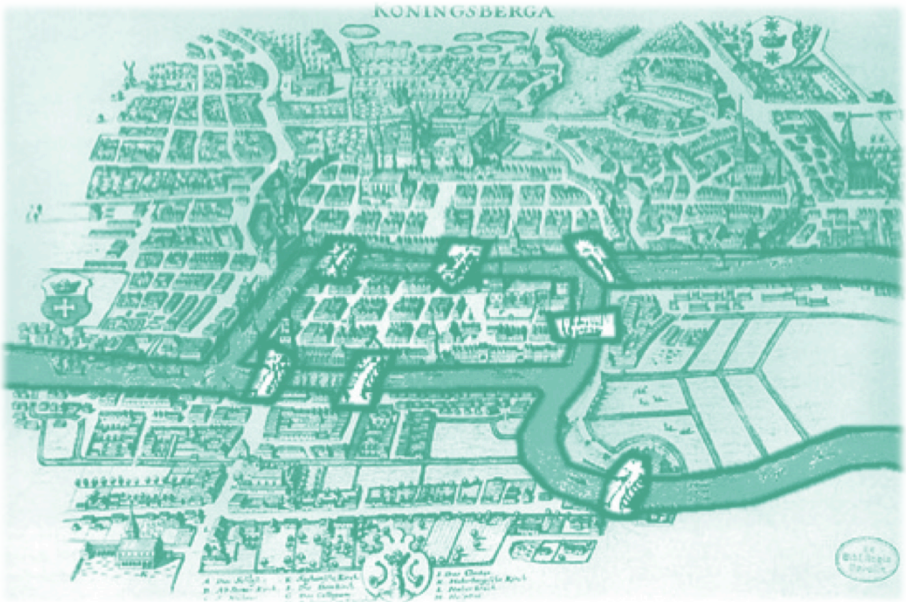
LIFE DATA EPIDEMIOLOGY

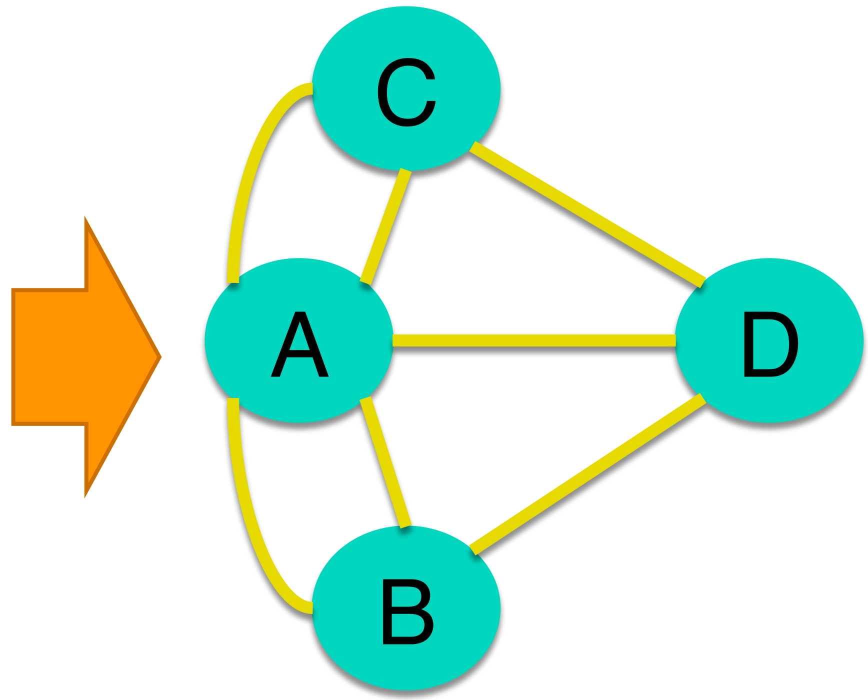
Lecture 7: Networks

Leonardo Badia

leonardo.badia@unipd.it

Networks as graphs





- Graph: $\mathcal{G}(\mathcal{V},\mathcal{E})$
  - Vertices (set $\mathcal{V}$): nodes, users, elements
  - Edges (set $\mathcal{E}$): links, arcs, hops, connections

Degree

- The **degree** k_j of a node j in an undirected network $\mathcal{G}(\mathcal{V},\mathcal{E})$ is the number of links it has to other nodes
  - or, differently said, #of nodes j is linked to
  - the average degree is $\langle k \rangle = (\sum_j k_j) / |\mathcal{V}|$
- The number of links $|\mathcal{E}| = L = ½ (\sum_j k_j)$
  - ½ because each link is counted twice
  - If $|\mathcal{V}| = N$, average degree $\langle k \rangle = 2L / N$

Why random networks?

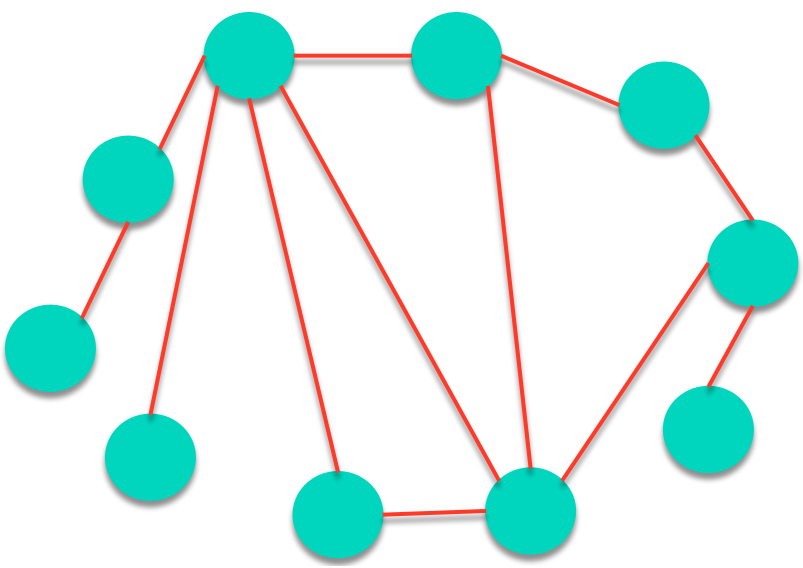
- It seems that many connections arising in real networks are unpredictable
- Idea: graph whose links between nodes are randomly generated with probability *p*
  - note: "**random**" = iid distributed
- This seems sensible as we often observe unexpected links

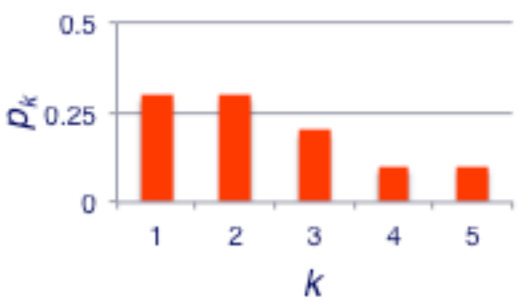
Random networks

- This simple model is often referred to as the **Erdős-Rényi model**
  - actually, the original model proposed by Erdős and Rényi involved a fixed set $\mathcal{V}$ and a fixed number of links randomly placed
  - the present model with a given probability *p* (where therefore the number of links is variable) is nevertheless very similar

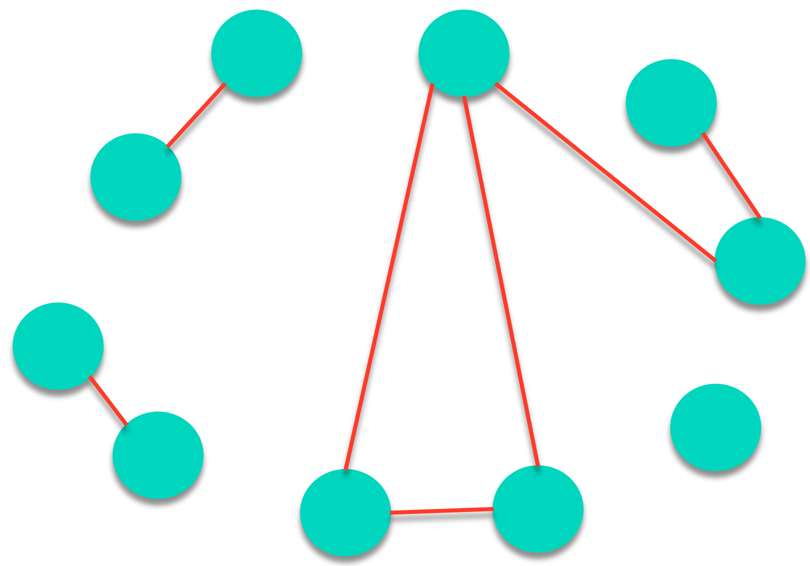
Random networks

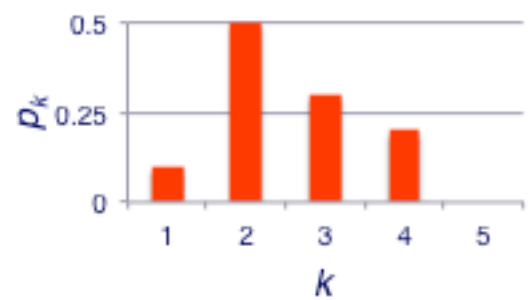
- Highly diverse networks created this way





L = 12





L = 7

How many links?

- Max number of links is $\binom{N}{2} = N(N-1)/2$
- The probability P_L that we have L links is:

$$P_L = \binom{\binom{N}{2}}{L} p^L (1-p)^{\binom{N}{2}-L} \quad \text{(binomial)}$$

- Hence: $\langle L \rangle = pN(N-1)/2, \ \langle k \rangle = p(N-1)$

Degree distribution

- Also binomial! Probability p_k that a node is connected to exactly k neighbors:

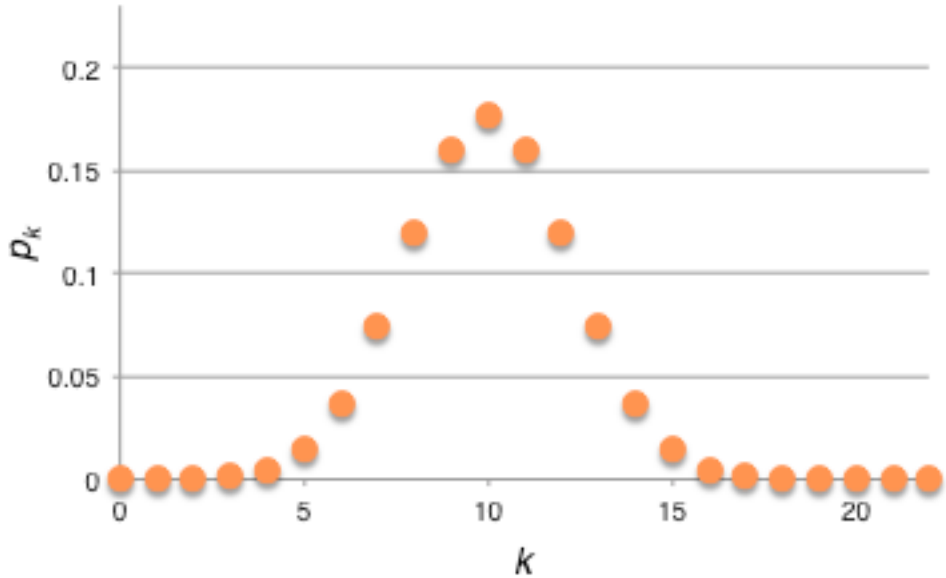
$$p_k = \binom{N-1}{k} p^k (1-p)^{N-1-k}$$

  - as already shown, $\langle k \rangle = p\,(N-1)$
  - also, variance of k: $\sigma_k^2 = p\,(1-p)\,(N-1)$

$$\frac{\sigma_k}{\langle k \rangle} = \sqrt{\frac{1-p}{p(N-1)}} \xrightarrow[N\to\infty]{} 0 \quad \text{(narrow for large } N\text{)}$$

Degree distribution

- Degree distribution - binomial



- $N = 21, \quad p = ½$

- Can be used the other way around

$$\langle k \rangle = p\,(N-1) \Rightarrow \quad p = \langle k \rangle / (N-1)$$

Degree distribution

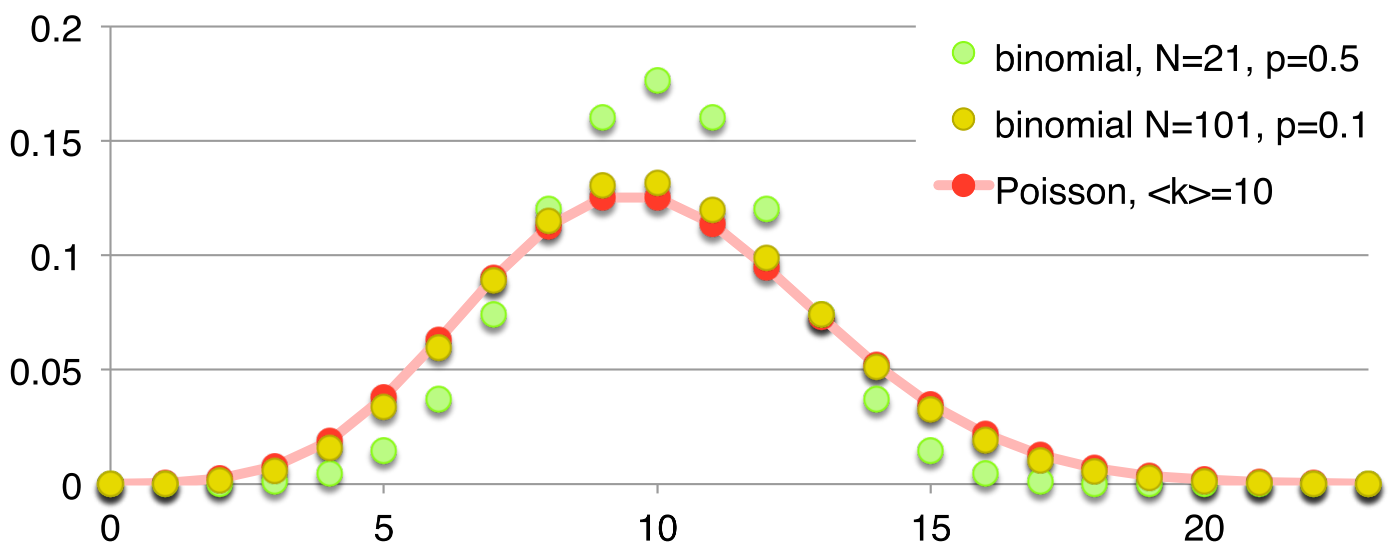
- Since real networks are sparse, $k \ll N$-1
  - hence p also small, binomial → Poisson

$$(1-p)^{N-1-k} \approx e^{(N-1-k)\log(1-\langle k \rangle/(N-1))} \xrightarrow[N\to\infty]{} e^{-\langle k \rangle}$$

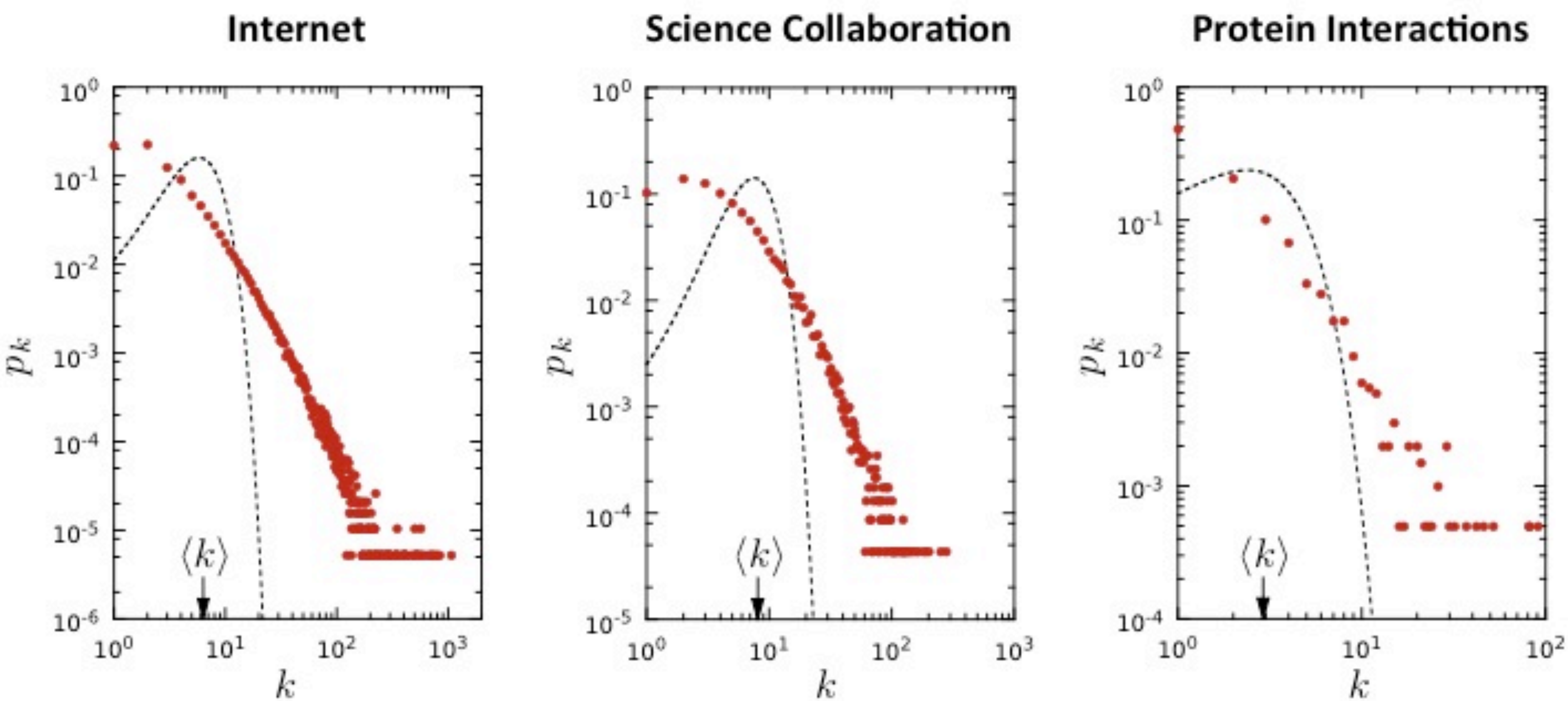
$$\binom{N-1}{k} \approx \frac{(N-1)^k}{k!} \qquad\qquad p_k = e^{-\langle k \rangle}\frac{\langle k \rangle^k}{k!}$$

Degree distribution

- Distribution is Poisson if *N* is large enough



Are real networks = Poisson?



Are real networks = Poisson?

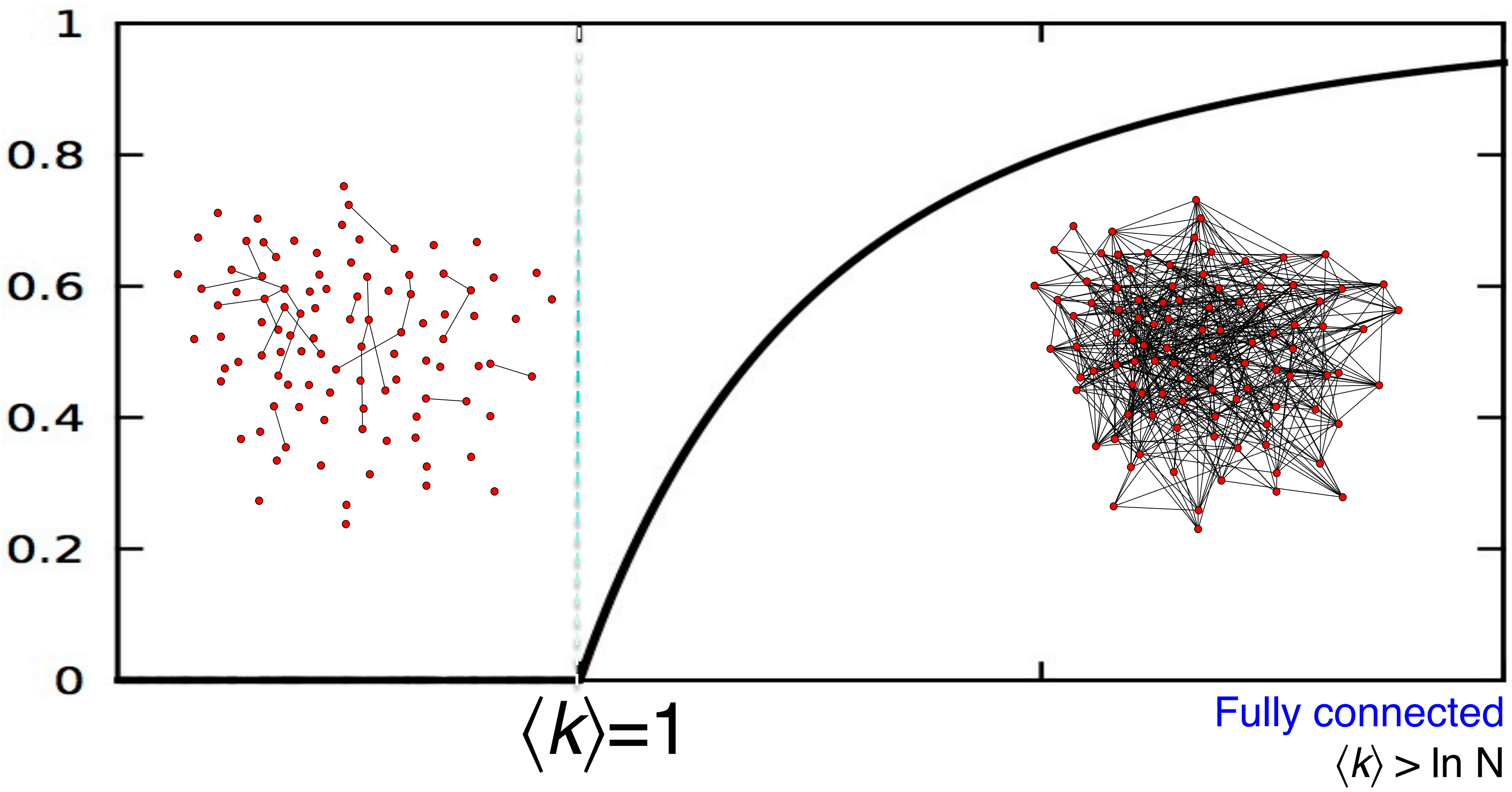
- A Poisson distribution is unlikely to have large values: so there are few nodes with high degree (called **hubs**)
  - this is in sharp contrast with experience: social networks often have hubs
- Random networks are generally deprived of hubs, which are common in reality

disconnected → network

- Consider a dynamic creation of a random network, i.e. links are added in sequence
  - it can be implemented by slowly raising p
  - adding links = transition from disconnected scenario to a significant huge component
  - Justification: for continuity, initially $\langle k \rangle = 0$ implying there is no network (only small disconnected components); but at the end $\langle k \rangle = N - 1$ and we have a complete graph

disconnected → network

- When does this transition happen?



disconnected → network

- At $\langle k \rangle = 1$ a giant component (GC) appears
- Why does the transition happen there?
  - call u the fraction of nodes $\notin$ GC
- Take a generic node i and see if it can reach the GC via another node j. No way if:
  - i and j are not connected (probability $1-p$)
  - or they are but j itself $\notin$ GC (probability pu)
- Result: $\text{Prob}[i \notin \text{GC}] = (1-p + pu)^{N-1}$
  but also: $\text{Prob}[i \notin \text{GC}] = u$

full connection

- When does the GC = the whole network?
  - a node is isolated from the GC with probability $(1-p)^{N_G} \simeq (1-p)^N$ if $|GC| \simeq N$
  - thus, we have $N(1-p)^N \simeq N e^{-Np}$ such nodes

- Switching point when $N e^{-Np} = 1$, i.e.:

$$p = \ln N / N$$

$$\langle k \rangle = \ln N$$

Small world

- A popular catchphrase of network science also known as "**six degrees of separation**"
  - we are more connected than what we think
- In a random network, distance between two randomly chosen nodes is generally **short**
- What does this mean?
  - Milgram experiment: try reaching an unknown individual on Earth;
  - estimate was 100 hops; on average, it took 6

Small world

- It looks surprising just because we are used to regular lattices
  - the “6 degrees” may even be overestimated
  - now we have social online networks and recent estimates tell that $\langle d \rangle = 4.75$ hops
  - so: level 1 of direct acquaintances, level 2 of friends of friends, and all other people in the world you can easily reach

Are real networks random?

feature	random networks	real networks	
degree distribution	binomial → Poisson, with no hubs	heavy tailed with some hubs	✗
connectedness (1)	if supercritical ($\langle k \rangle \geq 1$), we observe a GC	they have $\langle k \rangle \geq 1$ and usually a GC	✓
connectedness (2)	we need $\langle k \rangle \geq \ln N$ to have full connectivity	$\langle k \rangle \ll \ln N$ but we have full connectivity	✗
average path length	small, actually scales as $\ln N / \ln \langle k \rangle$	correct at least as order of magnitude	✓
clustering coefficient	independent of k_j decreases with N	increases with k_j independent of N	✗

Beyond Poisson

- Exponential distribution: can go to very big values but with vanishing probability
  - connected with memoryless generation

- Or, a "fat-tailed" distribution: the presence of nodes with high degree is significant
  - e.g. **power law**: $p_k \sim k^{-\gamma}$ with $\gamma > 1$ called the **exponent** of the power law
  - just proportional to and within a given range

Power laws

- Exact distribution requires normalization:

  $p_k = C_0\, k^{-\gamma}$ for $k \geq k_{min}$

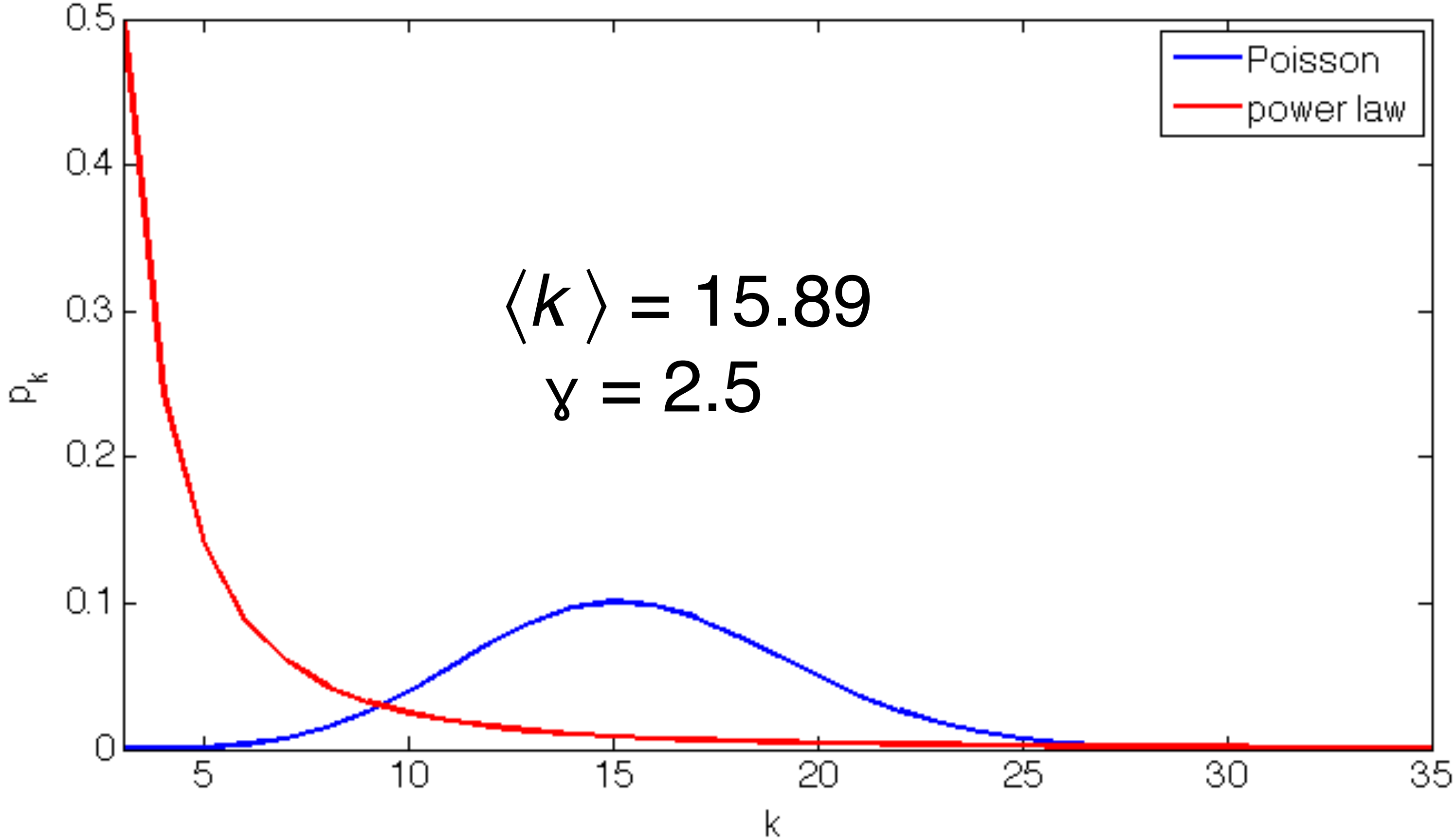
  then $C_0 = (\gamma - 1)\, k_{min}^{\gamma-1}$

  - note: this is for a continuous pdf
  - for discrete values, just more involuted

  - Also, cannot hold for any k (e.g. it goes to infinity for $k \rightarrow 0$) but only within $[k_{min}, k_{max}]$

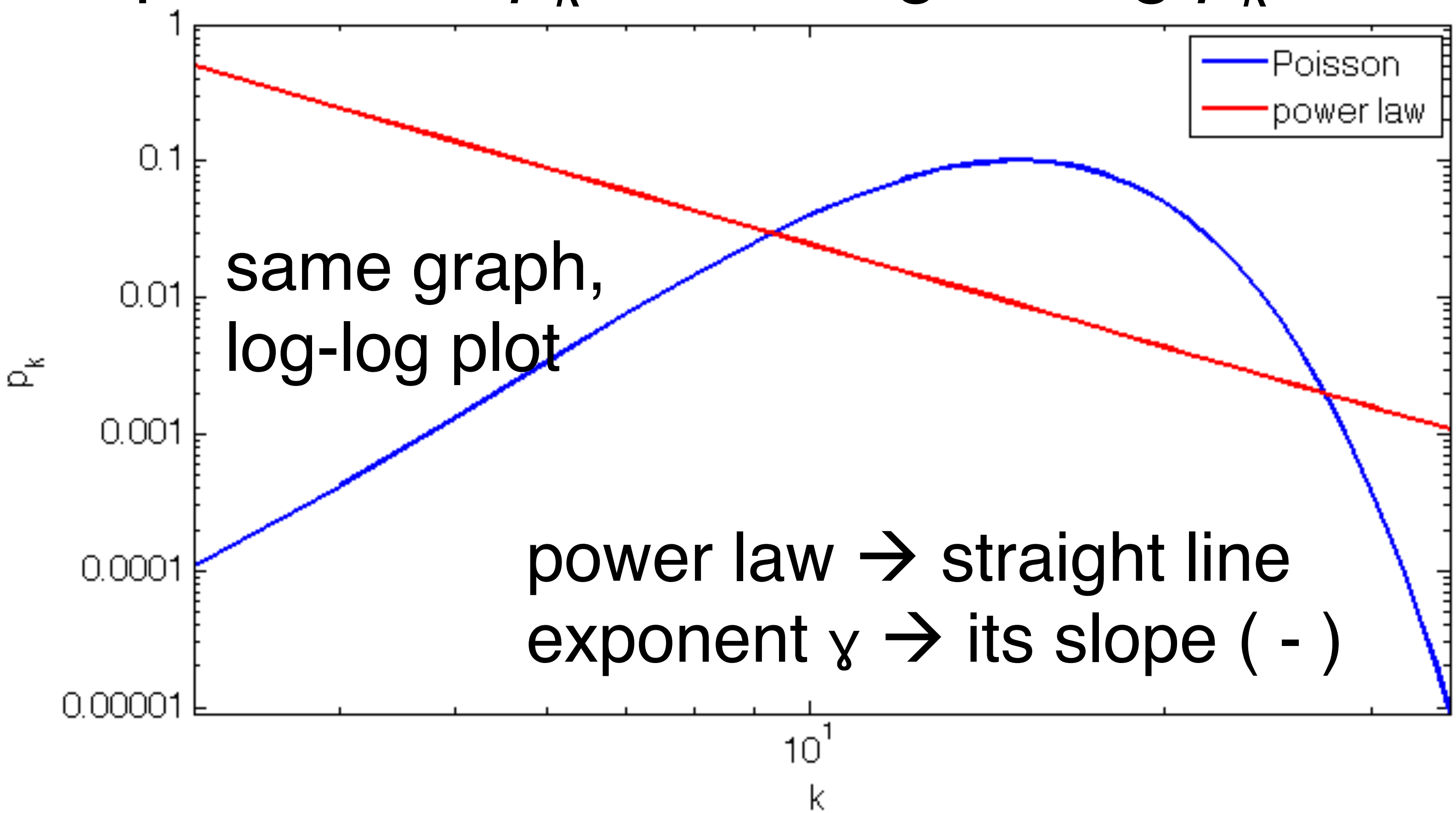
Power laws

- A power law distribution can have the same average value $\langle k \rangle$ than a Poisson



Power laws

- If we apply logarithms to both sides of the power law $p_k \sim k^{-\gamma}$ we get: $\log p_k \sim -\gamma \log k$



Hubs

- Why do we need to introduce that?
  - mostly since we want to account for **hubs**
  - random graphs lack hubs → uniformly connected, nodes have similar degrees

- Yet, many real world networks have nodes that are more connected than others
  - remember Pareto 80/20 rule, a well known empirical rule of many social sciences

Hubs

- As network size N is finite, distribution p_k is meaningless beyond some value k_{max}
- Actually, two different upper limits
  - for mathematical reasons (e.g., $0 \leq prob \leq 1$), $p_k \sim k^{-\gamma}$ only holds within range $[k_{min}, k_{max}]$
  - or we can find the practical cutoff of the distribution meant as the k_{max} s.t. we have vanishing probability of degree $k > k_{max}$

Hubs

- **Power-law**: $p_k = C_0\, k^{-\gamma}$ with $C_0 = (\gamma - 1)\, k_{min}^{\gamma-1}$
- Then condition $\int_{k_{max}}^{\infty} p_k \, dk = \frac{1}{N}$ translates to

$$(k_{min} / k_{max})^{\gamma-1} = N \rightarrow k_{max} = k_{min} \; N^{1/(\gamma-1)}$$

- Now the highest degree increases in N polynomially fast (albeit sublinearly)
  - it means that big hubs are present for big N

Scale-free networks

- Networks whose degrees follow a power law distribution are often called **scale-free**
- Why this name? Compute moments:
  - first moment (average): $\langle k \rangle = \int_{k_{\min}}^{\infty} k \cdot p_k \, \mathrm{d}k$
  - second moment: $\langle k^2 \rangle = \int_{k_{\min}}^{\infty} k^2 \cdot p_k \, \mathrm{d}k$

    that gives the variance as $\langle k^2 \rangle - \langle k \rangle^2$

Scale-free networks

- In general, the nth moment is

$$\langle k^n \rangle = \int_{k_{\min}}^{\infty} k^n \cdot p_k \, \mathrm{d}k = \int_{k_{\min}}^{\infty} C k^{n-\gamma} \, \mathrm{d}k$$

  and this integral converges only if $\gamma-1>n$

- This means that if $2 < \gamma < 3$ only the first moment is finite: the variance is infinite
  - hence the name “scale-free”, implying no inner structure in the degrees
  - random choices can pick very big hubs

Scale-free networks

- As a matter of fact, p_k only valid in $[k_{\min}, k_{\max}]$
  - correct, as network size N must be finite
  - and the variance cannot be > $k_{\max}{}^2$
- The nth moment in reality is

$$\langle k^n \rangle = \int_{k_{\min}}^{k_{\max}} k^n \cdot p_k \,\mathrm{d}k = C \frac{k_{\max}^{n-\gamma+1} - k_{\min}^{n-\gamma+1}}{n-\gamma+1}$$

- Still the nth moment is big for large networks as $k_{\max}$ increases with N
  - but is not infinite (just very big)

There are hubs nearby

- On a scale-free network, it is easier to find a **shortest path** towards a hub
  - because a hub is (by definition) better connected than other nodes
  - also the reason why the often quoted "six degrees" are probably fewer
- Note: in many social experiments, people avoided hubs (for entirely perceptual reasons, e.g., assuming they are busy)

Distances on scale-free

- For $\gamma<3$ we have an **ultra-small world:**
  - average distance $\langle d \rangle \uparrow$ but slower than ln N
  - very different from a random graph, where all nodes have similar degrees, thus most paths will have comparable length
  - here, most of the paths go through the few high degree hubs, reducing the distances
- From the quantitative point of view, we observe a stronger "small world" property

Distances on scale-free

- For $\gamma>3$ the network is scale-free but:
  - $\langle d \rangle$ increases as $\ln N$ (like random graphs)
  - $\langle k^2 \rangle$ is finite
  - we observe the same small world behavior that we identified for random graphs
- From the quantitative point of view, this kind of network is similar in many ways to a random graph

Connections to epidemics

- We can see a **network of infections**
  - who got infected by whom
  - note: this is actually a directed network, but many considerations still hold

- Then we have the following analogies:
  - Erdős-Rényi = homogeneous mixing
  - $\langle k \rangle = \mathbb{E}[\#\text{infecteds}]$ = coefficient R_0
  - and GC = epidemics over entire network

Connections to epidemics

- We also have **networks of contacts**
  - pre-existing structure on which the epidemics spread: topology is important
  - and also likely not random nor memoryless
- We have further analogies:
  - degree = risk structure
  - hubs = super-spreaders
  - and maybe different conditions for the disease to spread or for contrasting it