}{L(B)}$$

Per valutare la doppia somma, consideriamo un certo nodo $B \in V$ e cerchiamo di capire quante volte $B$ entra in gioco. Evidentemente $B$ compare in corrispondenza di ciascun nodo $A_k$ per cui $B \in M(A_k)$. Cioè $B$ compare esattamente $L(B)$ volte nella doppia somma. Ne consegue

$$\sum_{A \in V} PR(A) = (1-d) + d \sum_{B \in V} PR(B),$$

da cui segue la proprietà enunciata.

# Interpretazione geometrica di PR




- Si consideri un *random walk* sulla rete
- Lo spostamento da un nodo $A$ avviene lungo uno dei link uscenti da $A$ con ugual probabilità
- Se il nodo $A$ non ha link uscenti, si riparte da un nodo qualunque scelto con ugual probabilità
- Si può dimostrare che:

frazione di tempo trascorsa sul nodo $A \overset{t\to+\infty}{\to} PR(A)$

# Implementazione numerica di PR

- calcoliamo $\{PR(A)\}_{A\in V}$ ricorsivamente
- indichiamo con $t = 0, 1, 2, \ldots$ l'indice di ricorsione e con $\{PR_t(A)\}_{A\in V}$ i ranghi al passo $t$
- inizializziamo i ranghi in modo uniforme

$$PR_0(A) = \frac{1}{N}, \qquad A \in V$$

- quindi computiamo

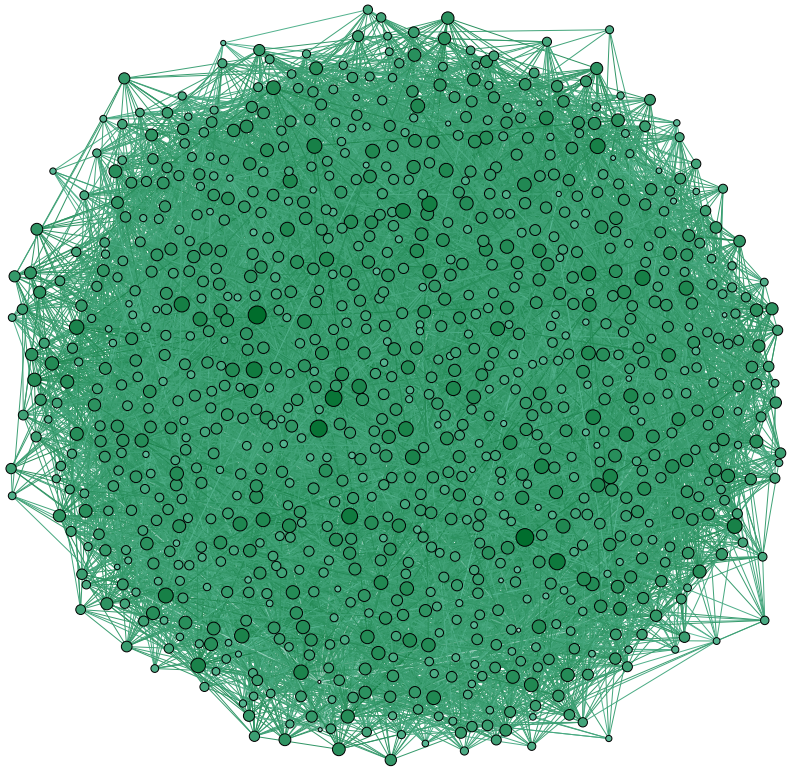
$$PR_{t+1}(A) = (1-d) + d \sum_{B\in M(A)} \frac{PR_t(B)}{L(B)}, \qquad A \in V$$

fino a convergenza, cioè fino a quando

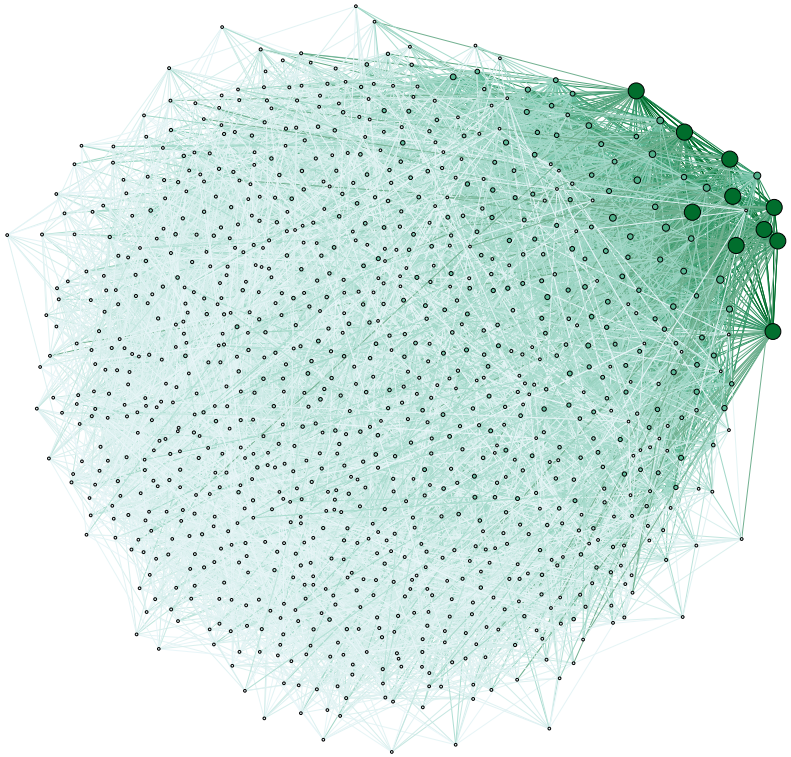
$$\sqrt{\sum_{A\in V} |PR_{t+1}(A) - PR_t(A)|^2} < \epsilon$$

per $\epsilon$ prefissato, per es. $\epsilon = 1.0 \times 10^{-8}$.

```
# ------------------------------------------------------------------------
# Page Ranking
# ------------------------------------------------------------------------

d = 0.85
eps = 1.0e-8

PageRank = []
OldPageRank = []
for j in range(N):
    PageRank.append(1.0/float(N))
    OldPageRank.append(1.0/float(N))

err = 1.0
iter = 0
print "\n"
While(err>eps):
    print "iteration = %d \t err = %8.6e" % (iter,err)
    for node in range(N):
        PageRank[node] = (1-d)/float(N)
        for adj in InAdjVec[node]:
            PageRank[node] += d*PageRank[adj]/len(OutAdjVec[adj])
    err = 0.0
    for node in range(N):
        err += mt.pow(PageRank[node] - OldPageRank[node],2.0)
        OldPageRank[node] = PageRank[node]
    err = mt.sqrt(err)
    iter += 1

print "\n"
for node in range(N):
    print "PageRank[%02d] = %10.8e" % (node,PageRank[node])


filename="net.gexf"
fd = open(filename,'w')
write_preamble(fd)
write_nodes(fd)
write_edges(fd)
write_closure(fd)
```



- Una rete direzionale di Erdős-Rényi con $N = 1000$ e $p = 0.01$
- Il colore appartiene ad una scala che va dal bianco ($PR = 0$) al verde intenso ($PR = 1$)
- Anche la dimensione dei nodi (cerchietti) cresce con il relativo PR
- I valori di PR sono stati normalizzati col max. valore osservato



- Una rete direzionale di Barabási-Albert con $N = 1000$ ed $m = 10$ ($p = 0.01$)
- Il colore appartiene ad una scala che va dal bianco ($PR = 0$) al verde intenso ($PR = 1$)
- Anche la dimensione dei nodi (cerchietti) cresce con il relativo PR
- I valori dei PR sono stati normalizzati col max. PR osservato

## Considerazioni su PR

- L'ordinamento prodotto da PR è globale sul WWW
- La qualità dell'output di PR dipende dalla topologia della rete
- Se il WWW fosse una rete di ER, PR non funzionerebbe
- Connessione profonda tra Google e la struttura scale-free di Internet
- Esistono strategie per alterare il PR di un nodo sul WWW complessivamente note come spamdexing, search engine spam, search engine poisoning, black-hat SEO, search spam or web spam (es. link farms, cloaking, etc.)

## Domande:

- Come si campiona una rete di grandi dimensioni?
- Come si ricostruisce una rete di grandi dimensioni?

## Domande:

- Come si campiona una rete di grandi dimensioni?
- Come si ricostruisce una rete di grandi dimensioni?

Due tipi di reti:

- grafi densi
- grafi sparsi

# Domande:

- Come si campiona una rete di grandi dimensioni?
- Come si ricostruisce una rete di grandi dimensioni?

Due tipi di reti:

- grafi densi
- grafi sparsi

Introduciamo il grado medio (numero medio di primi vicini per nodo): $\langle d \rangle = \frac{2|E|}{|V|}$

Un grafo si dice denso se nel limite termodinamico ($|V| \to +\infty$) $\exists c > 0$: $\langle d \rangle \to c|V|$
Un grafo si dice sparso se nel limite termodinamico $\exists c > 0$: $\langle d \rangle < c$

# Domande:

- Come si campiona una rete di grandi dimensioni?
- Come si ricostruisce una rete di grandi dimensioni?

Due tipi di reti:

- grafi densi
- grafi sparsi

Introduciamo il grado medio (numero medio di primi vicini per nodo): $\langle d \rangle = \frac{2|E|}{|V|}$

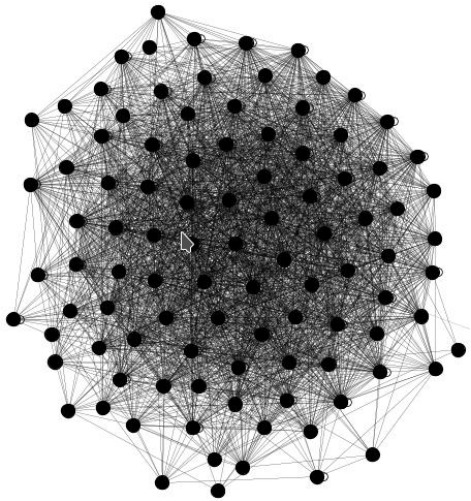
Un grafo si dice denso se nel limite termodinamico ($|V| \to +\infty$) $\exists c > 0$: $\langle d \rangle \to c|V|$
Un grafo si dice sparso se nel limite termodinamico $\exists c > 0$: $\langle d \rangle < c$

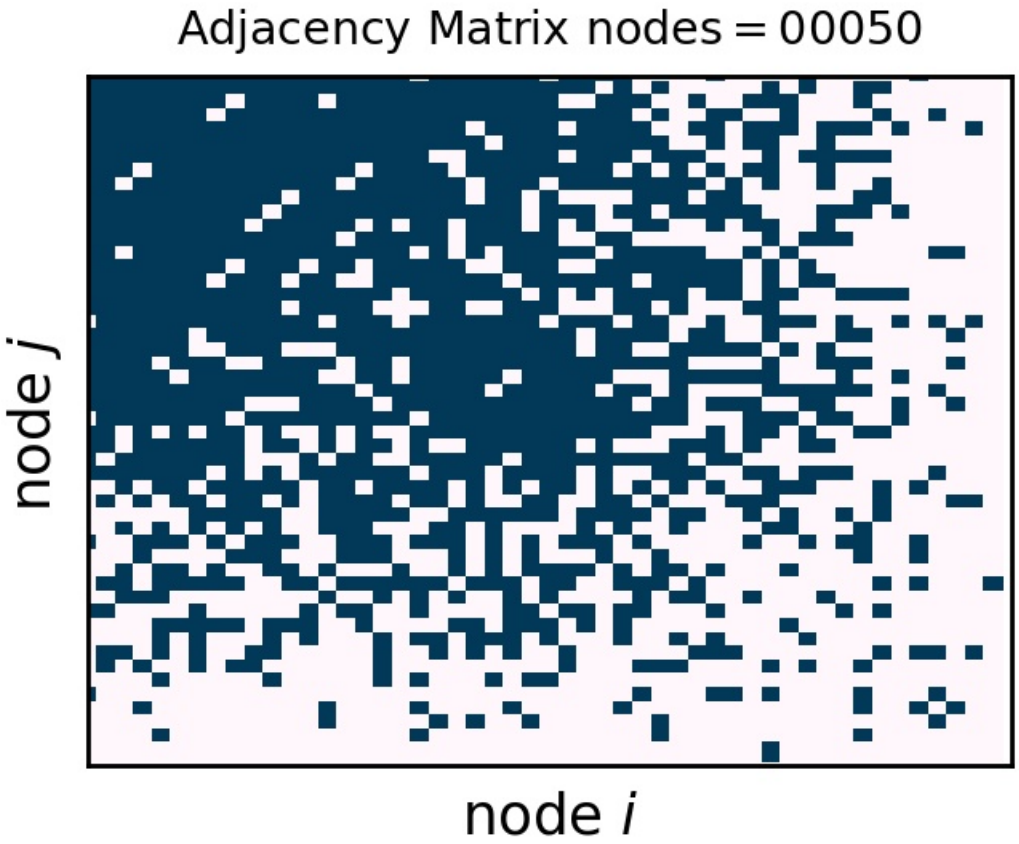
**Quello che segue si applica essenzialmente ai grafi densi (per il momento...)**

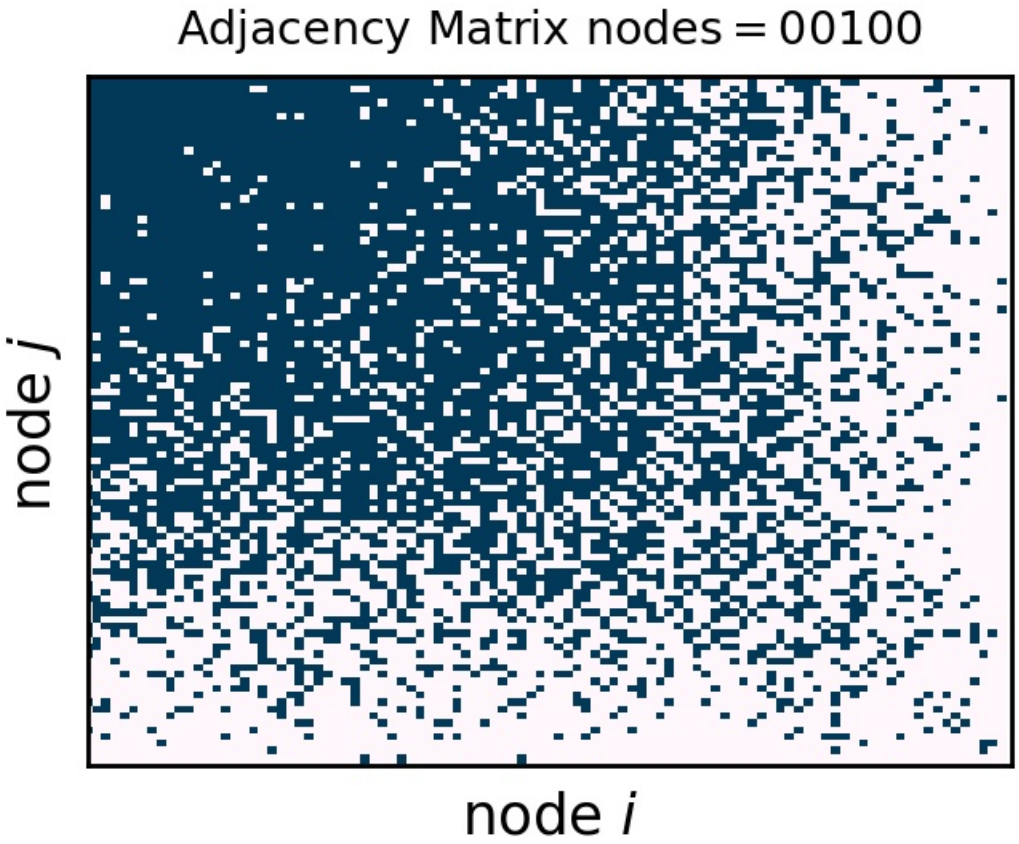
# Modello di crescita con attachment uniformemente random

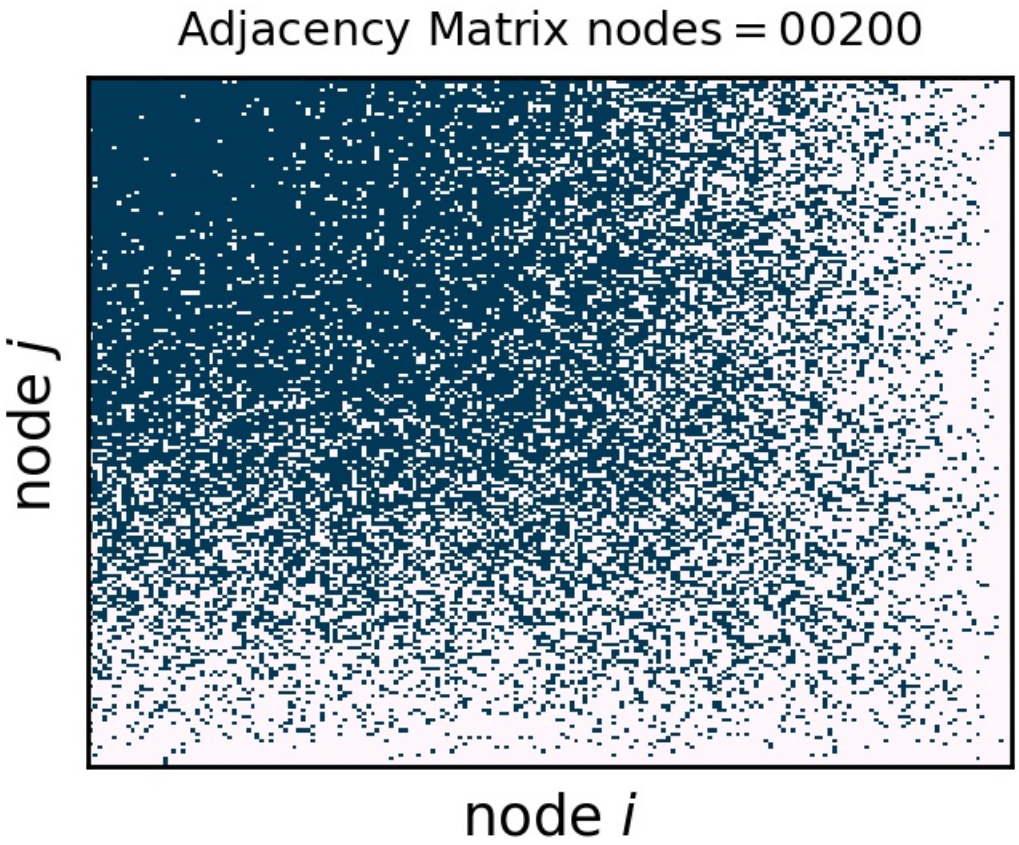
- Consideriamo inizialmente un singolo nodo
- Al passo $n$-mo aggiungiamo un nuovo nodo alla rete
- Al passo $n$-mo linkiamo ciascuna coppia di nodi con probabilità $1/n$

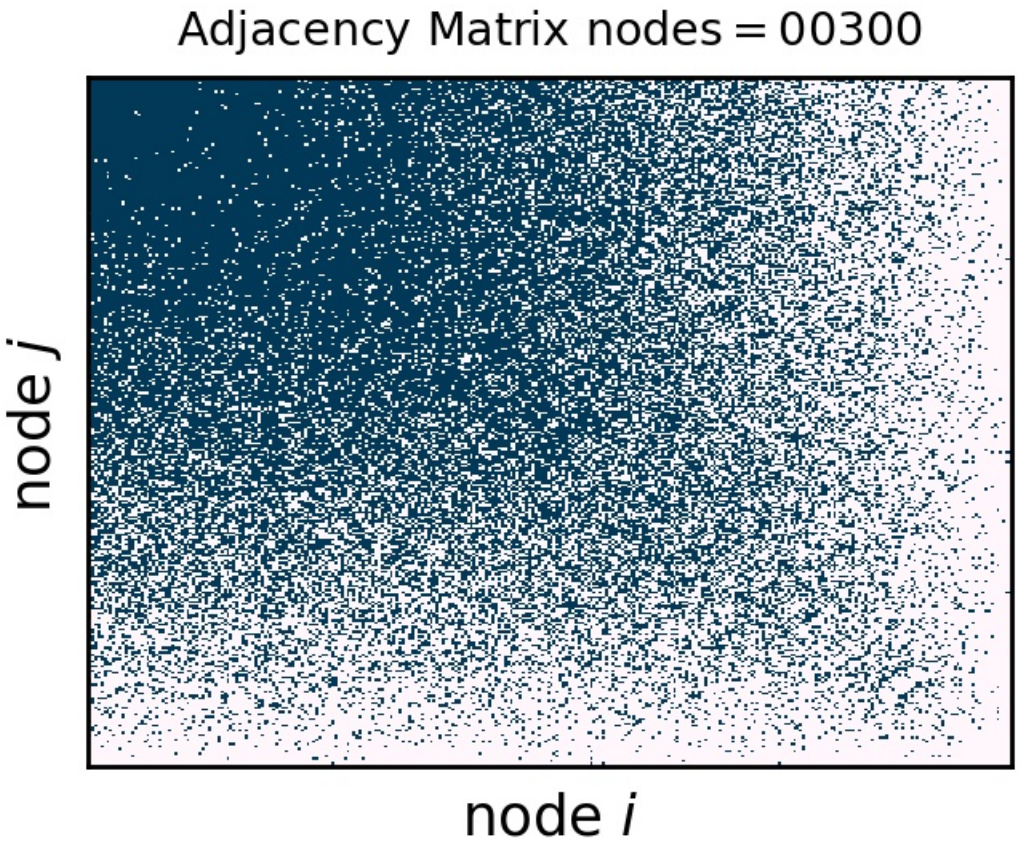
Otteniamo un grafo denso.

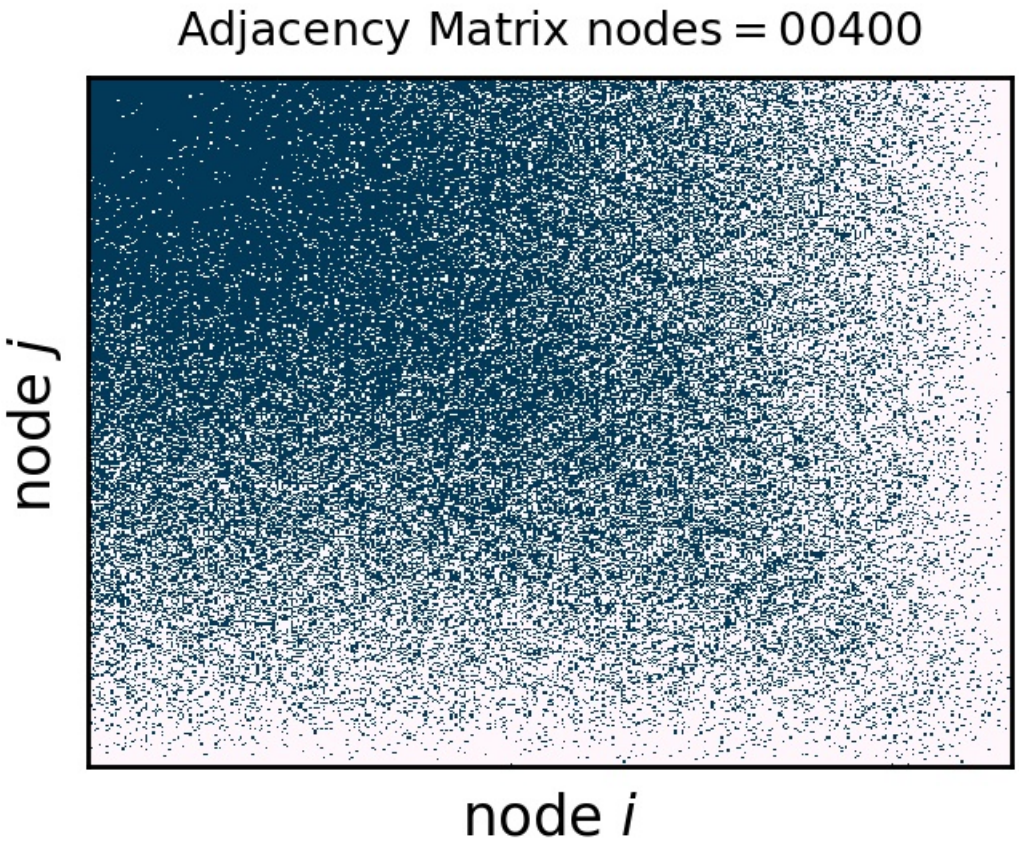


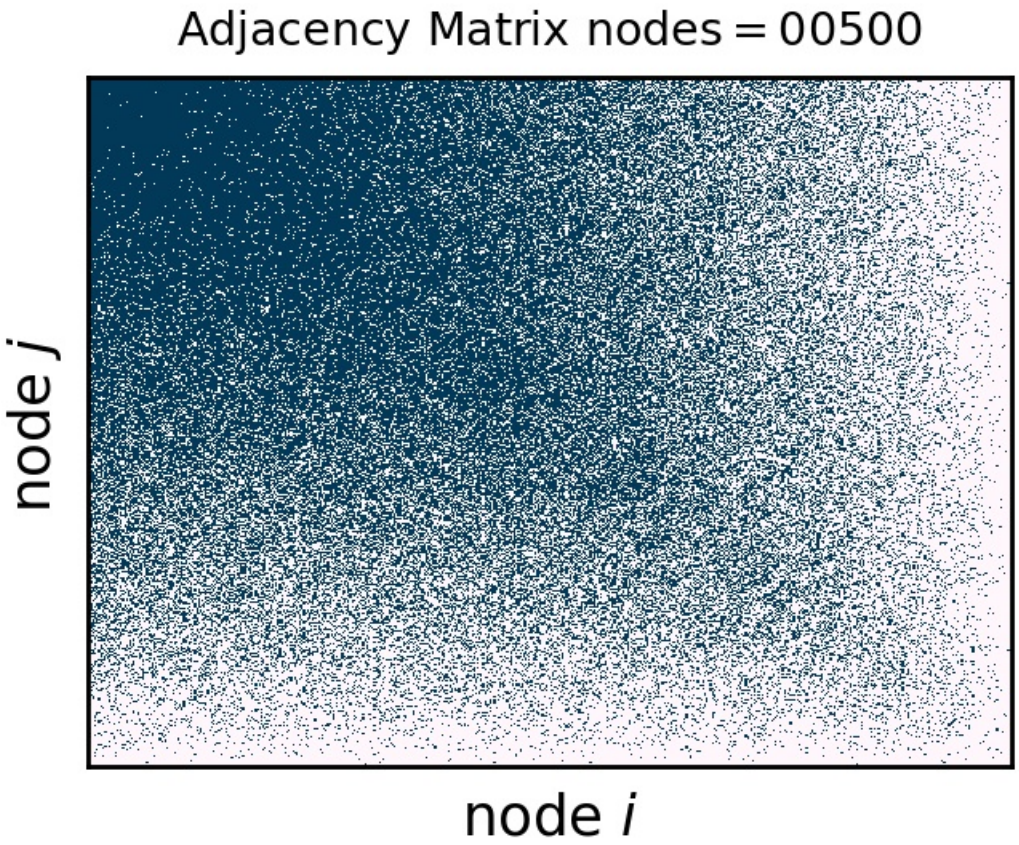

Adjacency Matrix nodes = 00050
node j
node i

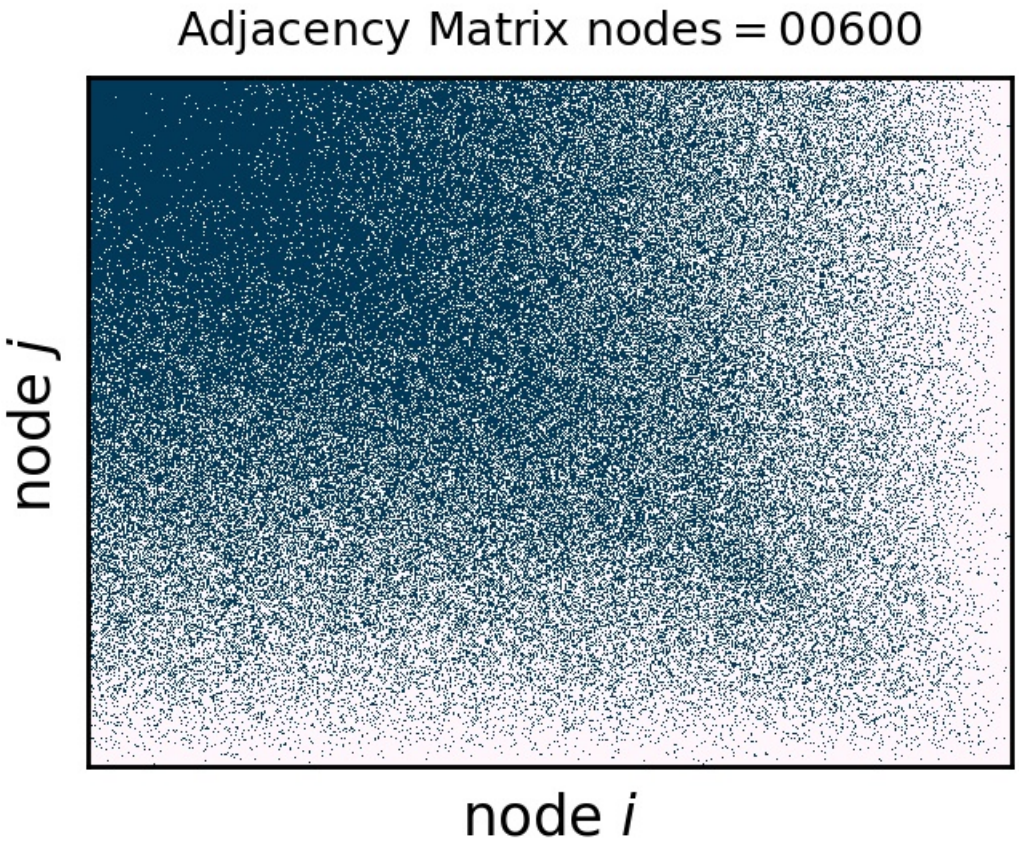

Adjacency Matrix nodes = 00100
node j
node i

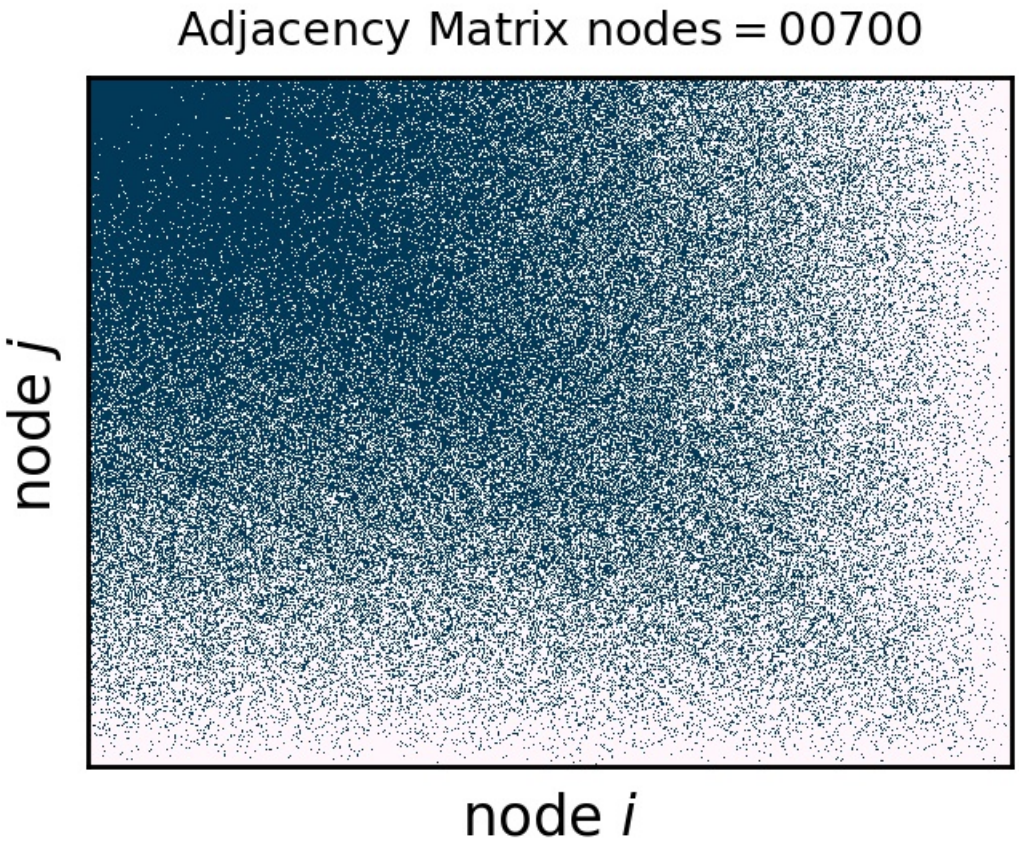

Adjacency Matrix nodes = 00200
node j
node i

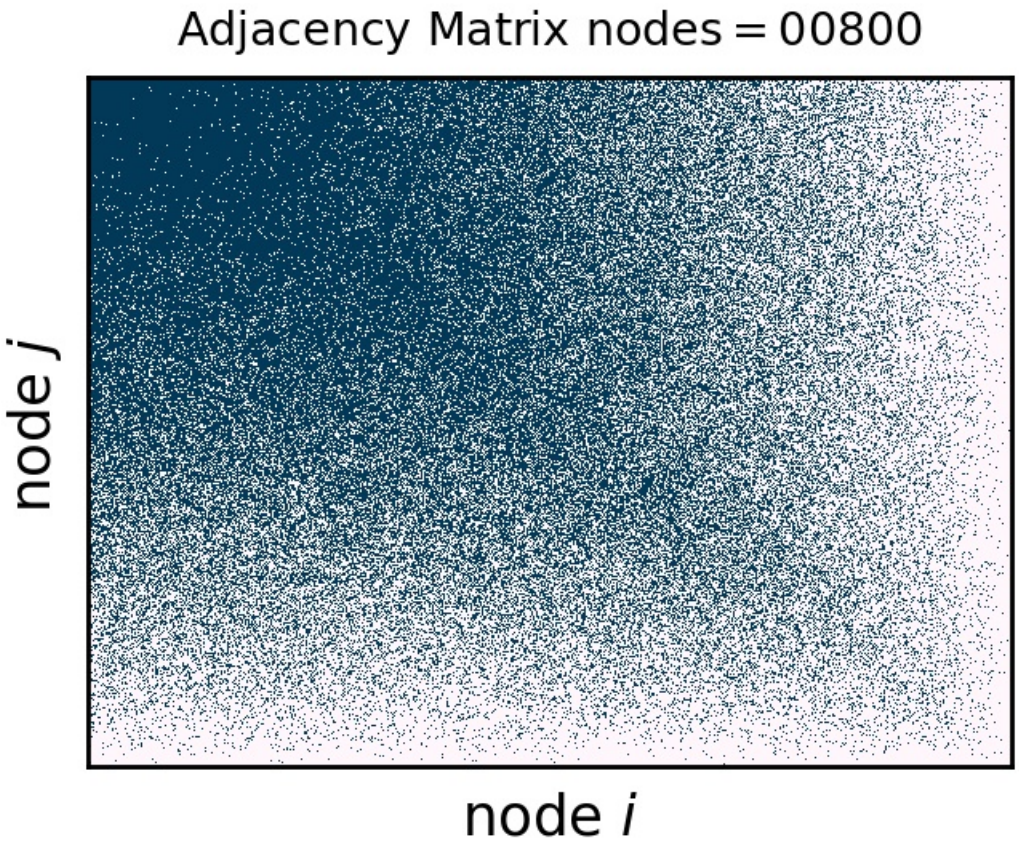

Adjacency Matrix nodes = 00300
node j
node i

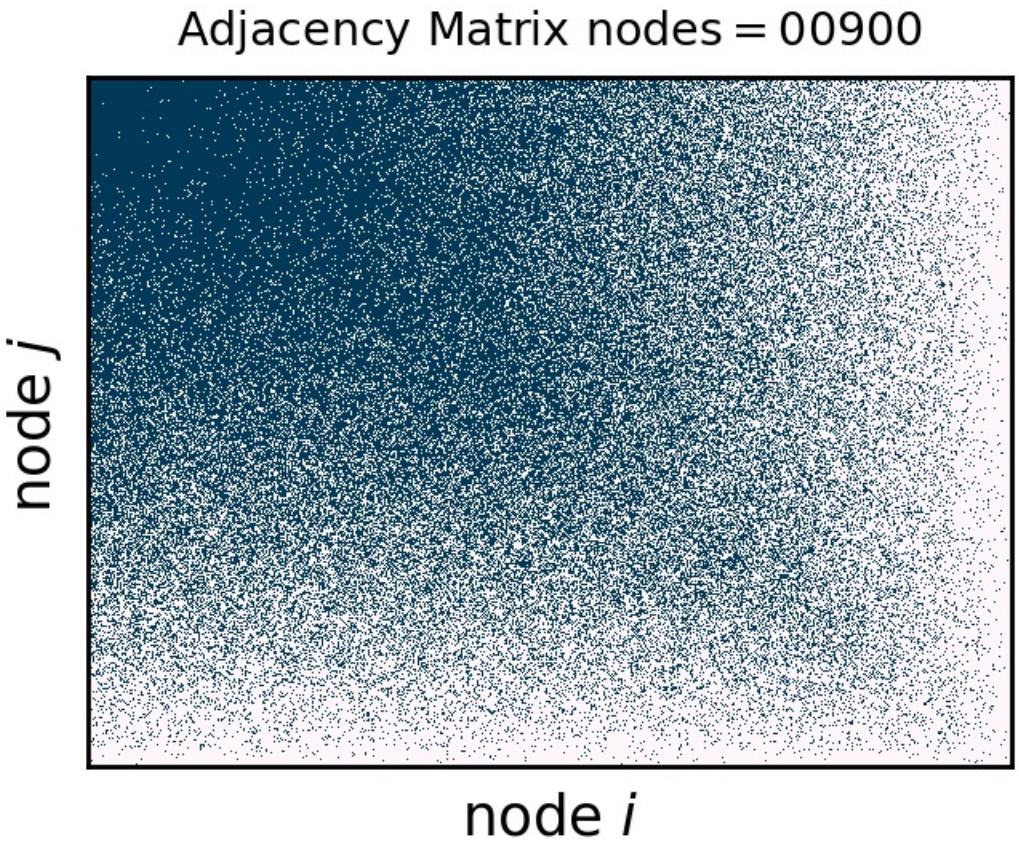

Adjacency Matrix nodes = 00400
node j
node i

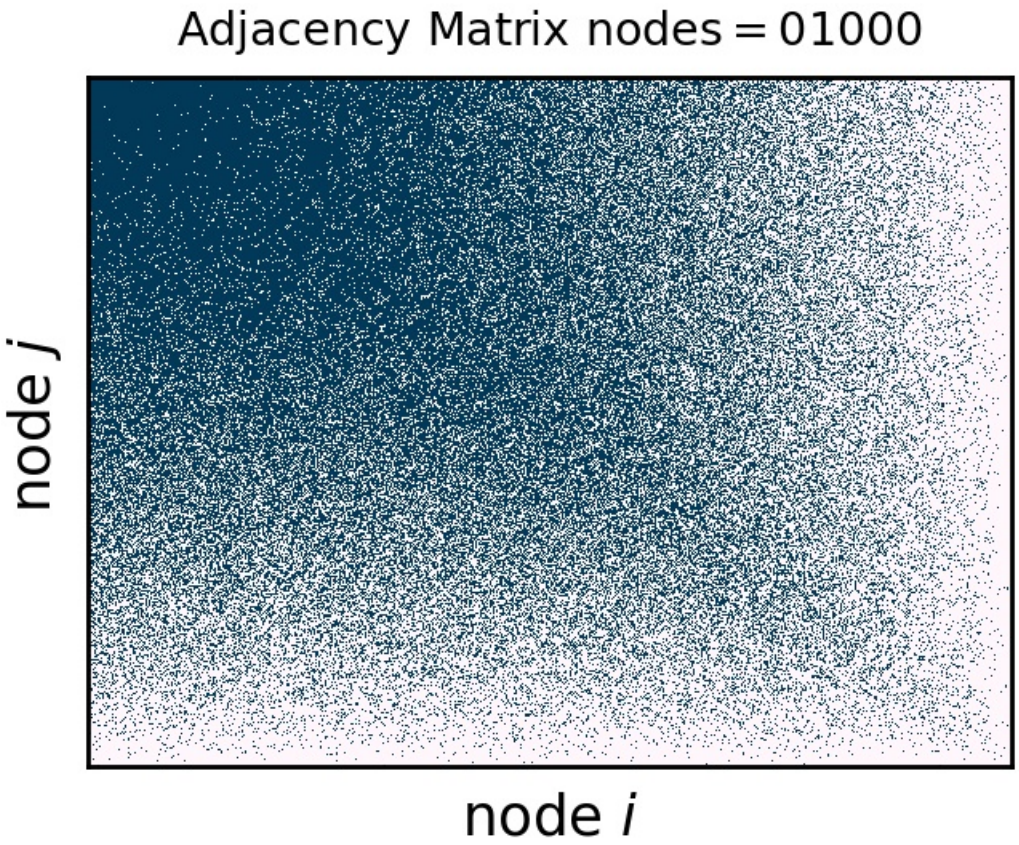

Adjacency Matrix nodes = 00500
node j
node i


Adjacency Matrix nodes = 00600
node j
node i


Adjacency Matrix nodes = 00700
node j
node i

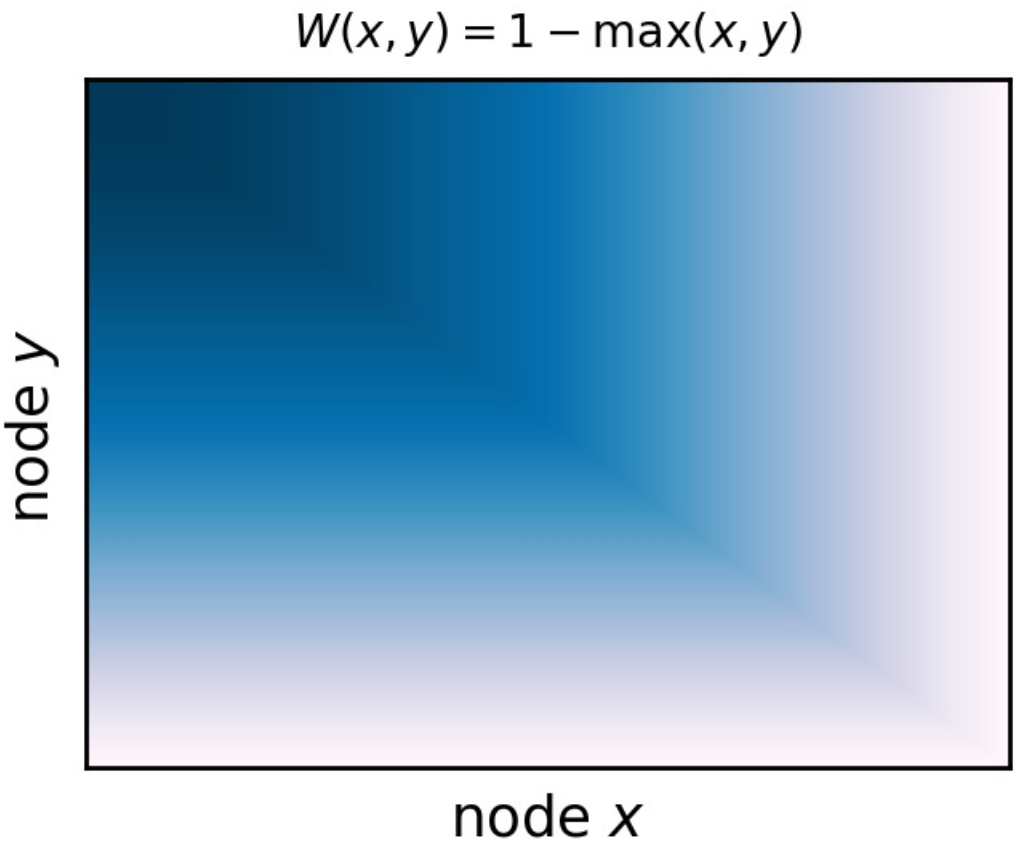

Adjacency Matrix nodes = 00800
node j
node i


Adjacency Matrix nodes = 00900
node j
node i

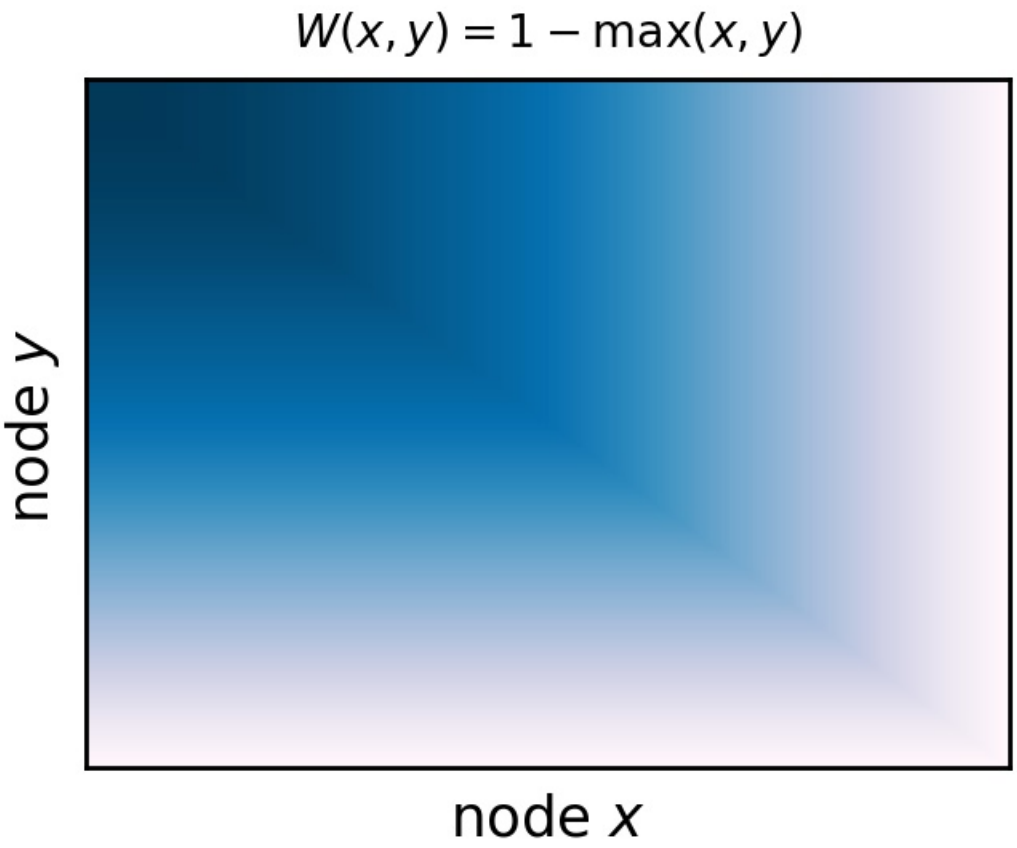

Adjacency Matrix nodes = 01000
node j
node i

Prendiamo la funzione $W(x, y) = 1 - \max(x, y)$

Prendiamo la funzione $W(x, y) = 1 - \max(x, y)$

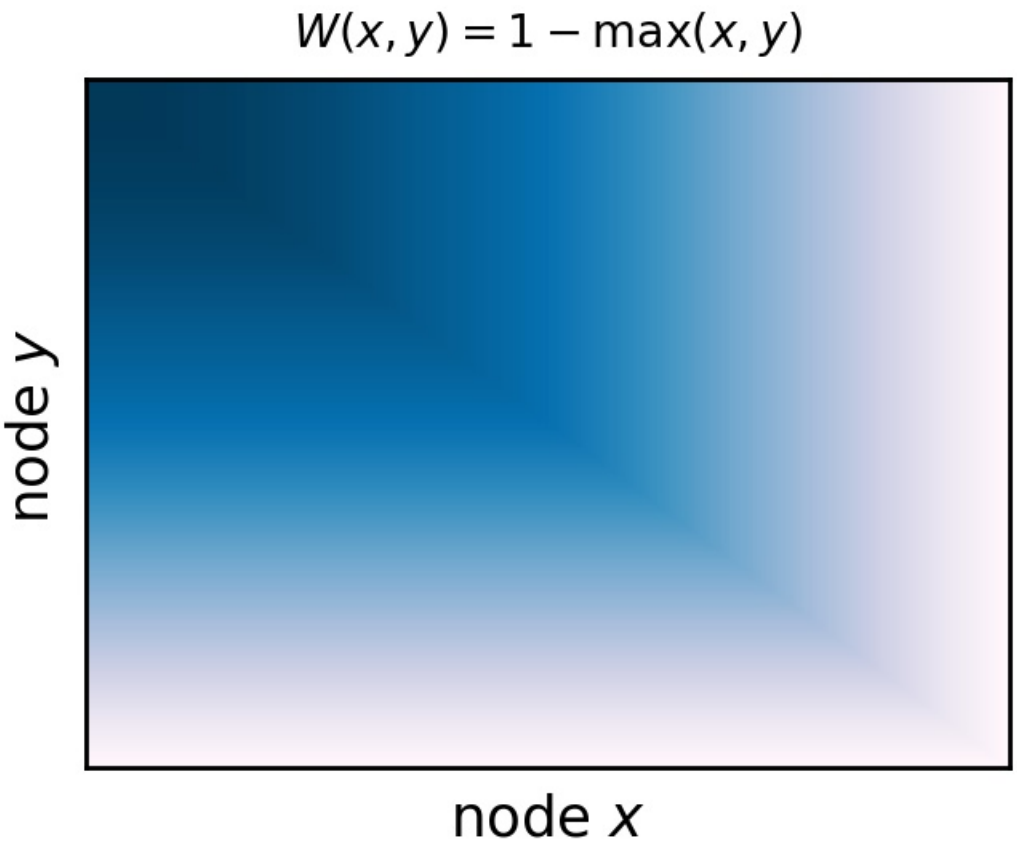


Prendiamo la funzione $W(x, y) = 1 - \max(x, y)$




Nel limite termodinamico la rete è rappresentata da una singola funzione, detta grafone **(graphon)**

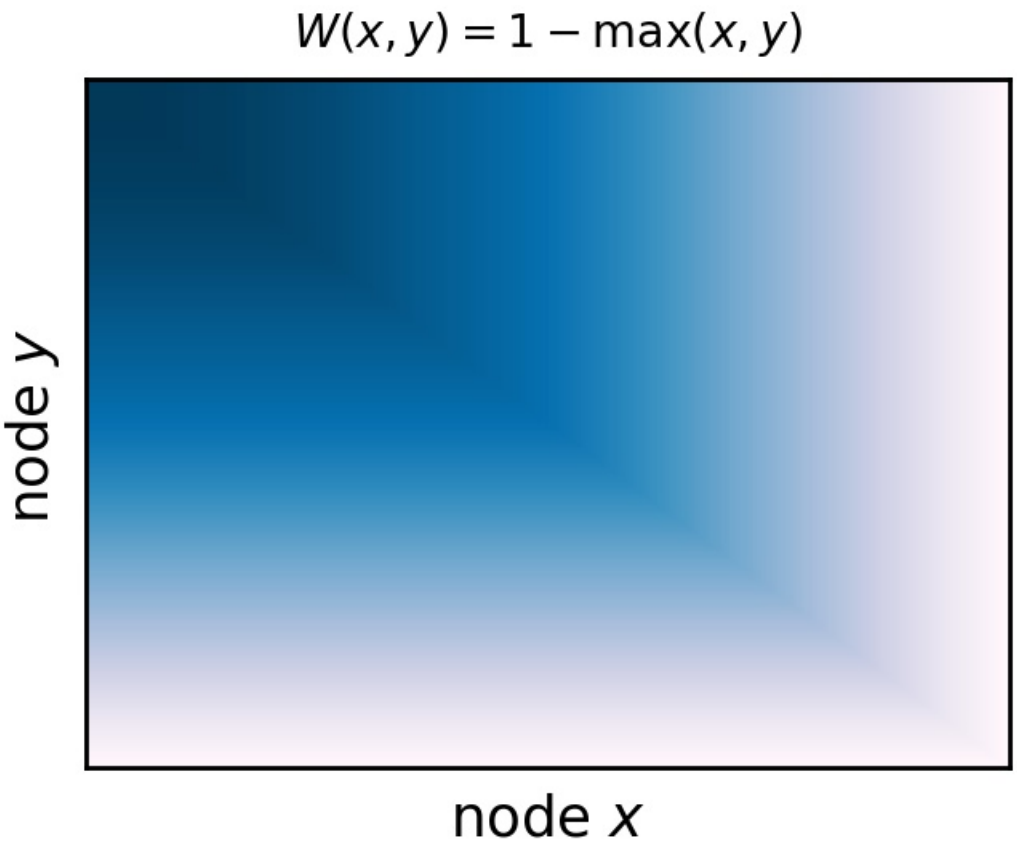
Prendiamo la funzione $W(x, y) = 1 - \max(x, y)$




Nel limite termodinamico la rete è rappresentata da una singola funzione, detta grafone **(graphon)**

**N.B.:** $W(x, y)$ è definita a meno di mappe biiettive che lasciano invariata la misura

Prendiamo la funzione $W(x, y) = 1 - \max(x, y)$




Nel limite termodinamico la rete è rappresentata da una singola funzione, detta grafone **(graphon)**

**N.B.:** $W(x, y)$ è definita a meno di mappe biiettive che lasciano invariata la misura
La convergenza al graphon è un teorema del limite centrale per reti

# Campionamento

# Campionamento

- Supponiamo di conoscere il graphon $W(x,y)$ di una rete (dobbiamo immaginare che la rete empirica appartenga ad una sequenza crescente)
- Vogliamo ottenere un campione della rete di size $n$
- Estraiamo numeri random $0 \le x_k \le 1,\ k = 1,\dots,n$ con distribuzione uniforme in $[0,1]$
- Disegnamo $n$ nodi e connettiamo i nodi $\ell$ e $k$ con prob. $W(x_\ell, x_k)$
- Si può dimostrare che il campione ottenuto è consistente, cioè ha proprietà simili al grafo di partenza

## Campionamento

- Supponiamo di conoscere il graphon $W(x,y)$ di una rete (dobbiamo immaginare che la rete empirica appartenga ad una sequenza crescente)
- Vogliamo ottenere un campione della rete di size $n$
- Estraiamo numeri random $0 \leq x_k \leq 1,\ k = 1,\ldots,n$ con distribuzione uniforme in $[0,1]$
- Disegnamo $n$ nodi e connettiamo i nodi $\ell$ e $k$ con prob. $W(x_\ell, x_k)$
- Si può dimostrare che il campione ottenuto è consistente, cioè ha proprietà simili al grafo di partenza

## Ricostruzione

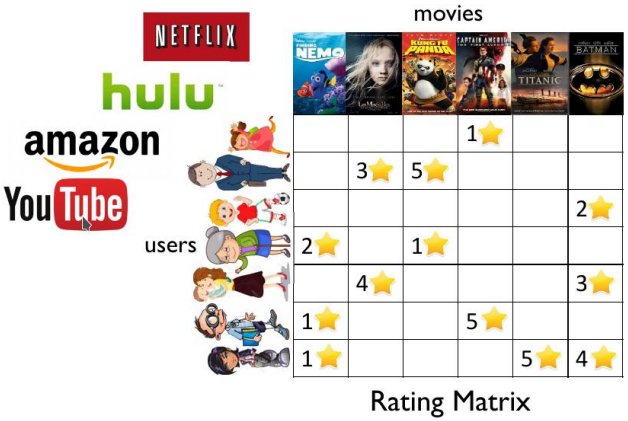
# Campionamento

- Supponiamo di conoscere il graphon $W(x,y)$ di una rete (dobbiamo immaginare che la rete empirica appartenga ad una sequenza crescente)
- Vogliamo ottenere un campione della rete di size $n$
- Estraiamo numeri random $0 \le x_k \le 1,\ k = 1,\dots,n$ con distribuzione uniforme in $[0,1]$
- Disegnamo $n$ nodi e connettiamo i nodi $\ell$ e $k$ con prob. $W(x_\ell, x_k)$
- Si può dimostrare che il campione ottenuto è consistente, cioè ha proprietà simili al grafo di partenza

# Ricostruzione

- Supponiamo di conoscere il graphon $W(x,y)$ di una rete di size $n$
- Vogliamo aggiungere nodi $m$ alla rete e connetterli in modo tale da non alterare le proprietà della rete rispetto al graphon
- Estraiamo numeri random $0 \le x_k \le 1,\ k = 1,\dots,n+m$ con distribuzione uniforme in $[0,1]$
- Aggiungiamo $m$ nodi al grafo, inizialmente sconnessi
- Per ogni coppia $(i,j)$ dove $1 \le i \le n$ e $n+1 \le j \le m$, connettiamo $(i,j)$ con prob. $W(x_i, x_j)$
- Per ogni coppia $(i,j)$ dove $n+1 \le i,j \le m$, connettiamo $(i,j)$ con prob. $W(x_i, x_j)$
- Si può dimostrare che la rete ottenuta è consistente, cioè ha proprietà simili alla rete di partenza

# Esempio: collaborative filtering




- Molte compagnie di servizi on-line costruiscono una rete bipartita utenti-prodotti
- Ad ogni (utente,prodotto) è associato un rating se il cliente ha acquistato il prodotto
- La rete ha dimensioni molto grandi (si pensi ad Amazon, YouTube, etc.)
- La matrice di adiacenza ha molti vuoti, cioè è largamente ignota rete sparsa
- Si può stimare il graphon $W(x, y)$ della rete ed effettuare una ricostruzione
- A quel punto le compagnie hanno un modo per suggerire prodotti nuovi ad utenti

# Summary

- Everywhere we look, we see large-scale networks – technological, social, economic, biological
- Modeling and analysis of these networks uses approaches from graph theory, combinatorics, probability, game theory, (local) algorithms
- Results include new theories, theorems, experimental predictions

  … even new business models

  … and new (personalized) drug therapies