Performance Evaluation

Lecture 1: Complex Networks

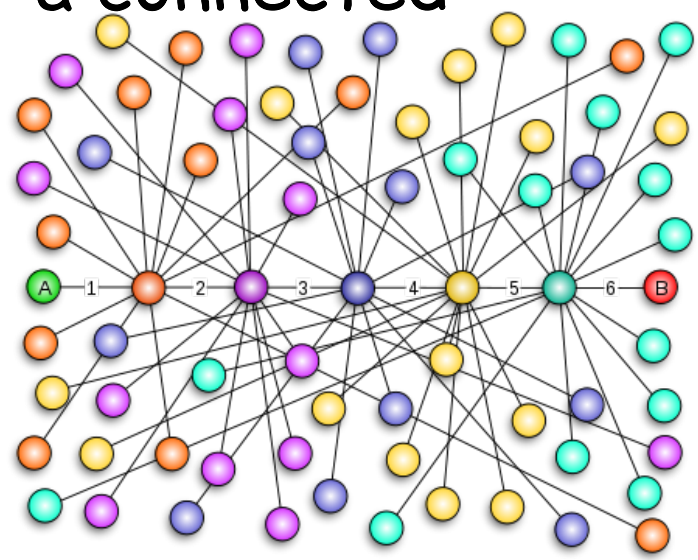
Giovanni Neglia
INRIA – EPI Maestro
10 December 2012

Outline

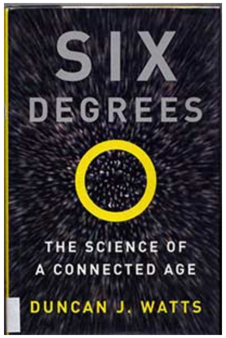
- Properties of Complex Networks
  - Small diameter
  - High Clustering
  - Hubs and heavy tails
- Physical causes
- Consequences

Small Diameter (after Milgram's experiment)

Six degrees - the science of a connected age, 2003



Six degrees of separation is the idea that everyone is on average approximately six steps away, by way of introduction, from any other person in the world, so that a chain of "a friend of a friend" statements can be made, on average, to connect any two people in six steps or fewer.



Small Diameter, more formally

- A linear network has diameter N-1 and average distance $\Theta(N)$
  - How to calculate it?
- A square grid has diameter and average distance $\Theta(\text{sqrt}(N))$
- Small World: diameter $O((\log(N))^a)$, $a>0$
- Lessons from model: long distance random connections are enough

Erdös-Rényi graph

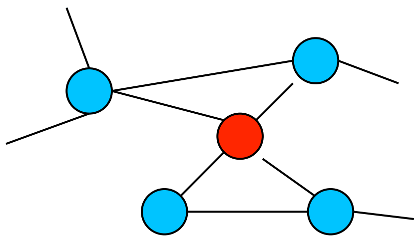
- A ER graph G(N,q) is a stochastic process
  - N nodes and edges are selected with prob. q
- Purpose: abstract from the details of a given graph and reach conclusions depending on its average features

Erdös-Rényi graph

- A ER graph $G(N,q)$ is a stochastic process
  - N nodes and edges are selected with prob. q
  - Degree distribution: $P(d)=C^{d}_{N-1}\, q^{d}(1-q)^{N-1-d}$
    - Average degree: $\langle d\rangle=q\,(N-1)$
    - For $N\to\infty$ and Nq constant: $P(d)=e^{-\langle d\rangle}\langle d\rangle^{d}/d!$
    - $\langle d^2\rangle=\langle d\rangle(1+\langle d\rangle)$
  - Average distance: $\langle l\rangle\approx \log N/\log\langle d\rangle$
    - Small world

Clustering

- "The friends of my friends are my friends"
- Local clustering coefficient of node i
  - (# of closed triplets with i at the center) / (# of triplets with node i at the center) = (links among i's neighbors of node i)/(potential links among i's neighbors)



$C_i=2/(4*3/2)=1/3$

- Global clustering coefficient
  - (total # of closed triplets)/(total # of triplets)
    - # of triplets = 3 # of triangles
  - Or $1/N \Sigma_i C_i$

Clustering

- In ER
  - $C \approx q \approx \langle d \rangle / N$

Clustering

- In real networks

Network	Size	$\langle k \rangle$	ℓ	ℓ_{rand}	C	C_{rand}	Reference	Nr.
WWW, site level, undir.	153 127	35.21	3.1	3.35	0.1078	0.00023	Adamic, 1999	1
Internet, domain level	3015–6209	3.52–4.11	3.7–3.76	6.36–6.18	0.18–0.3	0.001	Yook *et al.*, 2001a, Pastor-Satorras *et al.*, 2001	2
Movie actors	225 226	61	3.65	2.99	0.79	0.00027	Watts and Strogatz, 1998	3
LANL co-authorship	52 909	9.7	5.9	4.79	0.43	1.8×10^{-4}	Newman, 2001a, 2001b, 2001c	4
MEDLINE co-authorship	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	Newman, 2001a, 2001b, 2001c	5
SPIRES co-authorship	56 627	173	4.0	2.12	0.726	0.003	Newman, 2001a, 2001b, 2001c	6
NCSTRL co-authorship	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	Newman, 2001a, 2001b, 2001c	7
Math. co-authorship	70 975	3.9	9.5	8.2	0.59	5.4×10^{-5}	Barabási *et al.*, 2001	8
Neurosci. co-authorship	209 293	11.5	6	5.01	0.76	5.5×10^{-5}	Barabási *et al.*, 2001	9
E. coli, substrate graph	282	7.35	2.9	3.04	0.32	0.026	Wagner and Fell, 2000	10
E. coli, reaction graph	315	28.3	2.62	1.98	0.59	0.09	Wagner and Fell, 2000	11
Ythan estuary food web	134	8.7	2.43	2.26	0.22	0.06	Montoya and Solé, 2000	12
Silwood Park food web	154	4.75	3.40	3.23	0.15	0.03	Montoya and Solé, 2000	13
Words, co-occurrence	460.902	70.13	2.67	3.03	0.437	0.0001	Ferrer i Cancho and Solé, 2001	14
Words, synonyms	22 311	13.48	4.5	3.84	0.7	0.0006	Yook *et al.*, 2001b	15
Power grid	4941	2.67	18.7	12.4	0.08	0.005	Watts and Strogatz, 1998	16
C. Elegans	282	14	2.65	2.25	0.28	0.05	Watts and Strogatz, 1998	17

Good matching for avg distance,
Bad matching for clustering coefficient

How to model real networks?

Regular Graphs have a high clustering coefficient

but also a high diameter

Random Graphs have a low diameter

but a low clustering coefficient

--> Combine both to model real networks: the Watts and Strogatz model

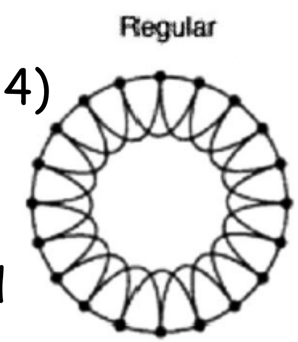
Regular Graph (k=4)

Long paths

$L = n/(2k)$

Highly clustered

$C=3/4$



Random Graph (k=4)

Short path length

$L=\log_k N$

Almost no clustering

$C=k/n$

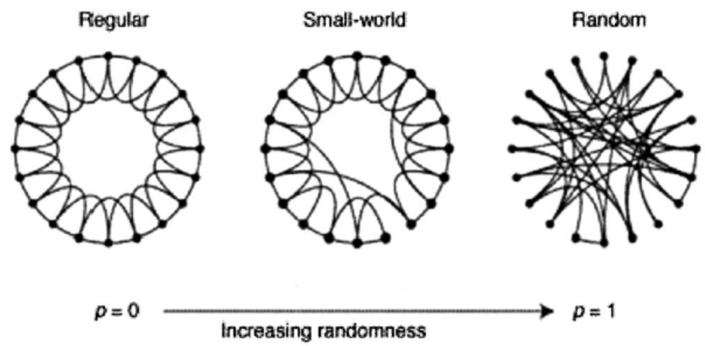
Regular ring lattice

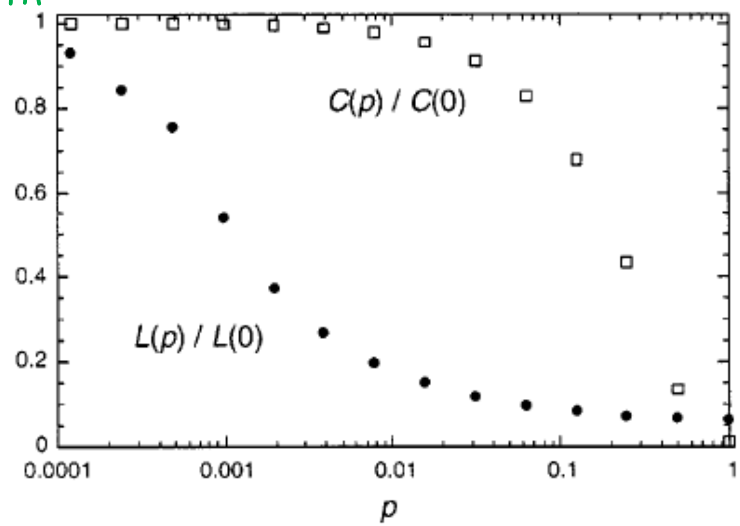
R. Albert and A.-L. Barabasi: Statistical mechanics of complex networks, Rev. Mod. Phys., Vol. 74, No. 1, January 2002

Watts and Strogatz model

Random rewiring of regular graph

- With probability *p* rewire each link in a regular graph to a randomly selected node
- Resulting graph has properties, both of regular and random graphs
  - --> High clustering and short path length





R. Albert and A.-L. Barabasi: Statistical mechanics of complex networks, Rev. Mod. Phys., Vol. 74, No. 1, January 2002

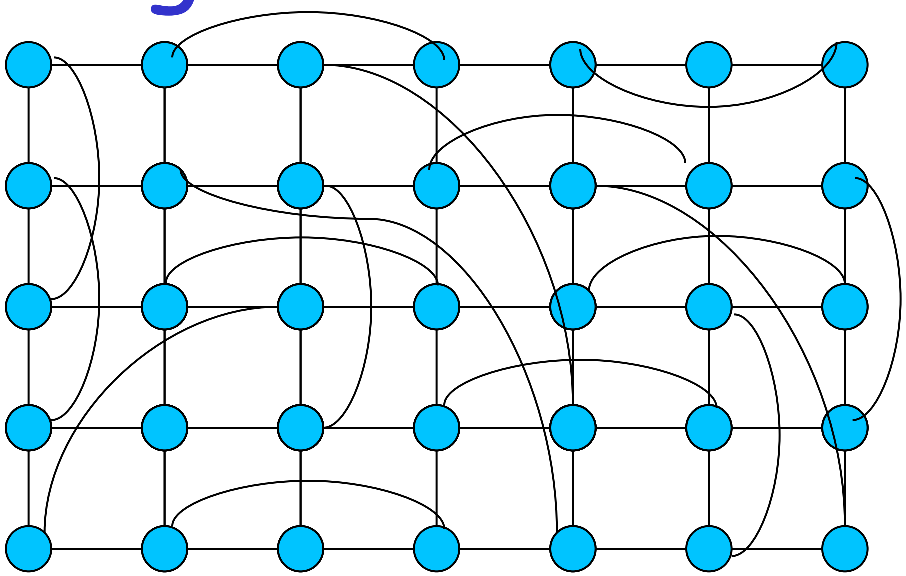
Small World

- Usually to denote
  - small diameter + high clustering

Intermezzo: navigation

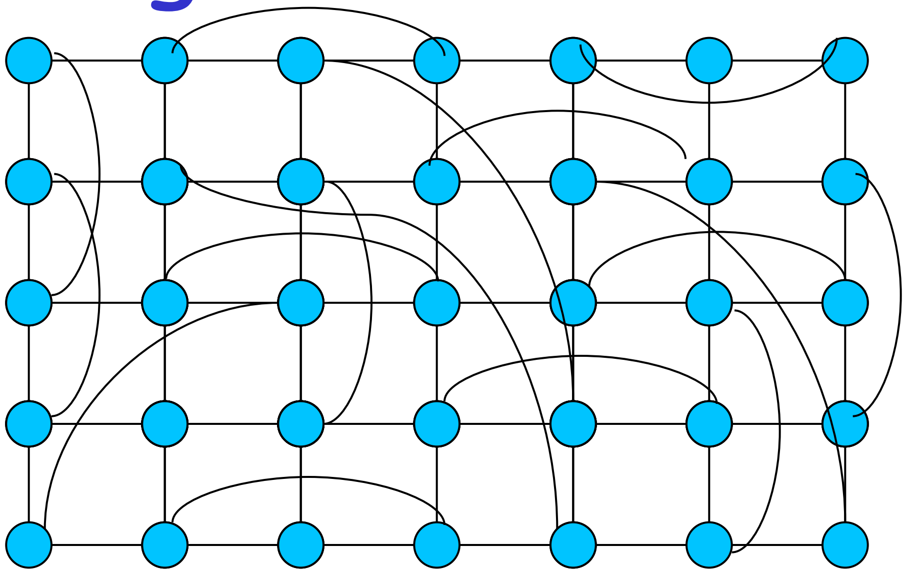
- In Small world nets there are short paths $O((\log(N))^a)$
- But can we find them?
  - Milgram's experiment suggests nodes can find them using only local information
  - Standard routing algorithms require O(N) information

Kleinberg's result



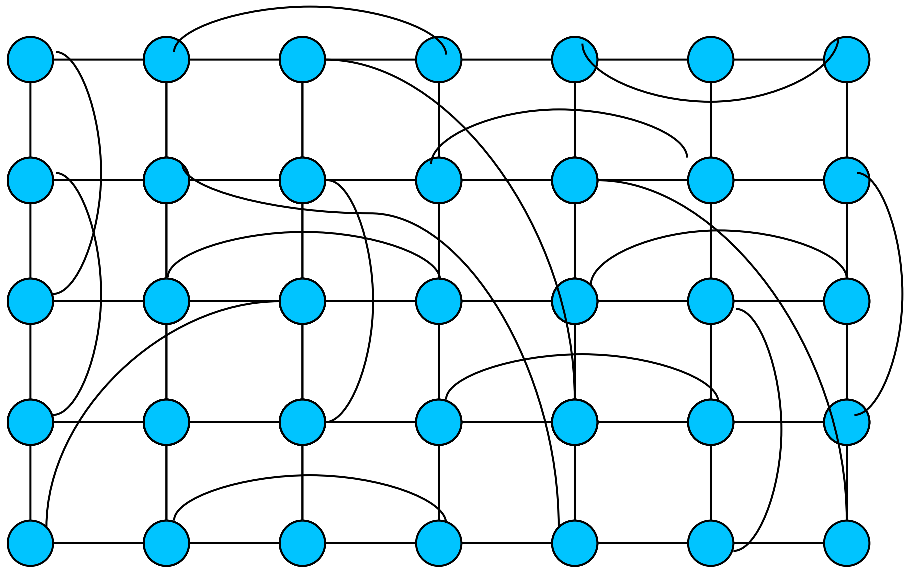
- Model: Each node has
  - Short-range connections
  - 1 long-range connection, up to distance r with probability prop. to $r^{-\alpha}$
  - For $\alpha=0$ it is similar to Watts-Strogatz model: there are short-paths

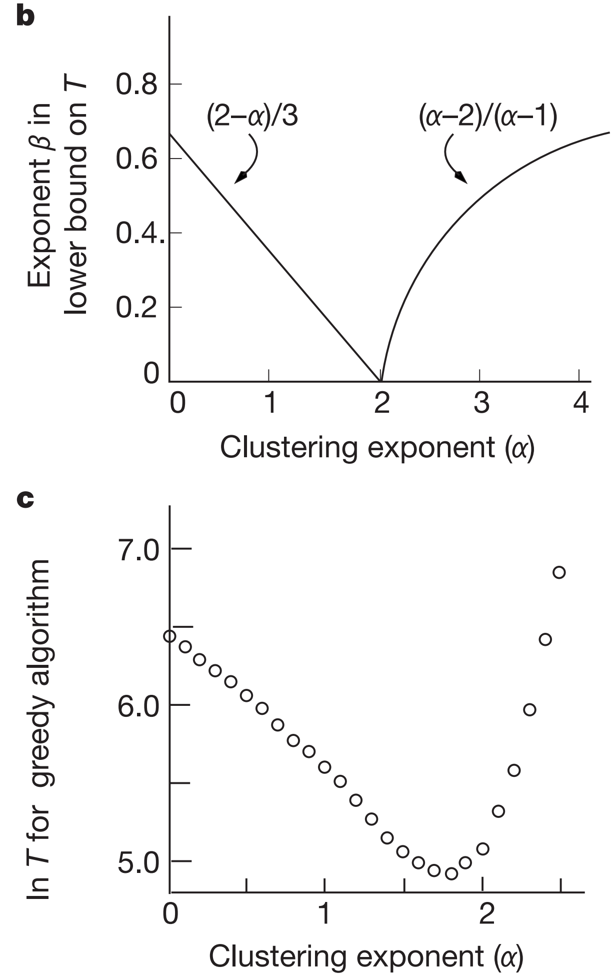
Kleinberg's result



- If $\alpha=2$ the greedy algorithm (forward the packet to the neighbor with position closest to the destination) achieves avg path length $O((\log(N))^2)$

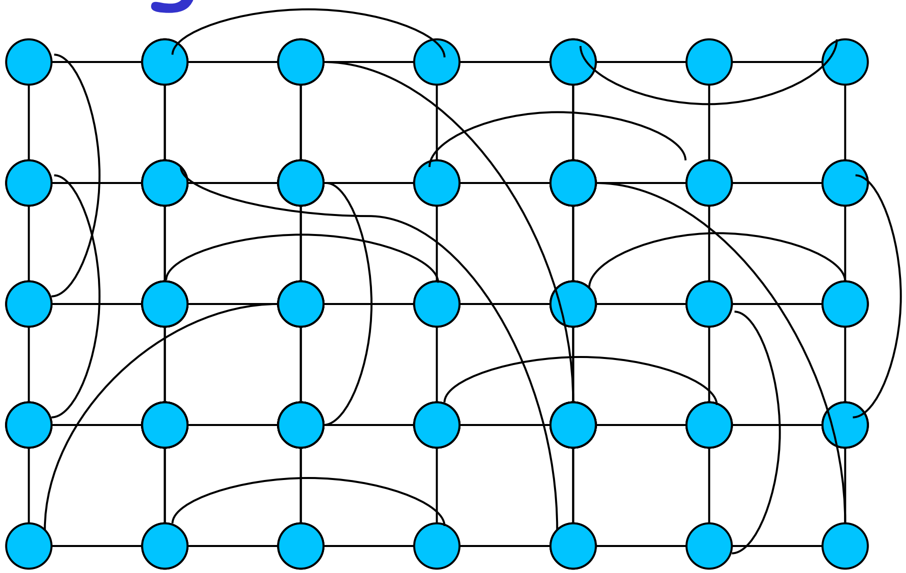
Kleinberg's result





- If α<>2 no local information algorithm can take advantage of small world properties
  - avg path length $\Omega(N^{\beta/2})$
    - where $\beta=(2-\alpha)/3$ for $0<=\alpha<=2$, $\beta=(\alpha-2)/(\alpha-1)$, for $\alpha>2$

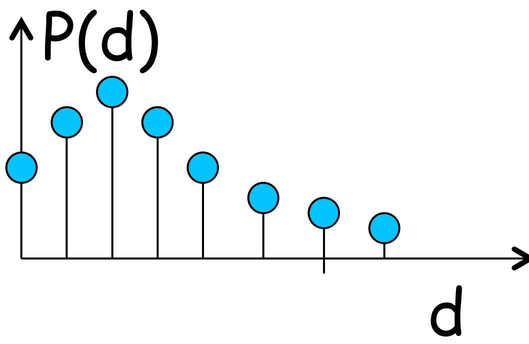
Kleinberg's result



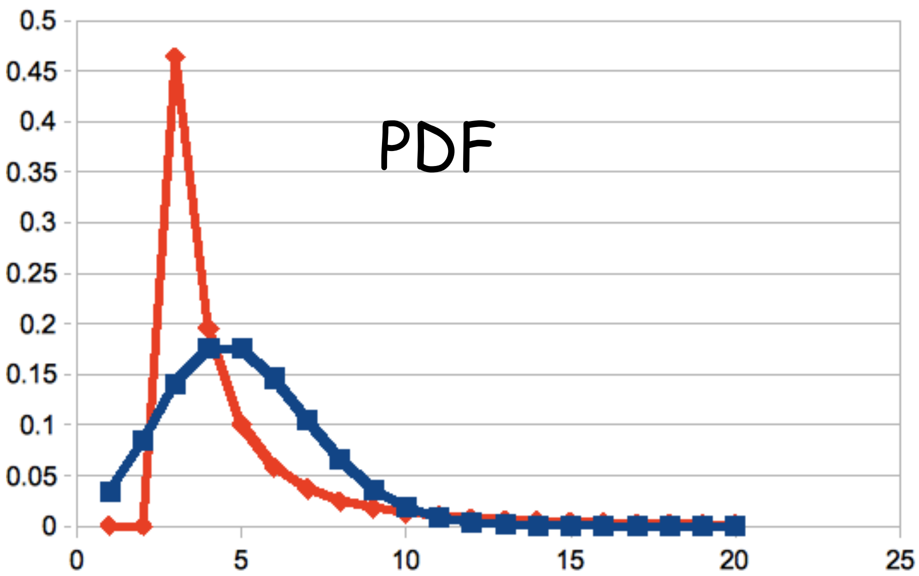
- Conclusions
  - The larger α the less distant long-range contacts move the message, but the more nodes can take advantage of their "geographic structure"
  - α=2 achieved the best trade-off

Hubs

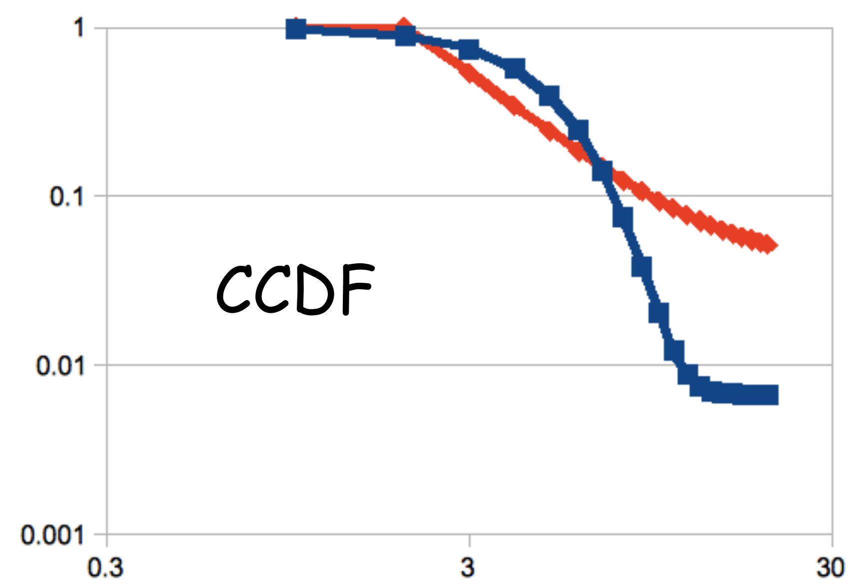
- 80/20 rule
  - few nodes with degree much higher than the average
  - a lot of nodes with degree smaller than the average
  - (imagine Bill Clinton enters this room, how rapresentative is the avg income)
- ER with N=1000, $\langle d \rangle = 5$, $P(d) \approx e^{-\langle d \rangle} \langle d \rangle^d / d!$
  - #nodes with d=10: $N*P(10) \approx 18$
  - #nodes with d=20: $N*P(20) \approx 2.6\ 10^{-4}$



Hubs

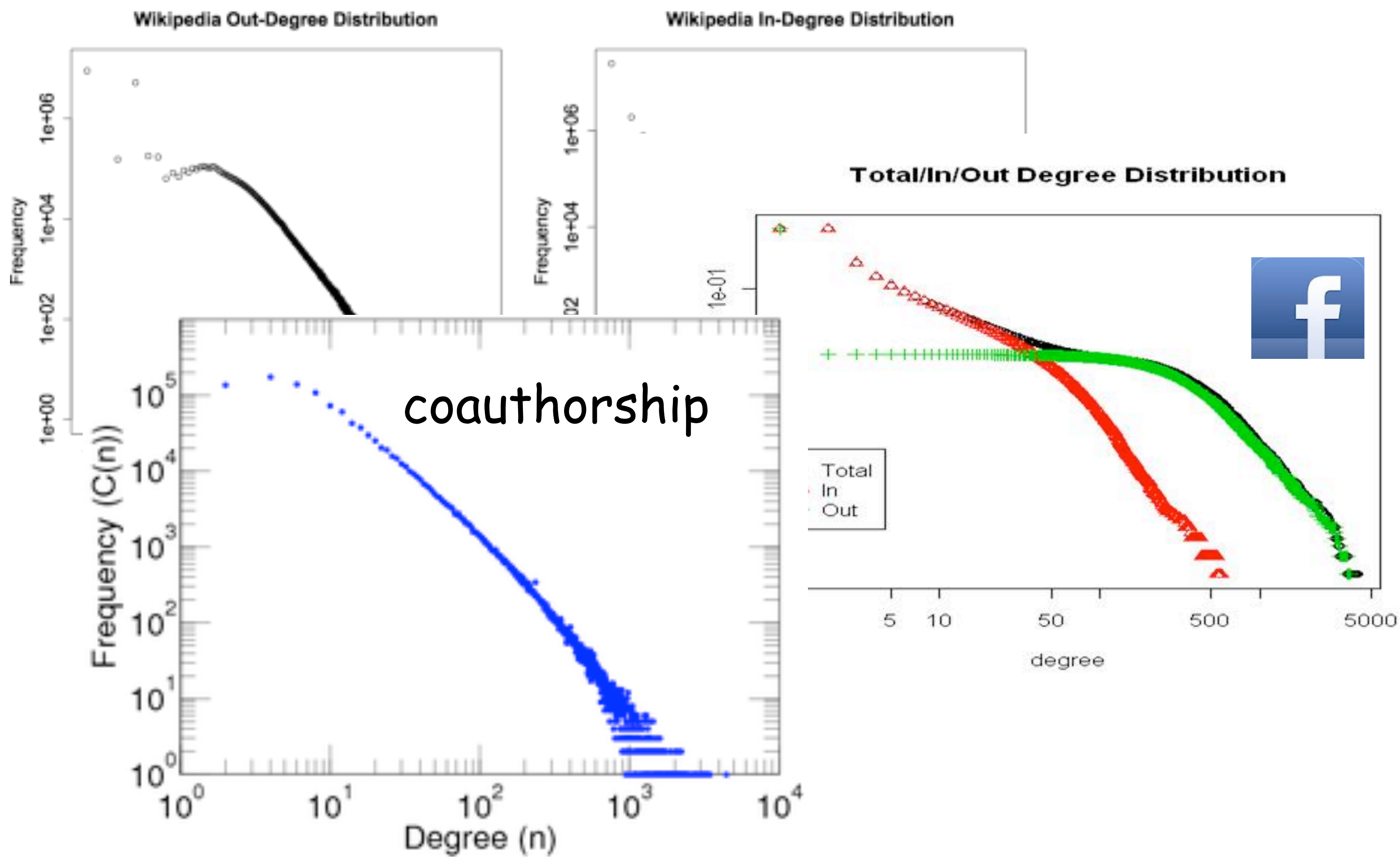


ER
Power law

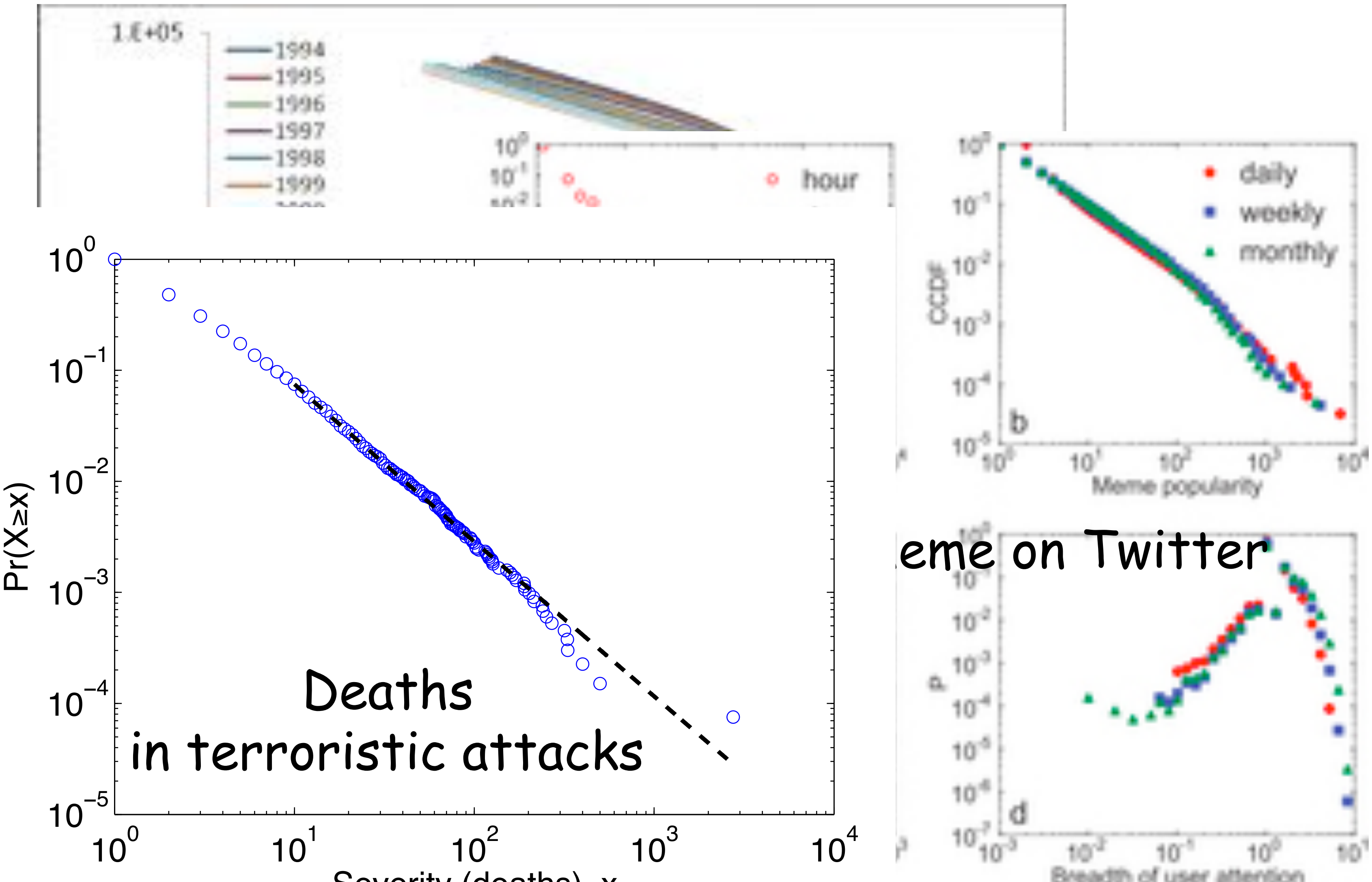


Power law:
$P(d) \sim d^{-\alpha}$

Power law degree distributions



… and more


1.E+05
1994
1995
1996
1997
1998
1999
hour
daily
weekly
monthly
CCDF
b
Meme popularity
eme on Twitter
P
d
Breadth of user attention
Pr(X≥x)
Severity (deaths), x
Deaths
in terroristic attacks

Power Law

- Where does it come from?
  - Albert-Barabasi's growth model
  - Highly Optimized Model
  - And other models
    - See Michael Mitzenmacher, A Brief History of Generative Models for Power Law and Lognormal Distributions

Albert-Barabasi's model

- Two elements
  - Growth
    - m_0 initial nodes, every time unit we add a new node with m links to existing nodes
  - Preferential attachment
    - The new node links to a node with degree k_i with probability

$$\Pi(k_i) = \frac{k_i}{\sum_{j=1,N} k_j}$$

Albert-Barabasi's model

- Node i arrives at time t_i, its degree keeps increasing
- With a continuum approximation:

$$\frac{\partial k_i}{\partial t} = \frac{m k_i}{\sum_{j=1,N} k_j} = \frac{m k_i}{2tm} = \frac{k_i}{2t} \longrightarrow k_i(t) = m\left(\frac{t}{t_i}\right)^{\beta}, \beta = \frac{1}{2}$$

- Then degree distribution at time t is:

$$P(k_i(t) < k) = P(t_i > t\,\frac{m^{1/\beta}}{k^{1/\beta}})$$

Albert-Barabasi's model

- At time t there are m_0+t nodes, if we consider that the t nodes are added uniformly at random in [0,t], then

$$P(t_i > x) = \frac{t - x}{t + m_0}$$

$$P(k_i(t) < k) = \frac{t}{t + m_0}\left(1 - \frac{m^{1/\beta}}{k^{1/\beta}}\right)$$

Albert-Barabasi's model

- The PDF is

$$P(k_i(t)=k)=\frac{\partial P(k_i(t)\le k)}{\partial k}=\frac{t}{t+m_0}\frac{1}{\beta}\frac{m^{1/\beta}}{k^{1/\beta+1}}$$

- For t->∞

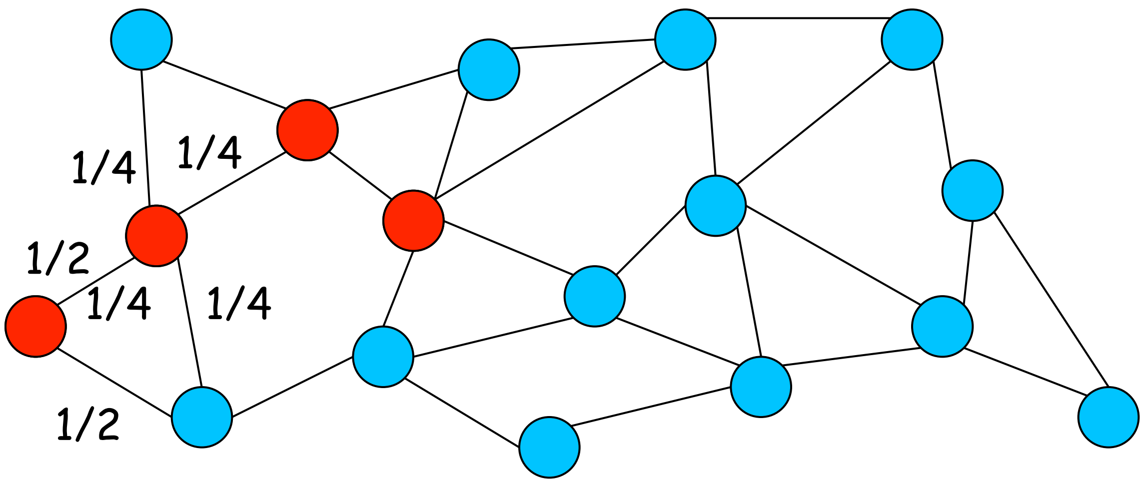
$$P(k_i(t)=k)\xrightarrow[t\to\infty]{}\frac{1}{\beta}\frac{m^{1/\beta}}{k^{1/\beta+1}}\propto k^{-\gamma},\ \ \gamma=3$$

Albert-Barabasi's model

- If $\Pi(k_i) \propto a + k_i$, $P(k) \propto k^{-\gamma}$, $\gamma = 3 + \frac{a}{m}$
- Other variants:
  - With fitness $\Pi(k) = \frac{\eta_i k_i}{\sum_{j=1,N} \eta_j k_j}$
  - With rewiring (a prob. p to rewire an existing connection)
  - Uniform attaching with "aging": A vertex is deactivated with a prob. proportional to $(k_i+a)^{-1}$

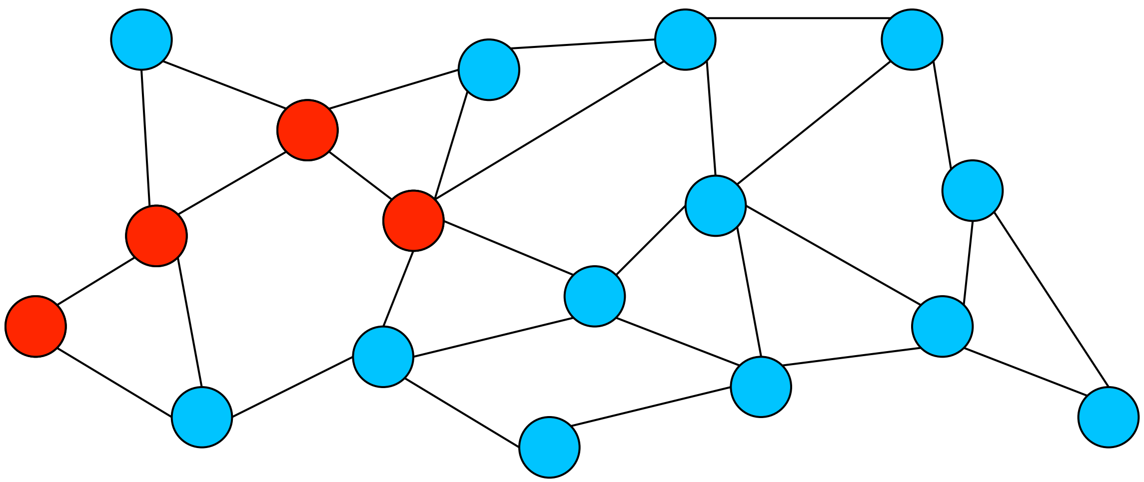
Back to Navigation: Random Walks

- What can we do in networks without a geographical structure?
  - Random walks



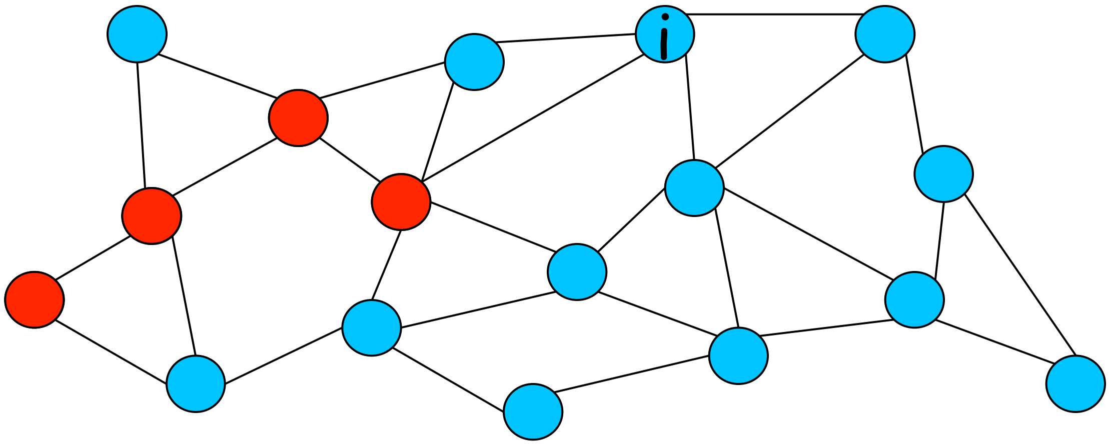
Back to Navigation: Random Walks

- How much time is needed in order to reach a given node?



Random Walks: stationary distribution

- $$\pi_i = \sum_{j \in N_i} \frac{1}{k_j} \pi_i$$
- $$\pi_i = \frac{k_i}{\sum_{i=1}^{N} k_j} = \frac{k_i}{2M}$$



- avg time to come back to node i starting from node i: $$\frac{1}{\pi_i} = \frac{2M}{k_i}$$
- Avg time to reach node i
  - intuitively $\approx \Theta(M/k_i)$

Another justification

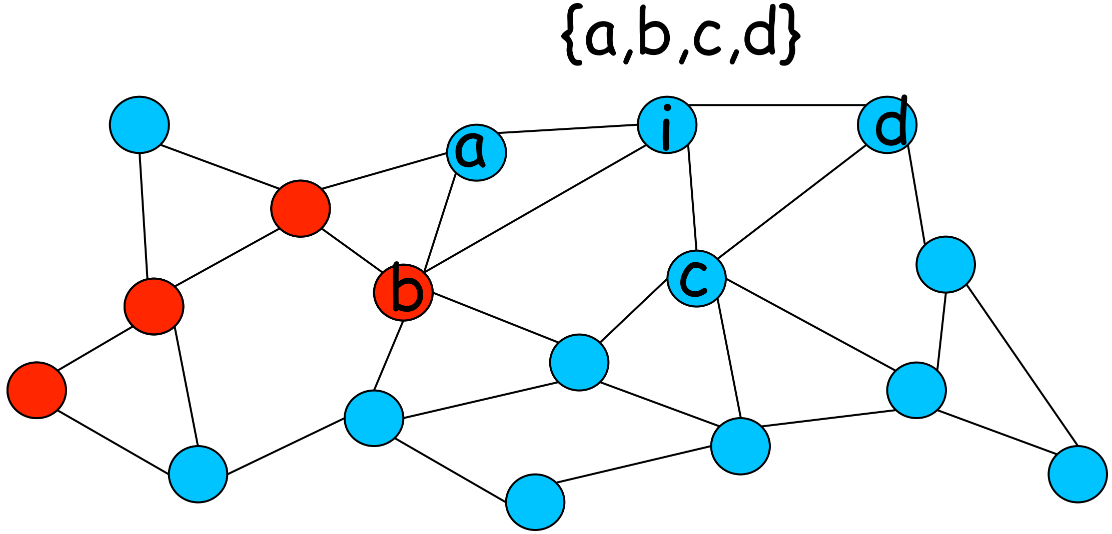
- Random walk as random edge sampling
  - Prob. to pick an edge (and a direction) leading to a node of degree k is $\frac{kp_k}{<k>}$
  - Prob. to arrive to a given node of degree k:

$$\frac{kp_k}{p_k N <k>} = \frac{k}{2M}$$

  - Avg. time to arrive to this node 2M/k

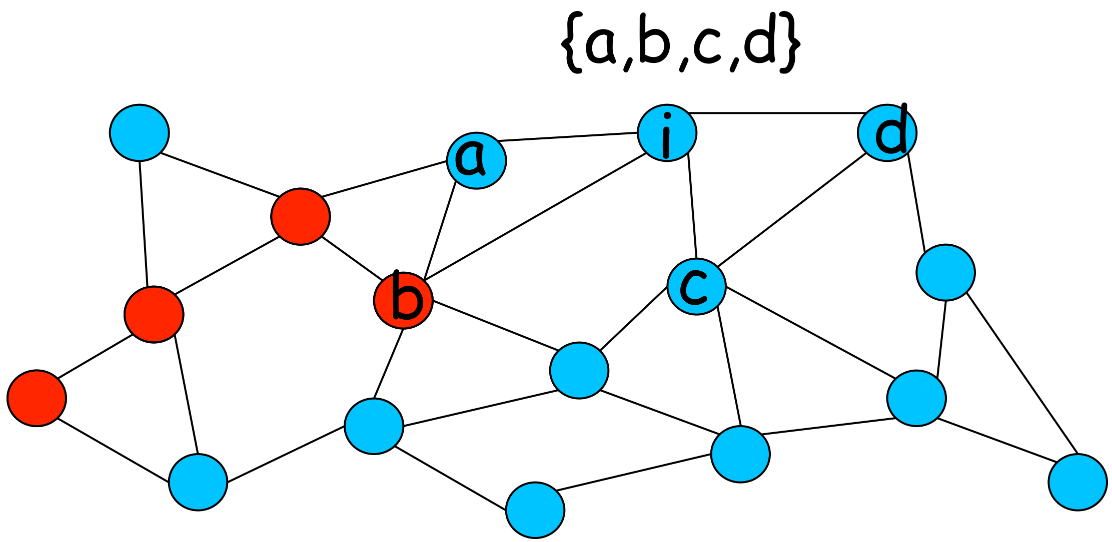
Distributed navigation
(speed up random walks)

- Every node knows its neighbors



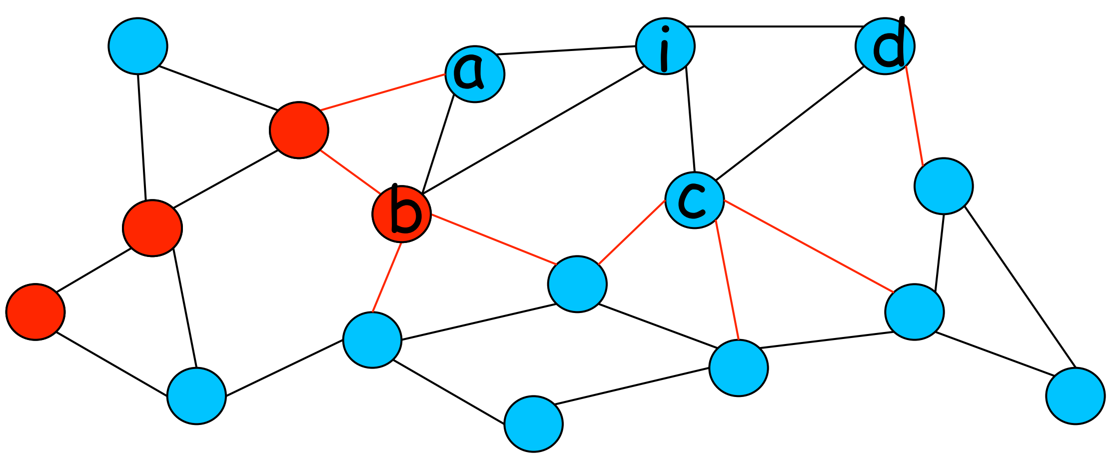
Distributed navigation
(speed up random walks)

- Every node knows its neighbors
- If a random walk looking for *i* arrives in *a* the message is directly forwarded to *i*



Distributed navigation
reasoning 1

- We discover *i* when we sample one of the links of *i*'s neighbors
- Avg # of these links: $k_i \sum_k \left((k-1) \frac{k p_k}{<k>} \right) = k_i \left(\frac{<k^2>}{<k>} - 1 \right)$
- Prob. to arrive at one of them: $\frac{k_i}{2M} \left(\frac{<k^2>}{<k>} - 1 \right)$



Distributed navigation
reasoning 2

- Prob that a node of degree k is neighbor of node *i*

$$1-\left(1-\frac{k_i}{2M}\right)^{k-1} \approx \frac{k_i(k-1)}{2M}$$

- Prob that the next edge brings to a node that is neighbor of node i:

$$\sum_k \frac{k_i(k-1)}{2M}\frac{kp_k}{<k>} = \frac{k_i}{2M}\left(\frac{<k^2>}{<k>}-1\right)$$

Distributed navigation

- Avg. Hop# $\dfrac{2M}{k_i}\dfrac{<k>}{<k^2>-<k>}$
  - Regular graph with degree d: $\dfrac{2M}{d(d-1)}$
  - ER with <k>: $\dfrac{2M}{k_i(<k>-1)}$
  - Pareto distribution $\left(P(k)\approx\dfrac{\alpha x_m^{\alpha}}{x^{\alpha+1}}\right)$:

$$\approx\frac{2M}{k_i}\frac{(\alpha-2)(\alpha-1)}{x_m-(\alpha-2)(\alpha-1)}$$

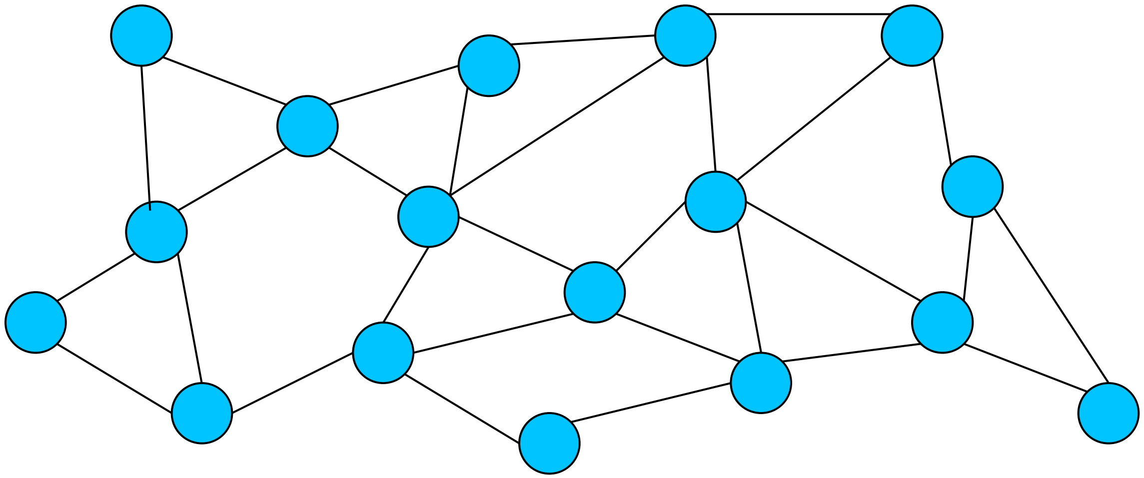
If α->2…

Distributed navigation

- Application example:
  - File search in unstructured P2P networks through RWs

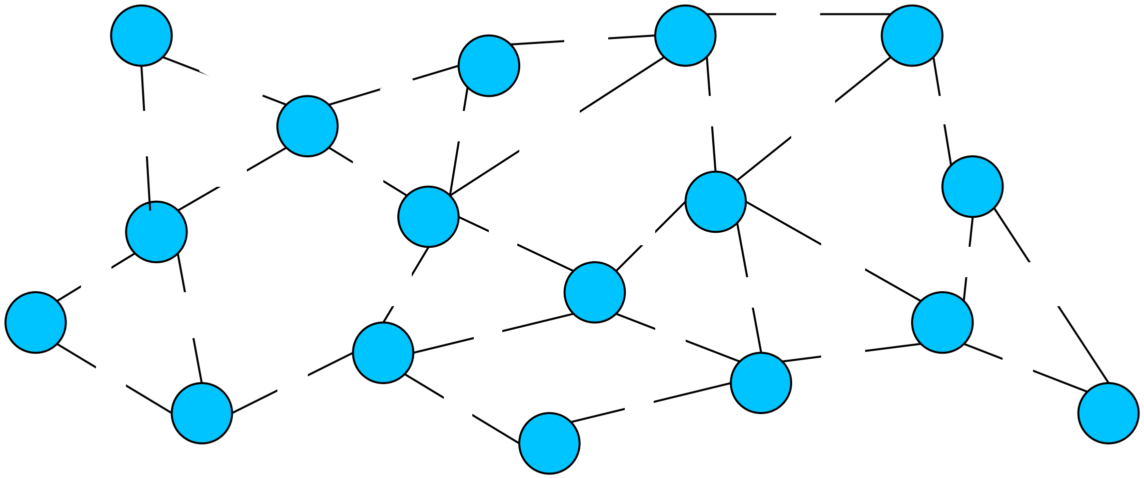
Configuration model

- A family of random graphs with given degree distribution



Configuration model

- A family of random graphs with given degree distribution
  - Uniform random matching of stubs



Configuration model

- A family of random graphs with given degree distribution
  - Uniform random matching of stubs