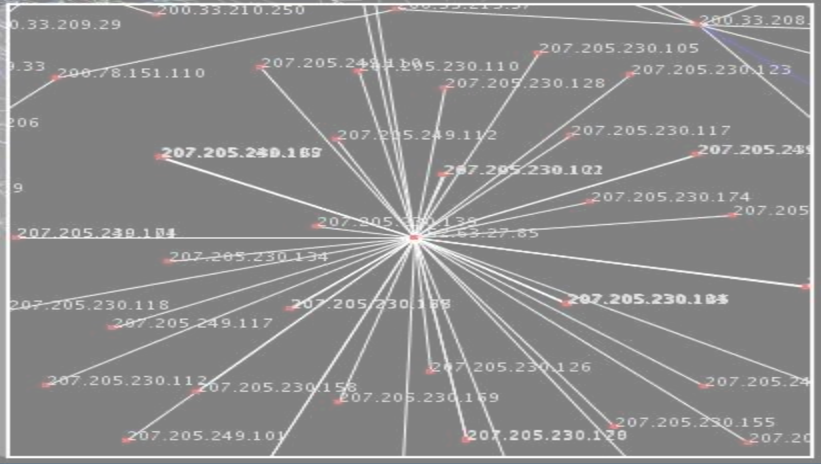
Topics in Algorithms and Data Science

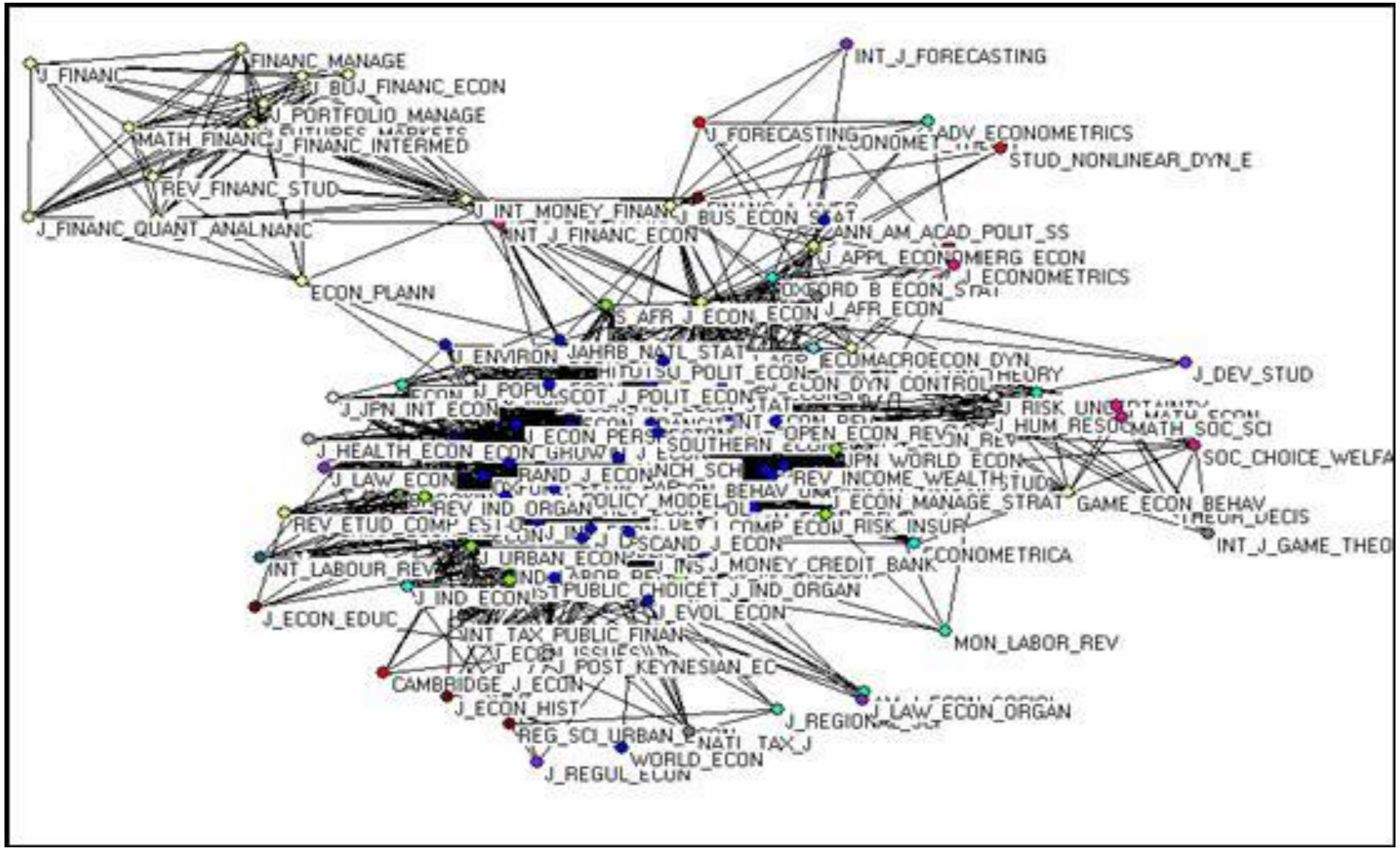
Random Graphs

Omid Etesami



Large graphs

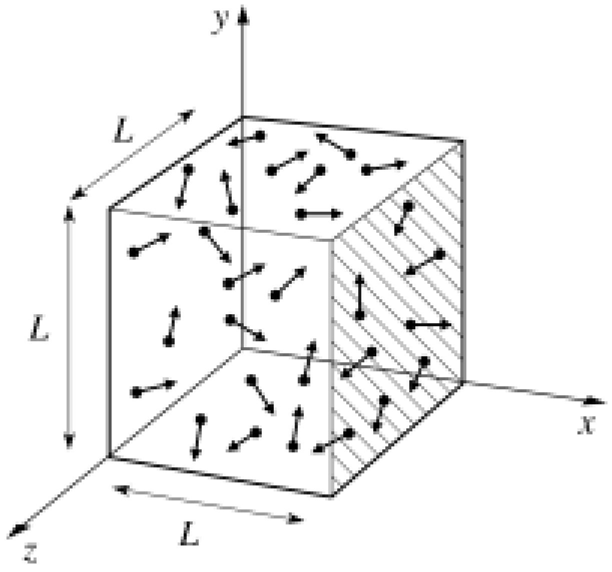
- World Wide Web
- Internet
- Social Networks
- Journal Citations
- …



Economics Journals Citations

Random graphs

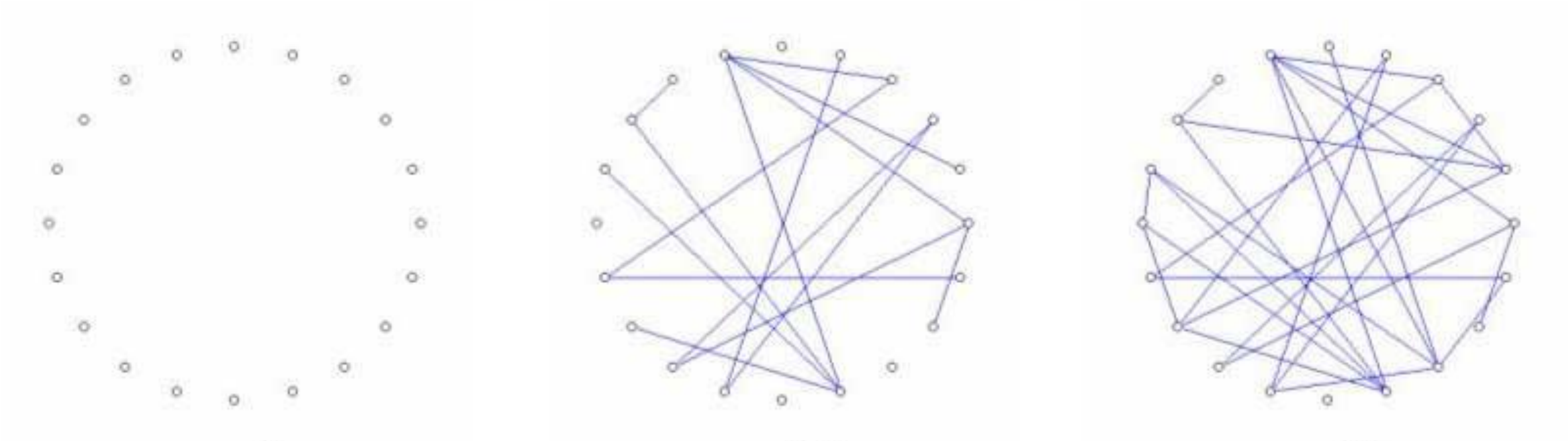
- Unlike traditional graph theory, we are interested in statistical properties of large graphs
- Similar to the shift in physics in late 19th century from mechanics to statistical mechanics



$G(n,p)$ graphs

Erdos-Renyi graphs

- *G(n, p)* random graph with *n* vertices
- Each edge appears with probability *p* independently of other edges

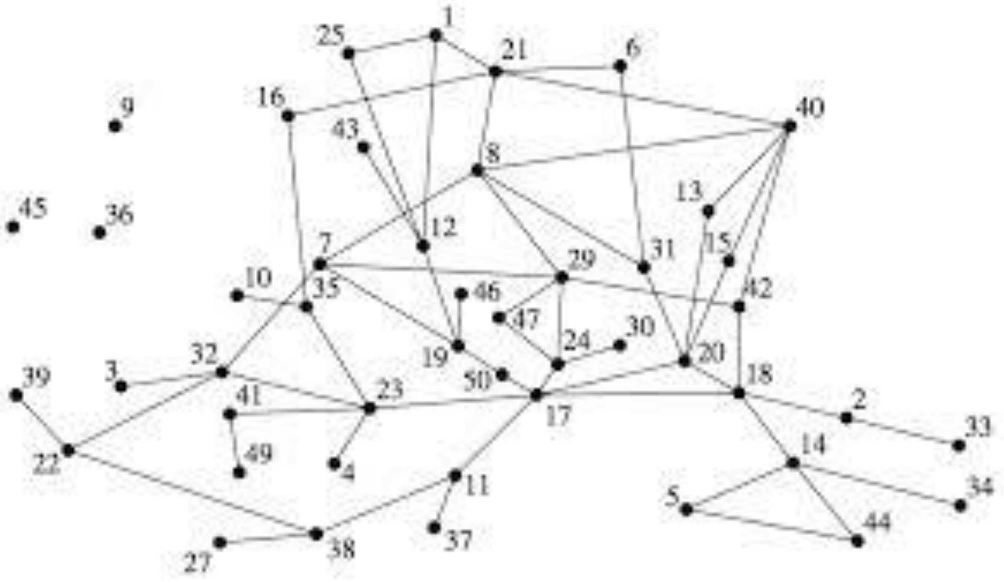


$p = 0$ (a) $p = 0.1$ (b) $p = 0.2$ (c)

Erdos-Renyi graphs with constant expected degree

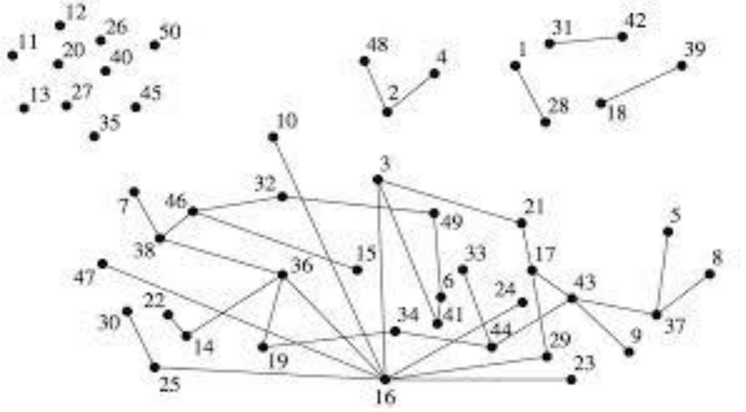
- The probability *p* may depend on *n.*

- If $p = d/n$, the expected degree is $(n-1)d/n \approx d$.

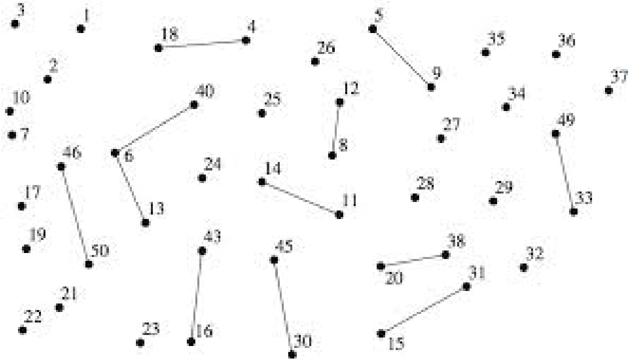


Global property emerges from independent choices

With no “collusion”, the following happens:



d > 1: with probability almost 1, there is a giant component of size *Ω(n)*



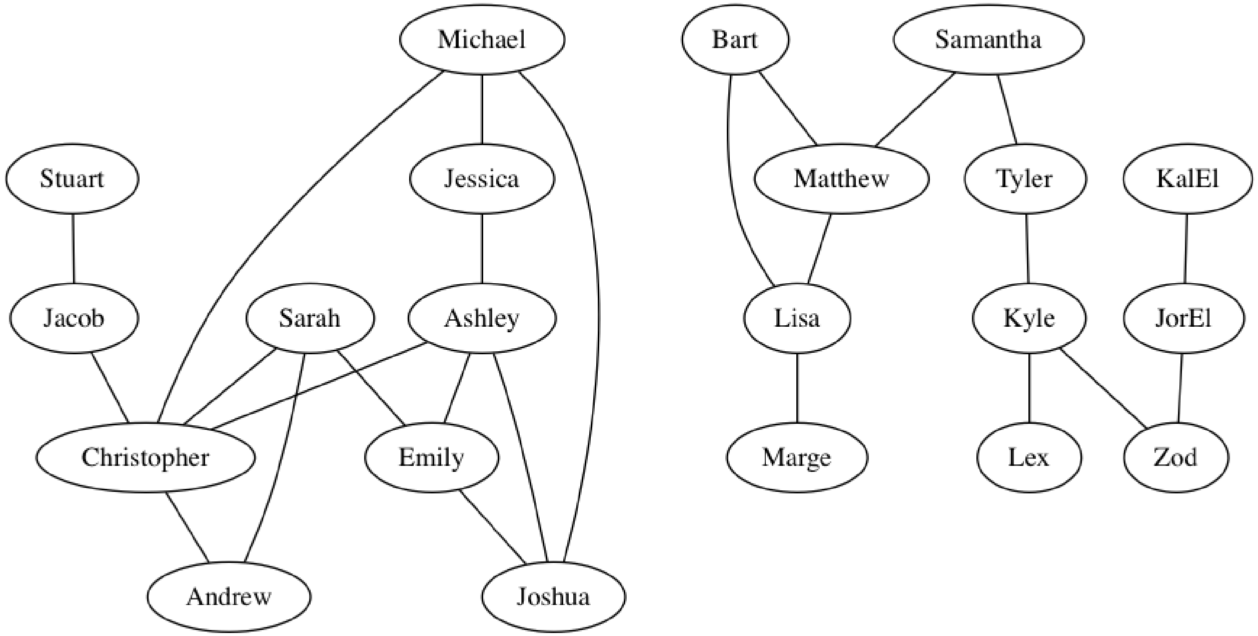
d < 1: with probability almost 1, each connected component is of size *o(n)*

Friendship graph

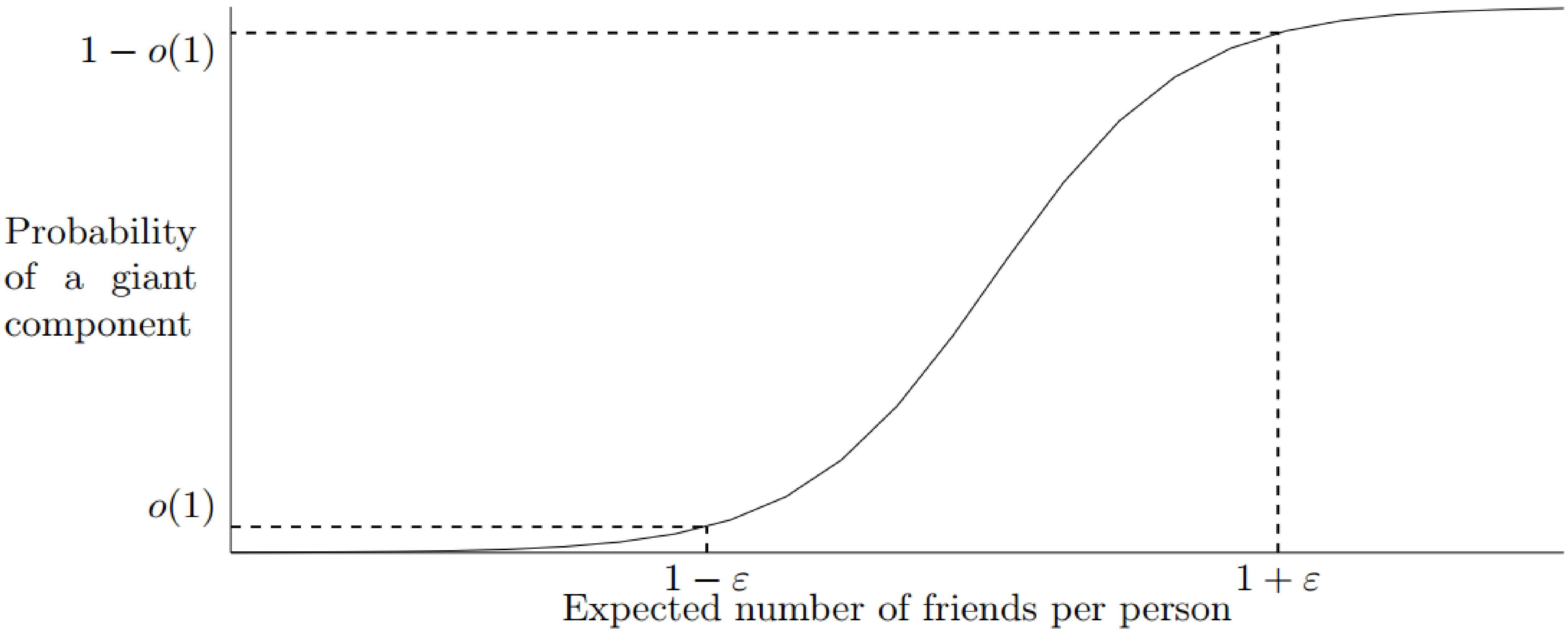
- vertices = people,   edges = knowing each other
- two persons in the same connected component if they indirectly know each other

- each pair of persons become friends with probability *p*

- average degree = expected # friends



Existence of giant component



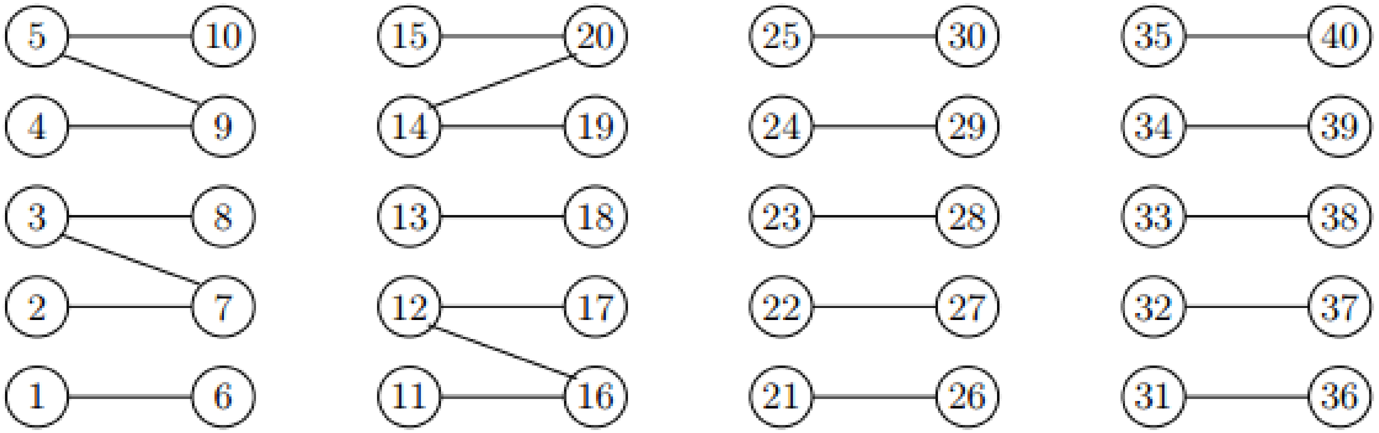
Random vs not random

The bottom graph looks more random.

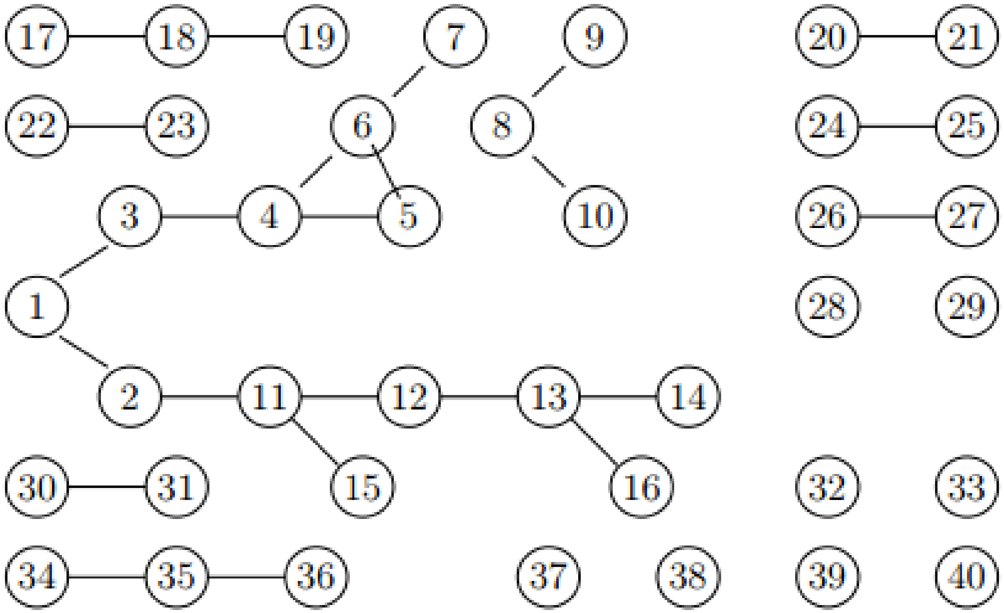
average degree > 1

so we expect a giant component.

Small components are mostly trees.



A graph with 40 vertices and 24 edges



A randomly generated $G(n, p)$ graph with 40 vertices and 24 edges

Degree distribution

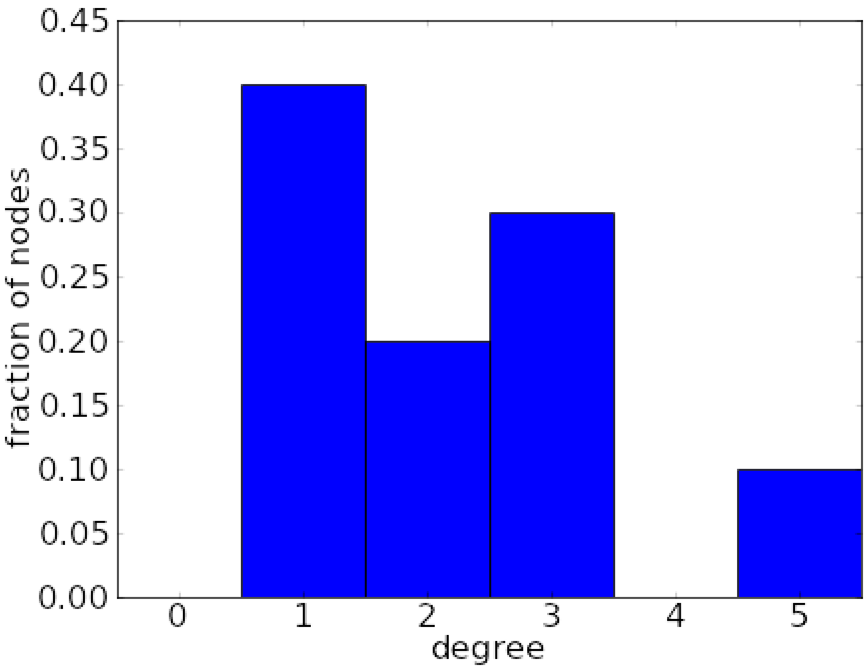
Degree distribution

is the number of vertices of each given degree.

Easy to calculate in real-world graphs.

In *G(n,p):* degree of each vertex is sum of
n-1 independent Bernoulli random variables,
resulting in the binomial distribution.

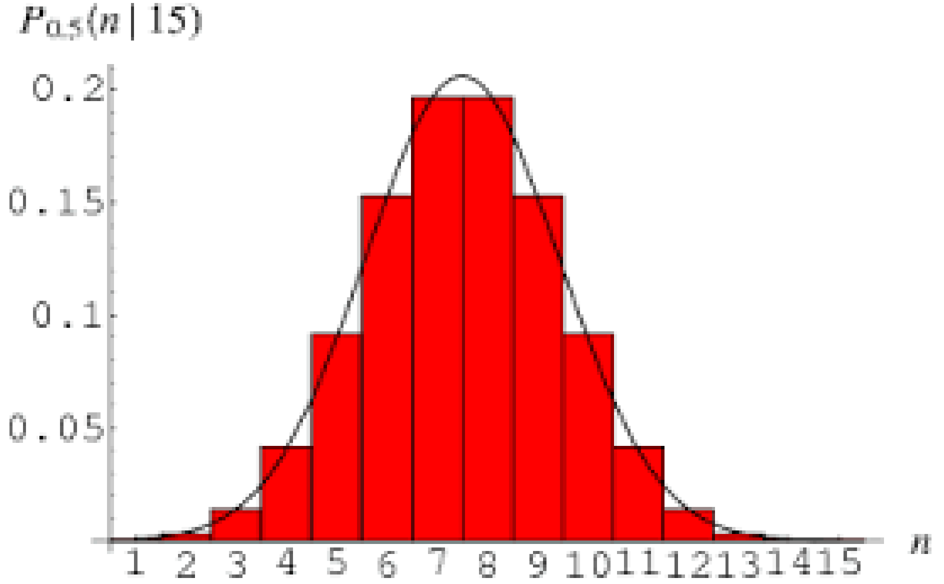
For large *n,* we replace *n-1* with *n.*



Example: *G(n, ½)*

$$\mathrm{Prob}(\deg = k) = \binom{n-1}{k}/2^{n-1} \approx \binom{n}{k}/2^n.$$

- Mean m = n/2 (sum of Bernoulli expected values)
- Variance σ^2 = n/4 (sum of Bernoulli variances)



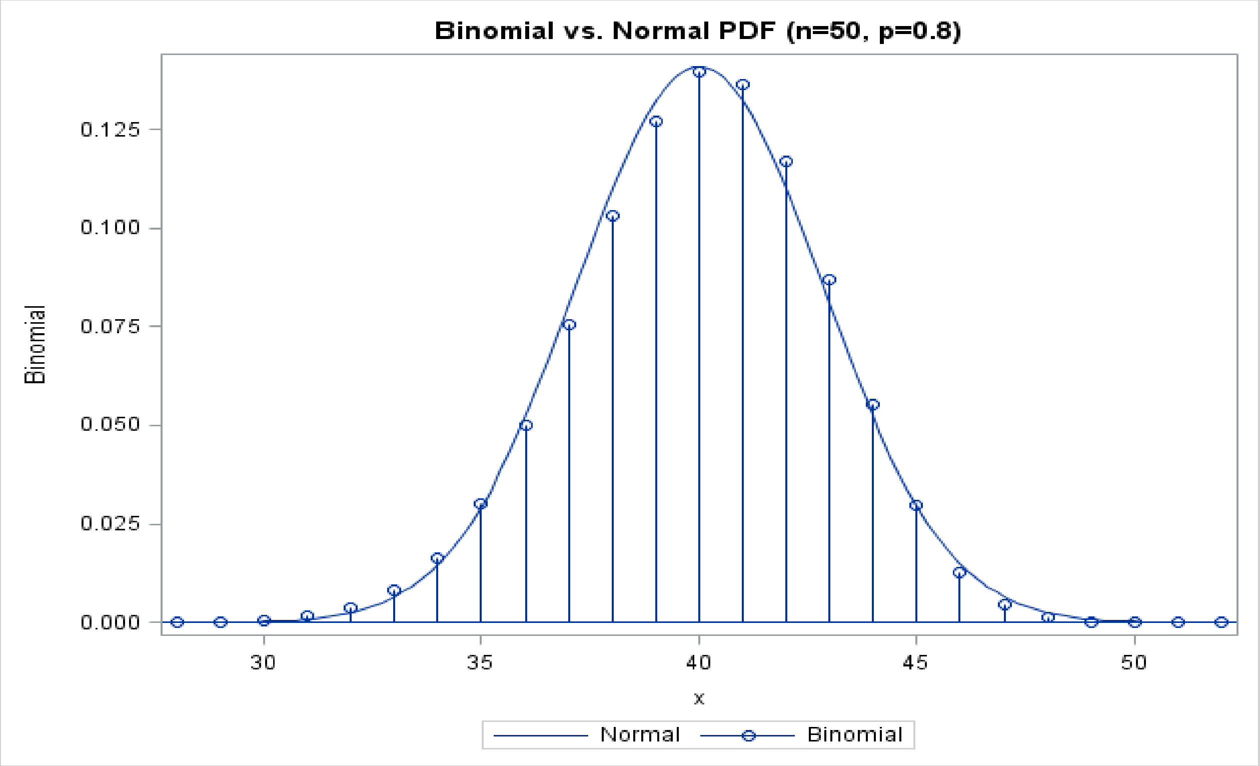
For each Ɛ > 0, almost surely

the degree of each vertex is within 1 ± Ɛ of n/2

$G(n,1/2)$ (continued): normal approximation

binomial distribution ≈ normal distribution of same mean and variance

$$\frac{1}{\sqrt{2\pi\sigma}} e^{\frac{-(k-m)^2}{2\sigma^2}}$$

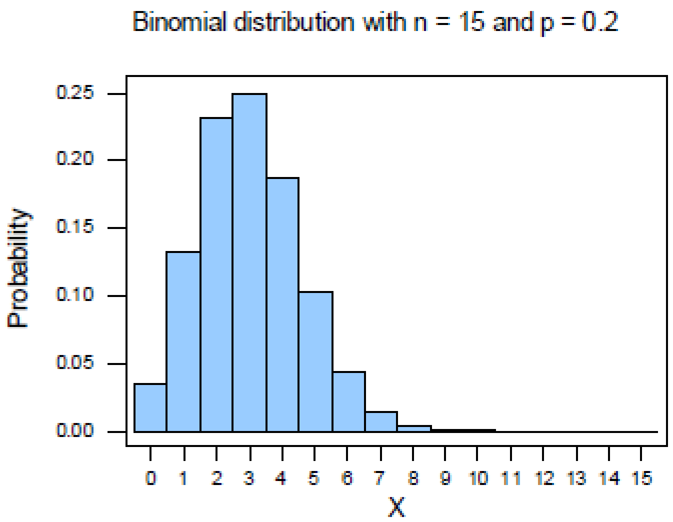


most mass have value mean ± $c\, n^{1/2}$ for constant c.

G(n, p) for general *p*

$$\text{Prob}(\text{degree} = k) = \binom{n-1}{k} p^k (1-p)^{n-k-1} \approx \binom{n}{k} p^k (1-p)^{n-k}$$

The approximation is valid for $k \approx np$.

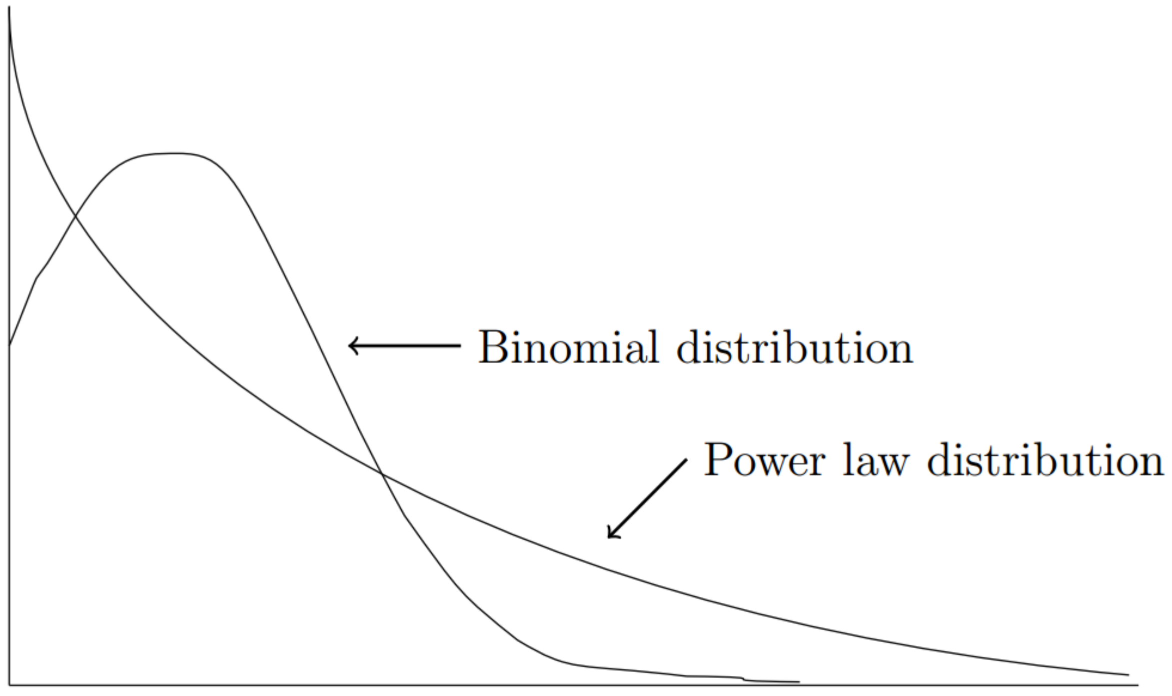


Real-world degree distributions

tail of a random variable = values far from mean (measured in number of standard variations)

- Tail of binomial distribution falls off exponentially fast
- Many graphs in applications have “heavy” tails

Models more complex than *G(n,p)* needed
 for real-world applications



Airline route graph

- Small cities have degree 1 or 2
- Major hubs have degree 100 or more

Power law distribution:
Pr(degree k) = c/k^r.

r often slightly less than 3.

Later in the course,
we see models that
give power law distributions.

Concentration of degree

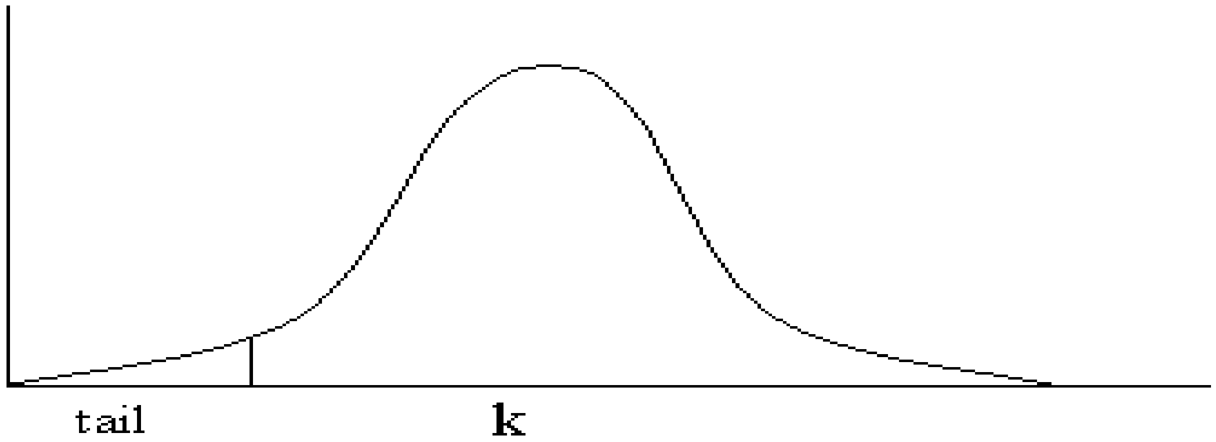
By Chernoff bounds (which we do not prove), for a fixed vertex v,

$$\Pr(|np - \deg(v)| \geq \alpha\sqrt{np}) \leq 2e^{-\alpha^2/3} \text{ when } 0 < \alpha < \sqrt{np}.$$

Let $\epsilon < 1$. Applying union bound to the above,

almost surely the degree of all vertices are in $[np(1-\epsilon), np(1+\epsilon)]$

if $p = \Omega(\ln n/(n\epsilon^2))$.



The lower bound on *p* is necessary:

When *p = 1/n,* vertices of degree Ω(log *n*/log log *n*) exist with high probability.

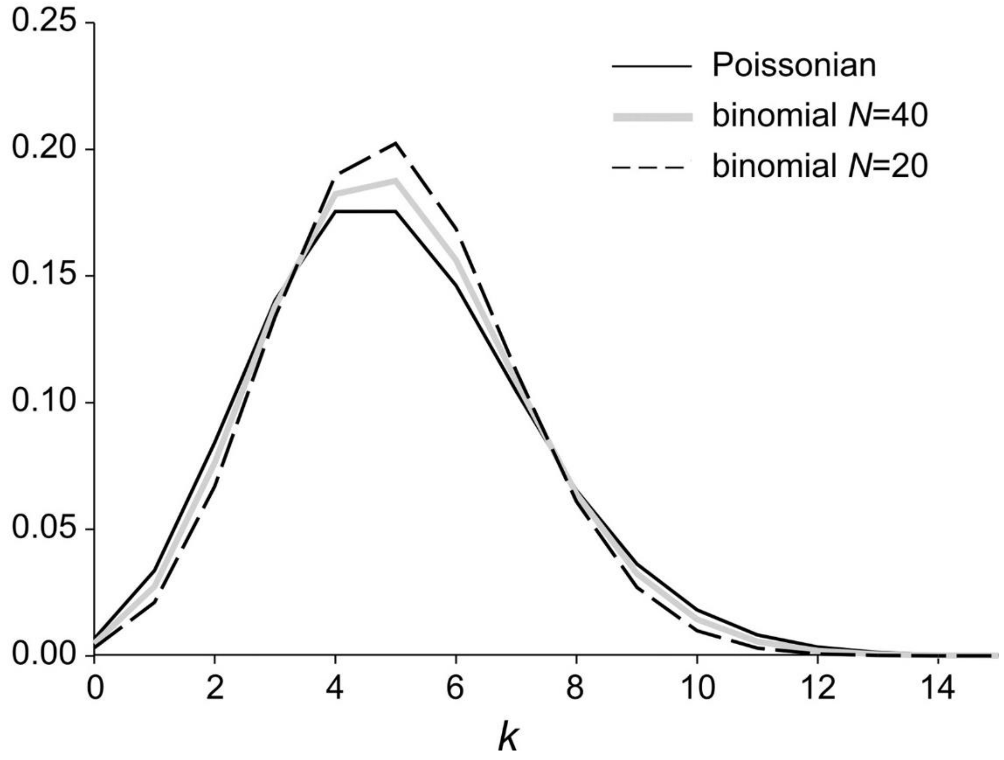
Graphs with constant expected value

When graphs have constant degree, *G(n, p=d/n)* for constant *d* is a better model.

In this case, the binomial distribution approaches the Poisson distribution.

$$\binom{n}{k} p^k (1-p)^{n-k} \approx \frac{n^k}{k!} (\tfrac{d}{n})^k (1-d/n)^{n-k} \approx d^k e^{-d}/k!$$

For $\binom{n}{k} \approx n^k/k!$ we need $k = o(\sqrt{n})$.

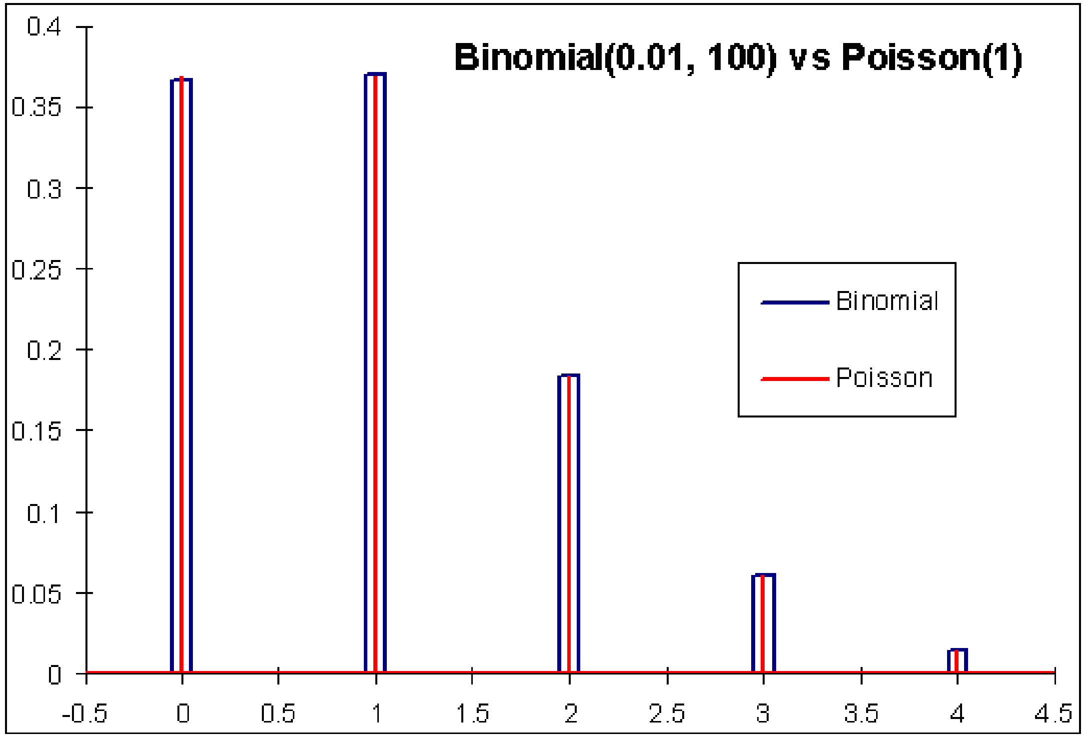


A vertex of high degree

When $p = 1/n$, we have

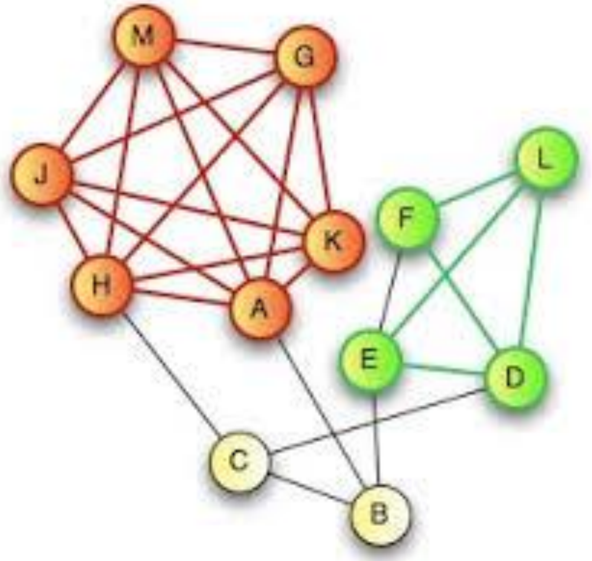
$$\Pr(k) = \binom{n}{k}(1/n)^k(1-1/n)^{n-k} \approx e^{-1}/k! \geq e^{-1}/k^k.$$

If $k = \ln n/\ln\ln n$,
we have $\Pr(k) \geq 1/(en)$.
(Without giving the proof)
with high probability
a vertex of degree k exists
(even though the degrees of
different vertices are not independent).



Today's open problem: finding max clique in *G(n, ½)*

- Almost surely *G(n, ½)* has a max clique of size ≈ $2 \lg_2 n$.



- Can you find it in polynomial time?

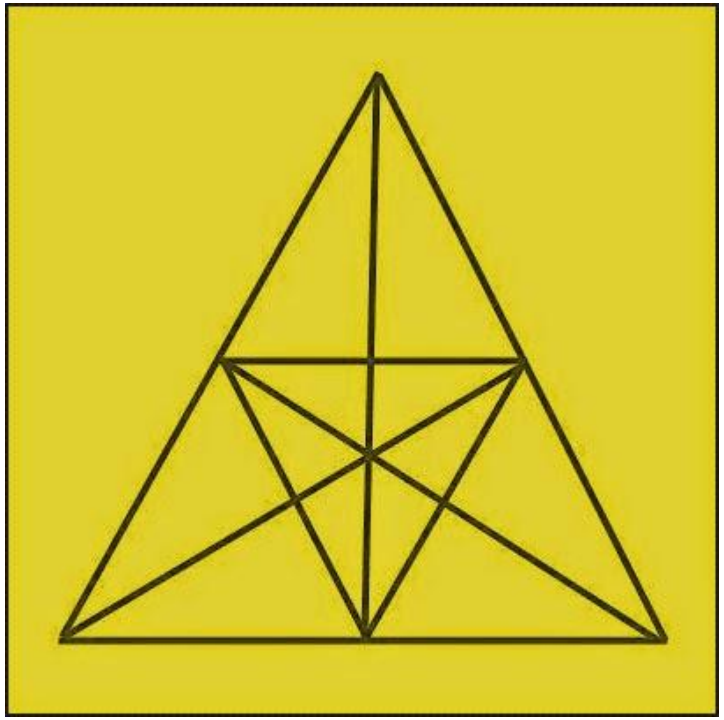
- Best current algorithm is greedy and
  finds only a clique of size ≈$\lg_2 n$.

- It is open if one can find a clique of size $(1 + \varepsilon) \lg_2 n$ for constant $\varepsilon > 0$.

Existence of triangles

Triangles in *G(n,d/n)*

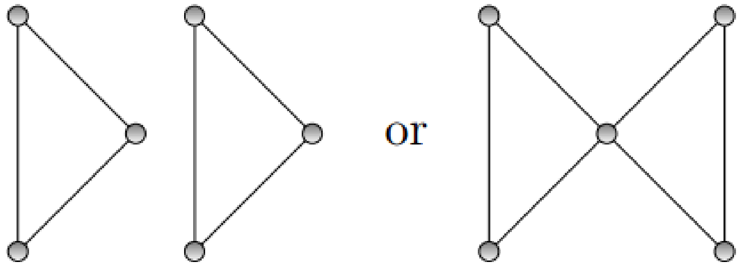
There are $\binom{n}{3}$ potential triangles.
Each is a triangle with probability $(d/n)^3$.
The expected number of triangles is $\binom{n}{3}(d/n)^3 \approx d^3/6$
by indicator random variables and linearity of expectation.



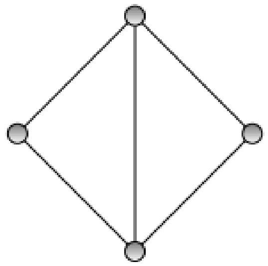
Second moment

To rule out the possibility that all triangles are on a small fraction of graphs, we bound the second moment of # triangles.

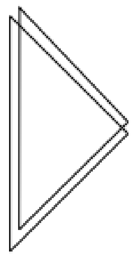
$$X = \sum_{ijk} \Delta_{ijk}.$$
$$E[X^2] = \sum_{ijk,i'j'k'} E[\Delta_{ijk}\Delta_{i'j'k'}].$$



The two triangles of Part 1 are either disjoint or share at most one vertex



The two triangles of Part 2 share an edge

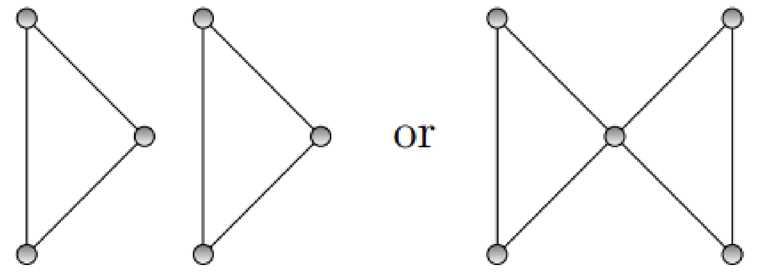


The two triangles in Part 3 are the same triangle

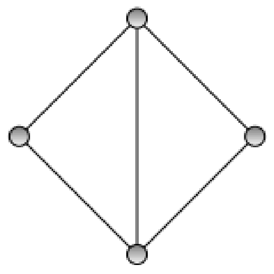
Splitting $E[X^2] = \sum_{ijk,i'j'k'} E[\Delta_{ijk}\Delta_{i'j'k'}]$. into three parts

- For Part 1, $E[\Delta_{i\,j\,k}\, \Delta_{i'j'k'}] = E[\Delta_{i\,j\,k}]\, E[\Delta_{i'j'k'}]$. Thus, the sum for Part 1 is at most $E^2[X]$.
- For part 2, the number of terms is $O(n^4)$, each term $(d/n)^5$.
- For part 3, the sum equals E[X].

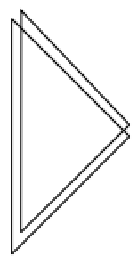
Thus, $\mathrm{Var}[X] = E[X^2] - E^2[X] \le d^3/6 + o(1)$.



The two triangles of Part 1 are either disjoint or share at most one vertex



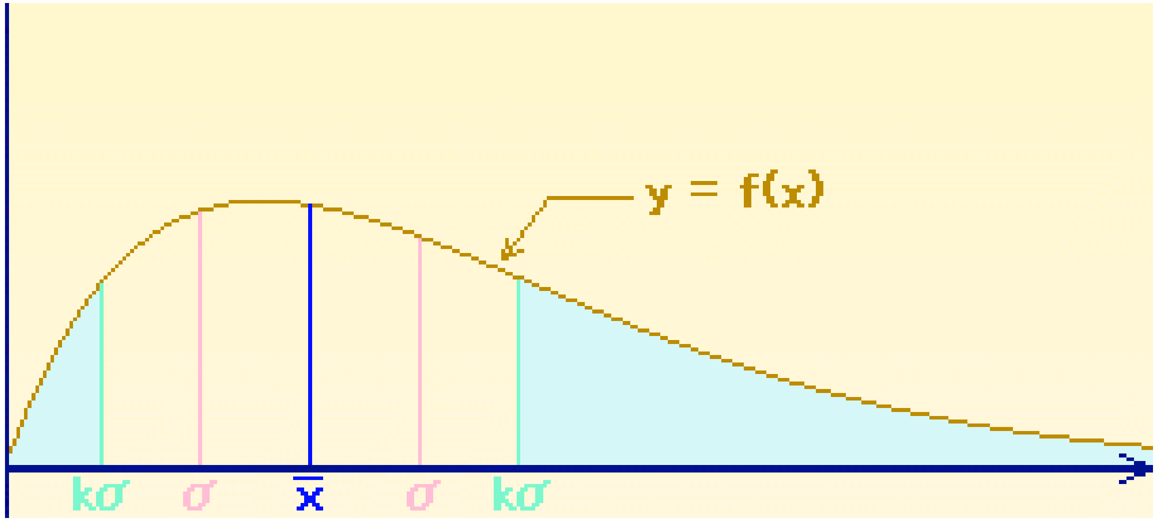
The two triangles of Part 2 share an edge



The two triangles in Part 3 are the same triangle

Chebyshev inequality

$\Pr[X = 0] \leq \Pr[|X - E[X]| \geq E[X]] \leq \mathrm{Var}[X] / E^2[X] \leq 6/d^3 + o(1)$.



When $d > 6^{1/3}$ there exists a triangle with constant nonzero probability.

Phase transitions

Phase transitions in physics

When temperature or pressure slightly increases, abrupt change in the phase of the matter happens,

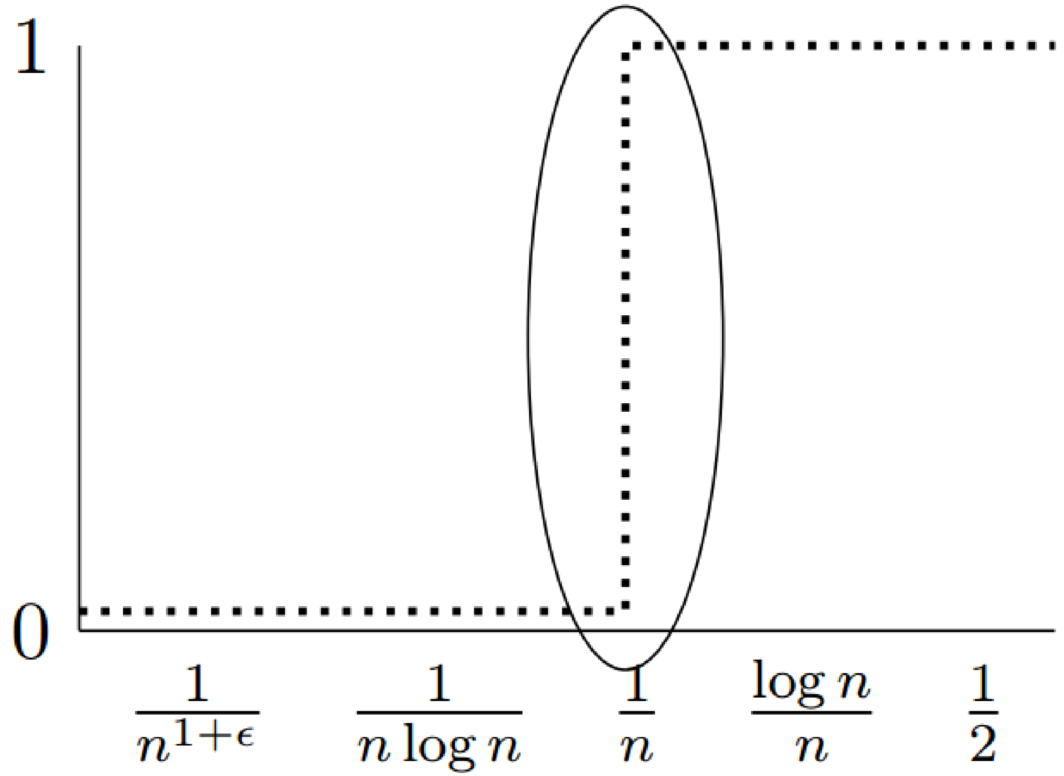
e.g. liquid -> gas.

Phase transition for random graphs

When the edge probability passes some threshold $p(n)$, there is an abrupt transition from not having a property to having that property.

- When $p_1(n) = o(p(n))$, almost surely
  $G(n,p_1)$ does not have the property.
- When $p_2(n) = \omega(p(n))$, almost surely
  $G(n,p_2)$ has the property.

- Example: for appearance of cycles,
  $p(n) = 1/n$.
- Example: for disappearance of isolated vertices,
  $p(n) = \log n / n$.

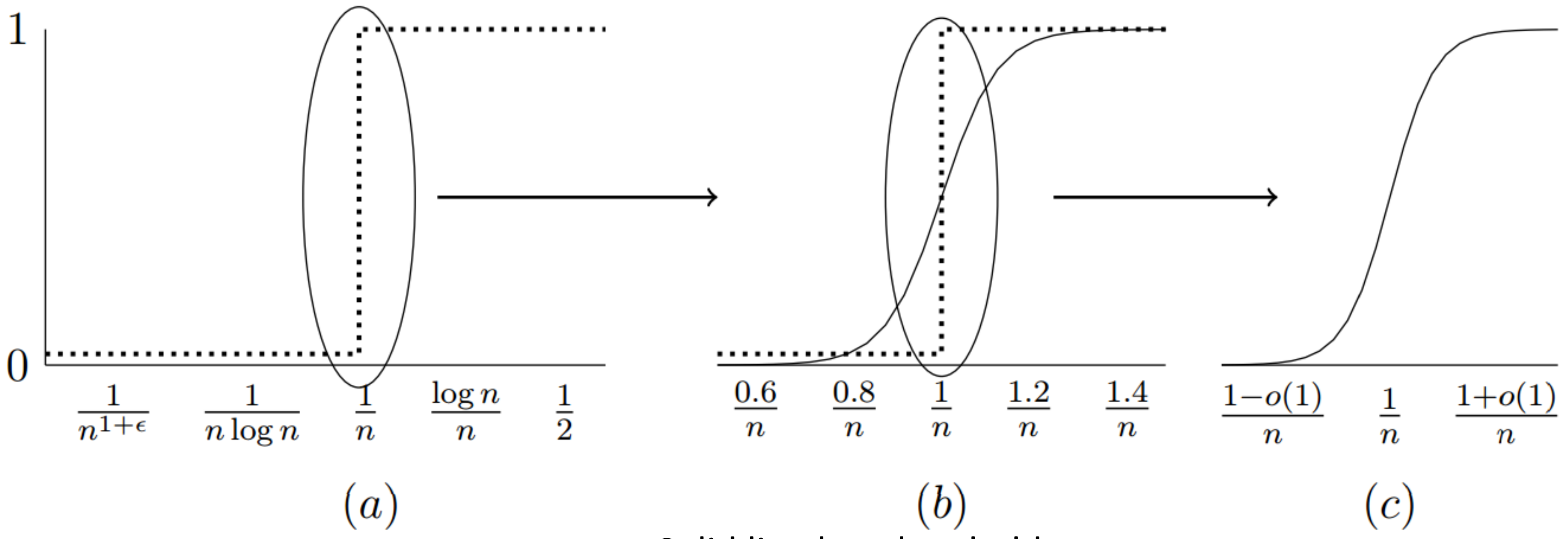


Sharp threshold

p(n) is called a *sharp* threshold if

- when $p_1(n) = p(n)(1\text{-}\Omega(1))$, almost surely $G(n,p_1)$ does not have the property;
- when $p_2(n) = p(n)(1+\Omega(1))$, almost surely $G(n,p_2)$ has the property.

Example: existence of a giant component has sharp threshold at $p(n) = 1/n$.



Dotted line has threshold.

Solid line has threshold;
dotted line has sharp threshold.

Solid line has sharp threshold.

1st and 2nd moment method

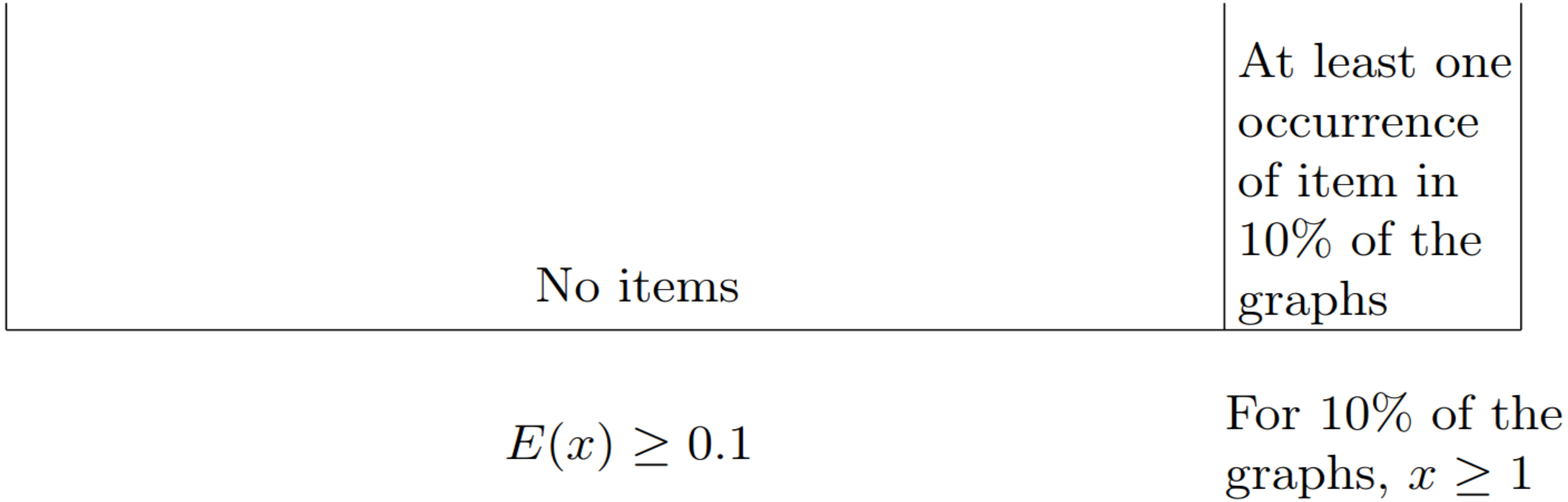
We already know that existence of a triangle has a threshold at $p(n) = 1/n$.

Let X be number of triangles.

Below threshold, $E[X] = o(1)$ so $\Pr[X > 0] = o(1)$ [Markov inequality, 1st moment]

Above threshold, $E[X^2] = E^2[X](1+o(1))$ so $\Pr[X = 0] = o(1)$ [Chebyshev, 2nd moment]

(That $E[X] = \omega(1)$ is not enough for the "above threshold" case.)



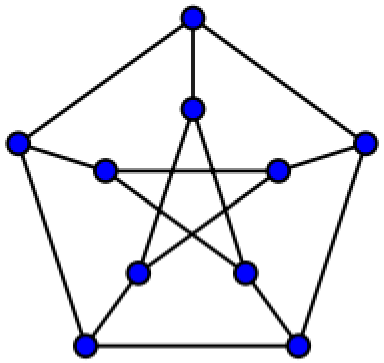
Graph diameter 2

Graph diameter 2 has a sharp threshold at

$$p = \sqrt{2 \ln n / n}$$

- Two vertices have a common neighbor if the size of their neighbors is approximately $n^{1/2}$. (Birthday paradox)

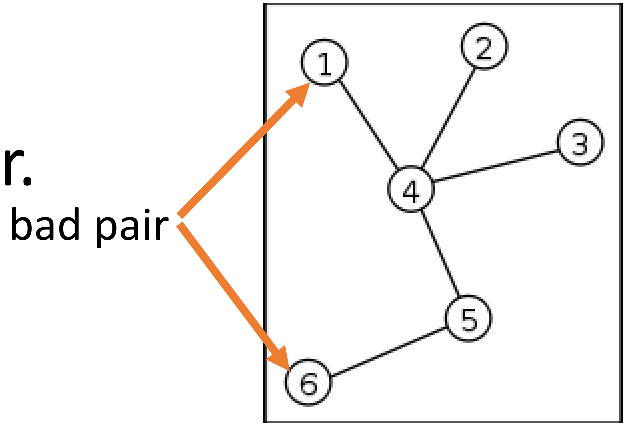
- The extra factor of $(\ln n)^{1/2}$ is to ensure all pairs of vertices have distance at most two.



Petersen has diameter 2

bad pairs

- *(i, j)* bad pair of vertices iff *dist(i,j) > 2.*
- I_{ij} indicator random variable for whether *(i, j)* bad pair.



$$x = \sum_{i<j} I_{ij} = 0 \text{ iff graph has diameter} \leq 2$$

$$\text{For } p = c\sqrt{\ln n / n}, \quad \mathrm{E}[x] = \binom{n}{2}(1-p)(1-p^2)^{n-2} \approx n^{2-c^2}/2.$$

- By first moment method,

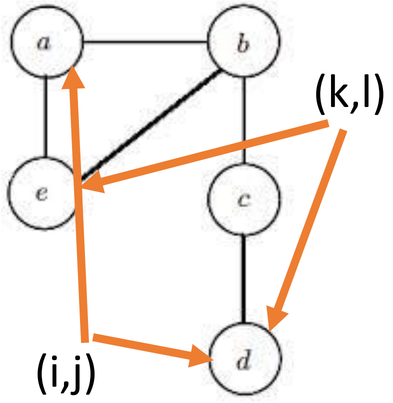
  if c > $2^{1/2}$, almost surely graph has diameter 2.

For c < $2^{1/2}$, we apply the second moment method.

$E[x^2] = \sum_{i<j,\ k<l} E[I_{ij}I_{kl}]$.

Split the sum into three parts according to $|\{i,j,k,l\}| = 2, 3, 4$.

Case 1. i,j,k,l all distinct: $E[I_{ij}I_{kl}] \le (1-p^2)^{2(n-4)} \approx n^{-2c^2}$.
Case 2. i,j,k,l has one repetition: $E[I_{ij}I_{kl}] \le (1-2p^2+p^3)^{n-3} \approx n^{-2c^2}$.
Case 3. $i=k, j=l$: $E[I_{ij}I_{kl}] = E[I_{ij}]$.



$$E[X^2] \le \frac{n^4}{4} n^{-2c^2}(1+o(1)) + O(n^3 n^{-2c^2}) + O(n^2 n^{-c^2})$$
$$= E[X]^2(1+o(1)).$$

Isolated vertices

The disappearance of isolated vertices has a sharp threshold at $p = \ln n / n$

In fact, at this point, the giant component has absorbed all small components of size ≥ 2,

so with the disappearance of isolated vertices, the graph becomes connected.

related to balls and bins

1^{st} and 2^{nd} moment when $p = c \ln n / n$

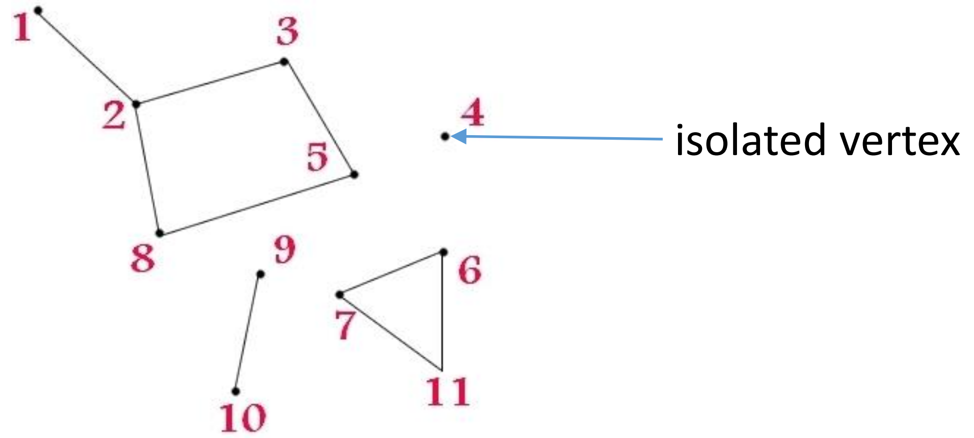
$x = I_1 + \ldots + I_n$, where I_j is indicator random variable for j being isolated.

$$E[x] = n(1-p)^{n-1} \approx n^{1-c}.$$

When $c > 1$, $E[x]$ tends to zero and we can using 1^{st} moment method.

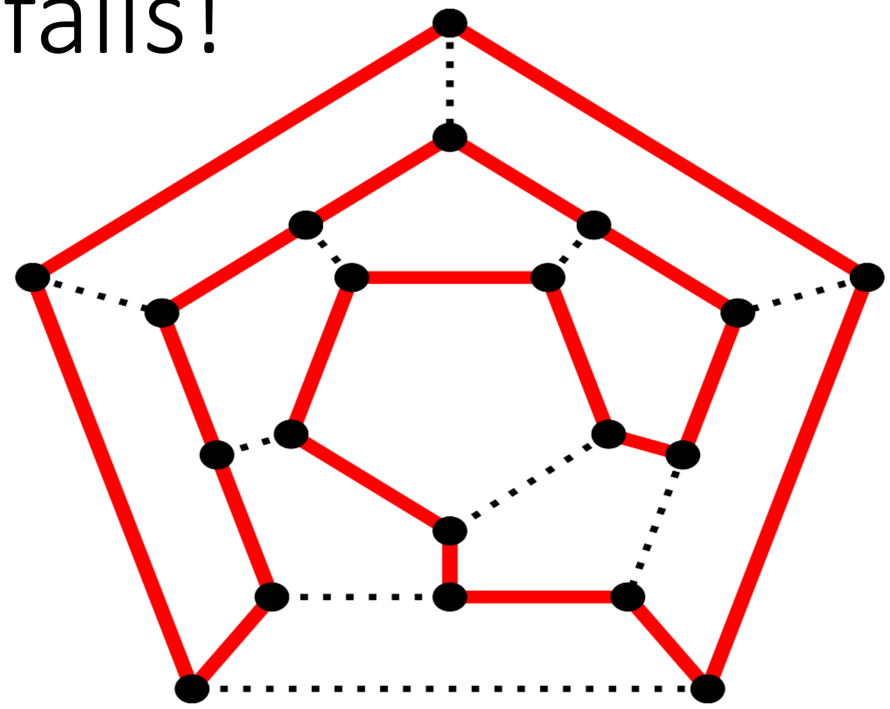
$$E[x^2] = \sum_{i,j} E[I_i I_j] = E[x] + n(n-1)(1-p)^{2(n-1)-1} \leq E[x] + n^{2-2c}.$$

For $c < 1$, an isolated vertex exists almost surely by 2^{nd} moment method.



Hamilton circuits

A situation where 1st moment fails!



Let x = # of Hamilton circuits

The value of *p* for which *E[x]* goes from zero to infinity is not the threshold for having a Hamilton cycle

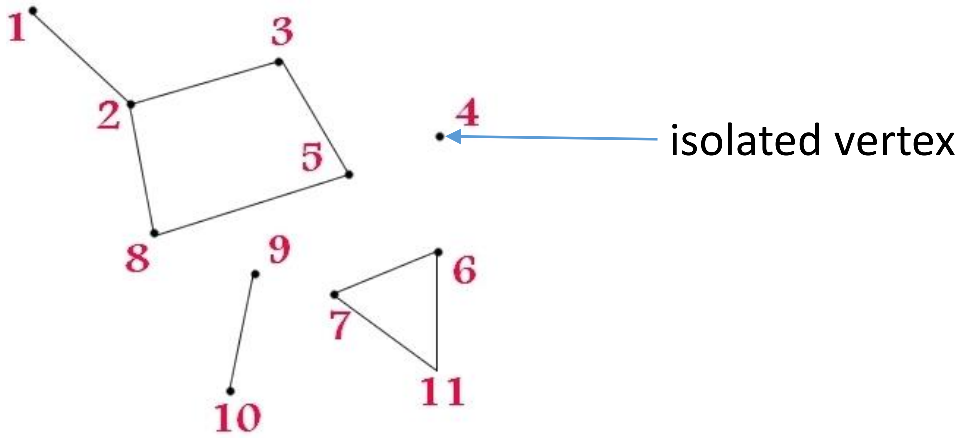
because Hamilton circuits are very concentrated on a small fraction of random graphs.

Expected # Hamilton circuits

For $p = d/n$,

$$E[x] = \frac{(n-1)!}{2}(d/n)^n \approx \Theta(n^{-1/2})(n/e)^n(d/n)^n = \begin{cases} o(1) \text{ if } d < e \\ w(1) \text{ if } d > e \end{cases}$$

but for constant *d,* isolated vertices exist and the graph is not even connected.

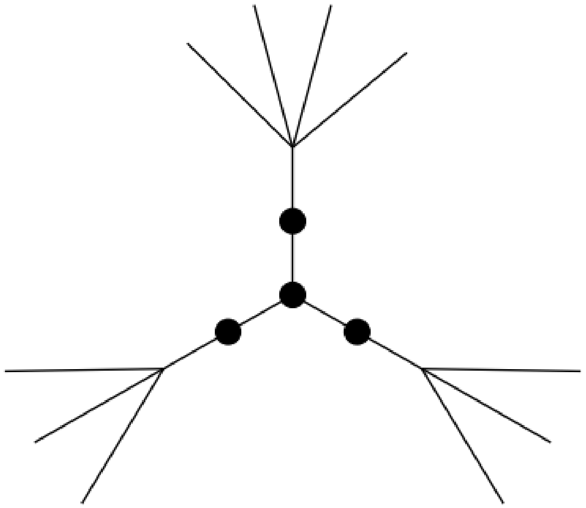


Actual threshold for Hamilton circuits

If $d = \ln n + \ln \ln n + \omega(1)$, almost surely $G(n, d/n)$ is Hamiltonian.
If $d = \ln n + \ln \ln n - \omega(1)$, almost surely $G(n, d/n)$ is not Hamiltonian.

Same threshold as the moment of disappearance of degree-1 vertices!

Why not a subgraph like this
(a degree-3 vertex connected to 3 degree-2 vertices)
happen at that moment?

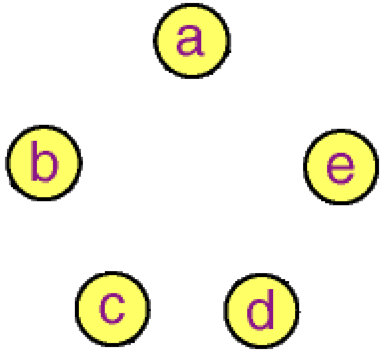


Frequency of degree 2 and 3 vertices is low. The probability that such a configuration of such vertices occur together is low.

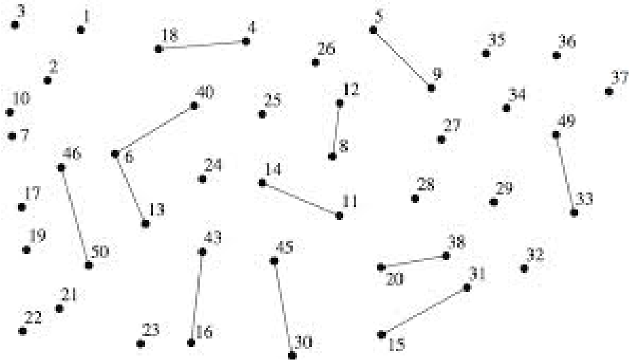
The giant component

The evolution of *G(n,p)* as *p* increases

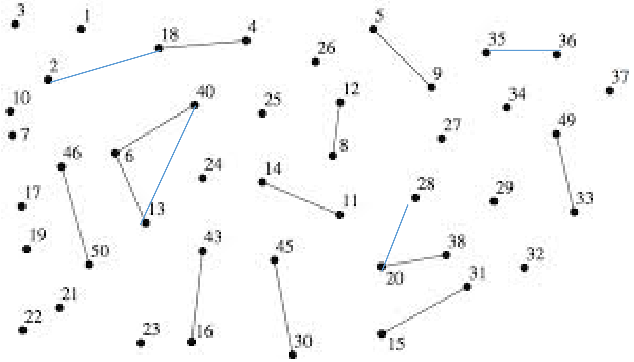
- $p = 0$: no edges



- $p = o(1/n)$: forest, i.e. no cycle



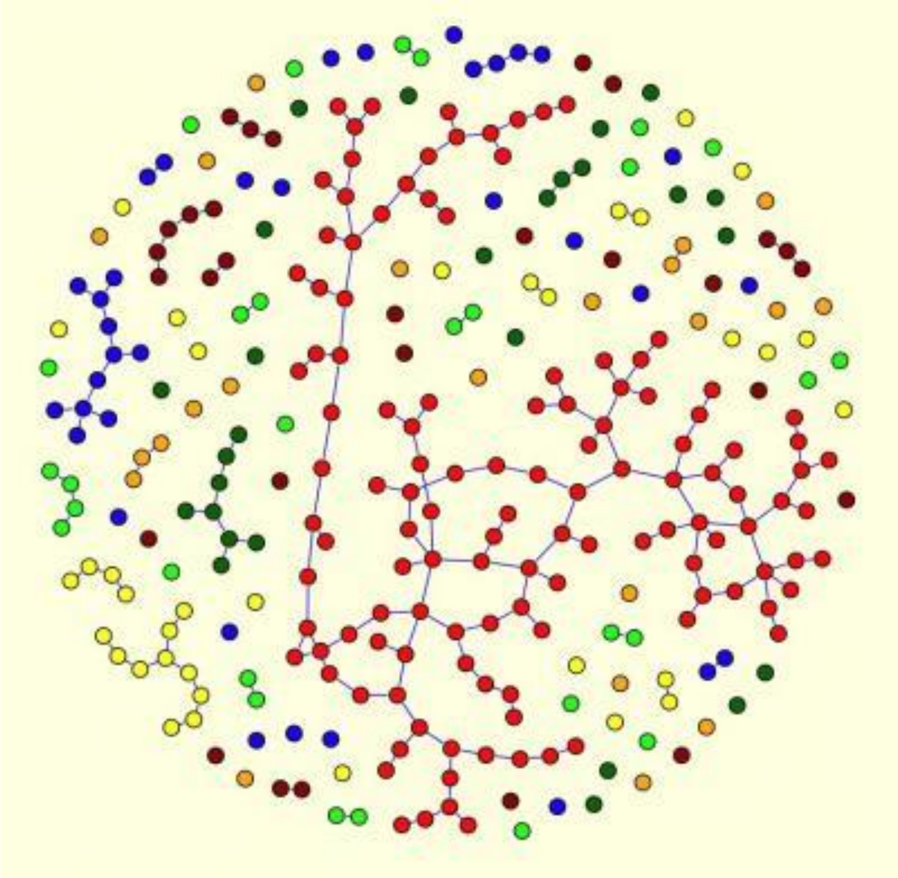
- $p = d/n$, d *constant* < 1:
  all components of size $O(\lg n)$,
  no component has more than one cycle,
  expected # components containing single cycles = O(1),
  there is a cycle with probability Ω(1)



The evolution of $G(n, p)$ as p further increases

- $p = 1/n$: for any function $f = \omega(1)$,
  tree of size $\geq n^{2/3}/f$ exists
  all components have size $\leq n^{2/3}f$

- $p = d/n$, d constant > 1:
  there exists a single giant component
  of size $\Omega(n)$



A giant component happens also in real graphs like portions of the web.

Example: protein interactions

- vertices = proteins,
- edges = proteins interact, i.e. two amino acids bind for an action
- 2735 vertices, 3602 edges: edges/vertices > ½

Size of component	1	2	3	4	5	6	7	8	9	10	11	12	$\cdots$	15	16	$\cdots$	1851
Number of components	48	179	50	25	14	6	4	6	1	1	1	0	0	0	1	0	1

- As more proteins added, the giant component absorbs the smaller components

Further examples of giant component

ftp://ftp.cs.rochester.edu/pub/u/joel/papers.lst

Vertices are papers and edges mean that two papers shared an author.

1	2	3	4	5	6	7	8	14	27488
2712	549	129	51	16	12	8	3	1	1

http://www.gutenberg.org/etext/3202

Vertices represent words and edges connect words that are synonyms of one another.

1	2	3	4	5	14	16	18	48	117	125	128	30242
7	1	1	1	0	1	1	1	1	1	1	1	1

The evolution of $G(n, p)$ as p increases even more

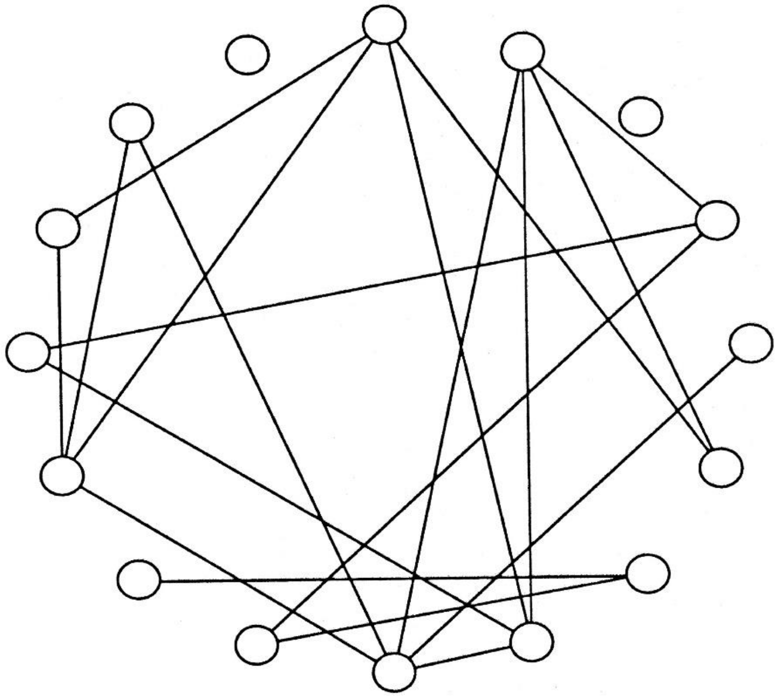
- $p = \ln n / (2n)$:

  all non-isolated vertices are absorbed in the giant component,

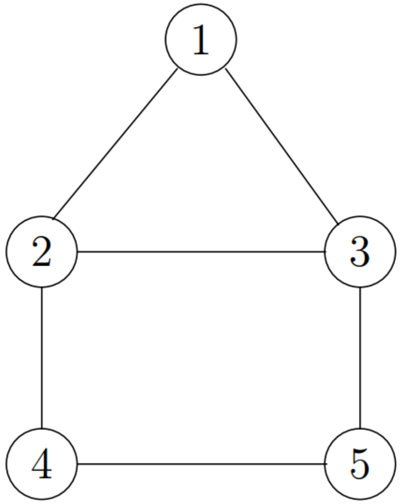
  i.e. graph consists of giant component + isolated vertices

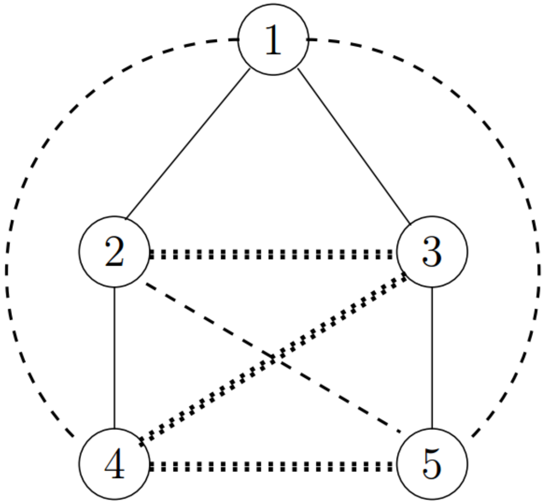
- $p = \ln n / n$: $G(n, p)$ becomes connected

- $p = 1/2$: $G(n, p)$ even has a clique of size $\approx 2 \lg_2 n$



Breadth-first search

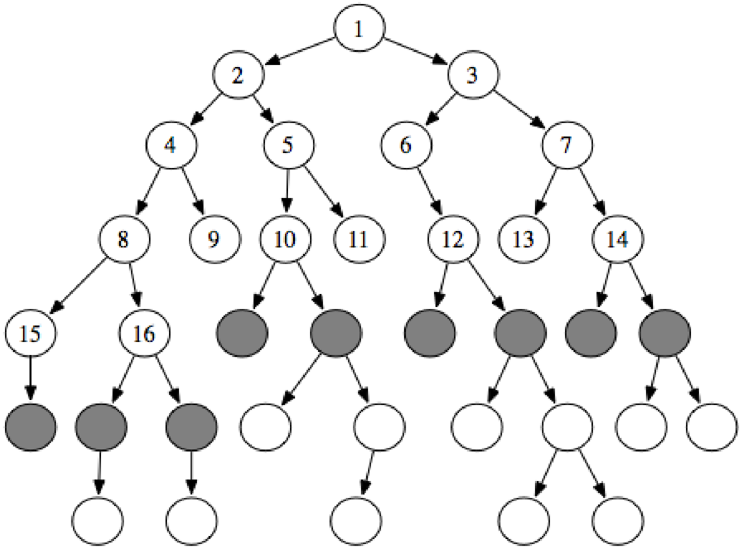




dotted line: unexplored edge
dashed line: edge does not exist
solid line: edge exists

- Generate an edge only when the BFS needs to know if the edge exists
- Start BFS from an arbitrary vertex and mark it discovered and unexplored
- frontier = set of discovered and unexplored vertices
- At each step select v from frontier, and explore it as follows: for each undiscovered vertex u, independently with probability $p = d/n$ add edge (v, u) and add u to the frontier
- BFS finishes when the frontier becomes empty, i.e. when the connected component has been entirely explored

A process equivalent to BFS



- S = {v}, i = 1
- While |S| - i >= 0

    add each vertex in V – S to S independently with probability p=d/n

    i++

If we replace while |S| - i >= 0 with while true,

any vertex other than *v* is not added to S at the first *i* steps w.p. exactly $(1- d/n)^i$.

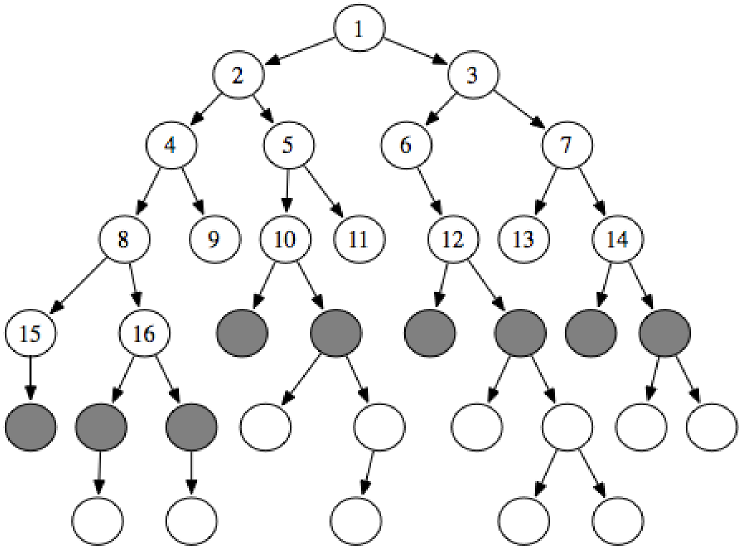
|S| after *i* iterations has distribution 1 + Binomial(n-1, $1 – (1-d/n)^i$).

For small *i*, the expected size of S is $\approx id$.

Rough analysis of the process

- The expected size of the "frontier", i.e. |S| - i,

  is approximately $\approx id - i = i\,(d - 1)$.

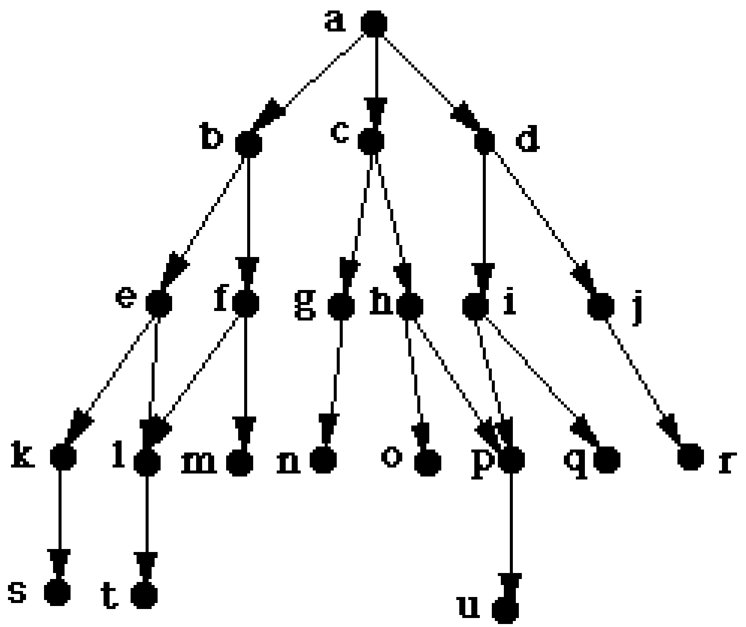


- For $d < 1$, the expected size of the "frontier" is negative.
- For $d > 1$, the expected size of the "frontier" increases, but the rate of discovering new vertices decreases when more vertices have been discovered.

  When $(d-1)/d$ fraction of vertices are discovered, this rate is 1.

  After that, the "frontier" shrinks.

Before threshold: d < 1



Thm. If *p = d/n*, with probability *1 – 1/n*, the sizes of all components are at most $\frac{4\ln n}{(1-d)^2}$

Proof: By union bound, it suffices to show for each vertex that w.p. ≤ $1/n^2$, its component is of size greater than $k = \frac{4\ln n}{(1-d)^2}$.

If component size is bigger, then |S| - k ≥ 1 at step *k*, i.e. random variable *Binomial(n-1, 1-(1-d/n)^k)* with mean at most *dk* is at least *k.* This happens with probability at most $(e^{1-d}d)^k$ by Chernoff bound:

$$\Pr(X > (1+\delta)\mu) < \left(\frac{e^\delta}{(1+\delta)^{(1+\delta)}}\right)^\mu .$$

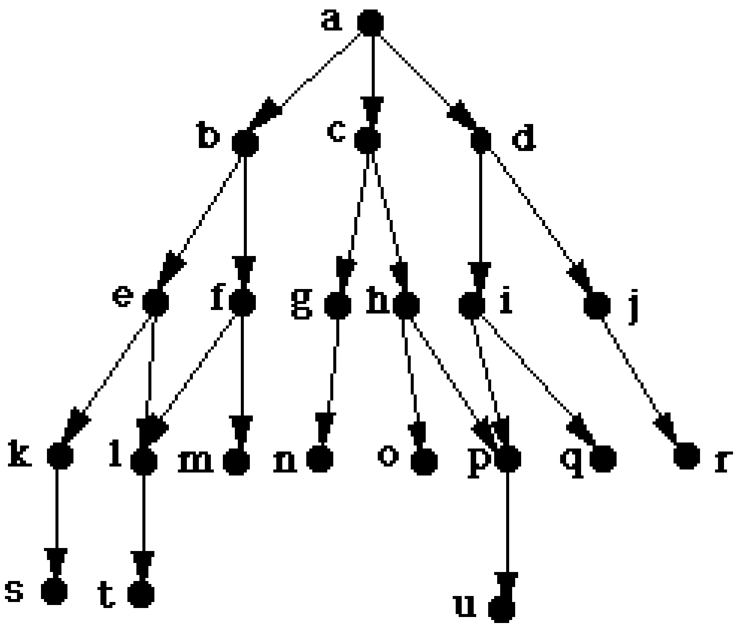
After threshold: d > 1

Thm. For each $d > 1$, there are constants c_1 and c_2 such that w.p. $\geq 1 - 1/n$, all component sizes are either $\leq c_1 \ln n$ or $\geq c_2 n$.

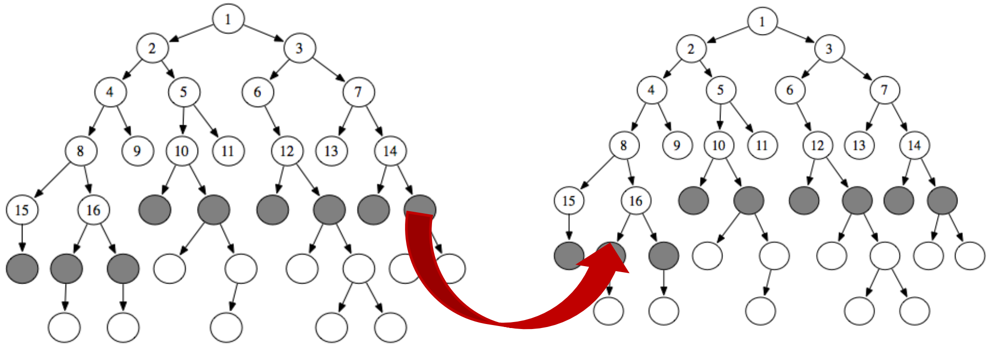
Proof: By union bound, it suffices to show for each vertex and $c_1 \ln n \leq i \leq c_2 n$ that the size of the component of that vertex is i w.p. at most $1/n^3$.

The probability is at most $Pr[Binomial(n-1, 1-(1-d/n)^i) = i]$.

The mean of the binomial variable is $id - O(i^2 d^2/n)$,
which is $i(1 + \Omega(1))$ for $i \leq c_2 n$ when c_2 is suitably small.
By Chernoff bound, the probability is at most $exp(-\Omega(i))$,
which is $\leq 1/n^3$ for $i \geq c_1 \ln n$ when c_1 is suitably large.



Two big components cannot coexist!



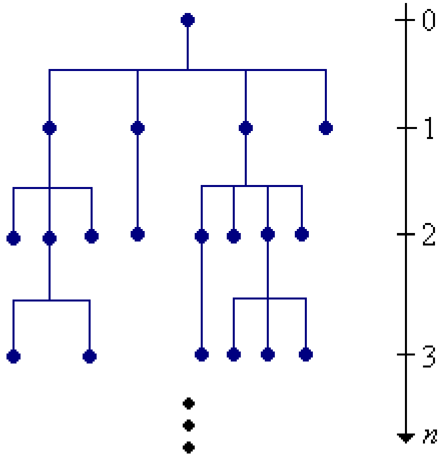
- Thm. Assume d > 1. The probability that at least two components of size ≥ $n^{2/3}$ exists is at most 1/n.

Proof.

- Let u and v be two vertices. Do BFS from both of them for $n^{2/3}$ steps.
- Either one of the BFSs finishes before that many steps, or the two BFS trees share vertices, or else w.p. ≥ $1 - 1/n^3$ by Chernoff bound both frontiers at step $i = n^{2/3}$ are of size $\Omega(n^{2/3})$.
- Since the frontier has not yet been explored, each pair of vertices from the two frontiers are independently connected with probability d/n.
- The probability that the two components are distinct is

$$\leq (1 - d/n)^{\Omega(n^{4/3})} \leq 1/n^3.$$

Branching process



- A method for creating a possibly infinite tree:

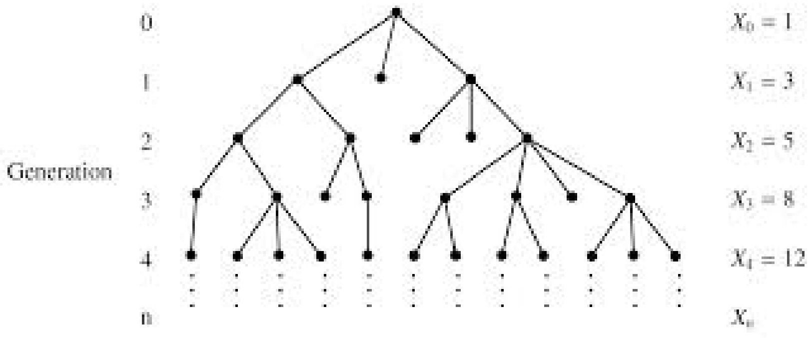
Let Y be a non-negative integer random variable

- Start from the root
- Choose a value according to the distribution of Y and spawn that many children
- For each of the root children, choose their # children independently according to the distribution of Y

…

Thm. If E[Y] > 1, extinction probability is < 1.

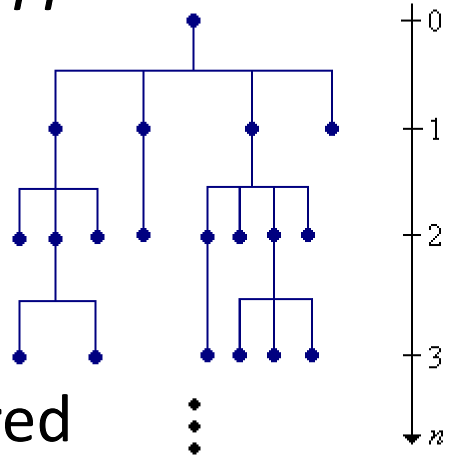
- We assume Y is bounded; otherwise truncate Y.
- Let p_i = Pr[Y = i].
- $p_0 < 1$.



- There exists $p_0 \leq \alpha < 1$ such that $f(\alpha) = \sum_i p_i \alpha^i \leq \alpha$ (because *f(1) = 1, f'(1) > 1*)
- By induction on *t*, Pr[extinction in t levels] ≤ α.

- Pr[extinction] = $\lim_{t \to \infty}$ Pr[extinction in t levels] ≤ α.

For d > 1, each vertex is with constant positive probability not in a component of size $\leq c_1 \ln n$



- Do BFS from vertex *v.*

- While # discovered vertices $\leq c_1 \ln n$,

  the distribution of # undiscovered neighbors of a vertex being explored

  dominates Binomial($n - c_1 \ln n$, d/n), which in turn

  dominates a random variable Y (depending on d but independent of n)

  with mean > 1.

- The probability that this branching process does not become extinct is positive independent of *n.*

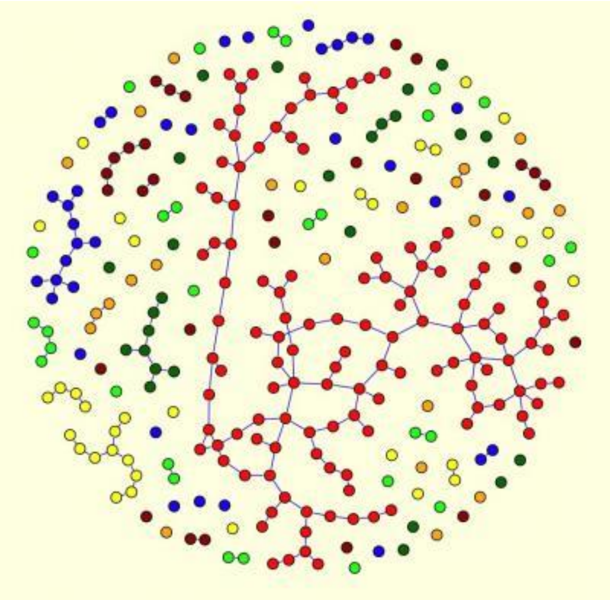
There exists a giant component when d > 1.

- Choose a vertex. With Ω(1) probability it is in a giant component.
- Otherwise, almost surely, it is in a component of size O(ln n).
- Remove that component from the graph.
- The remaining graph is an Erdos-Renyi graph, still with average degree 1 + Ω(1).
- Now repeat the above for the remaining graph.

You can do the above for ω(1) steps.

Then almost surely a giant component is found.

For another proof using second moment, see the textbook.



Size of the giant component?

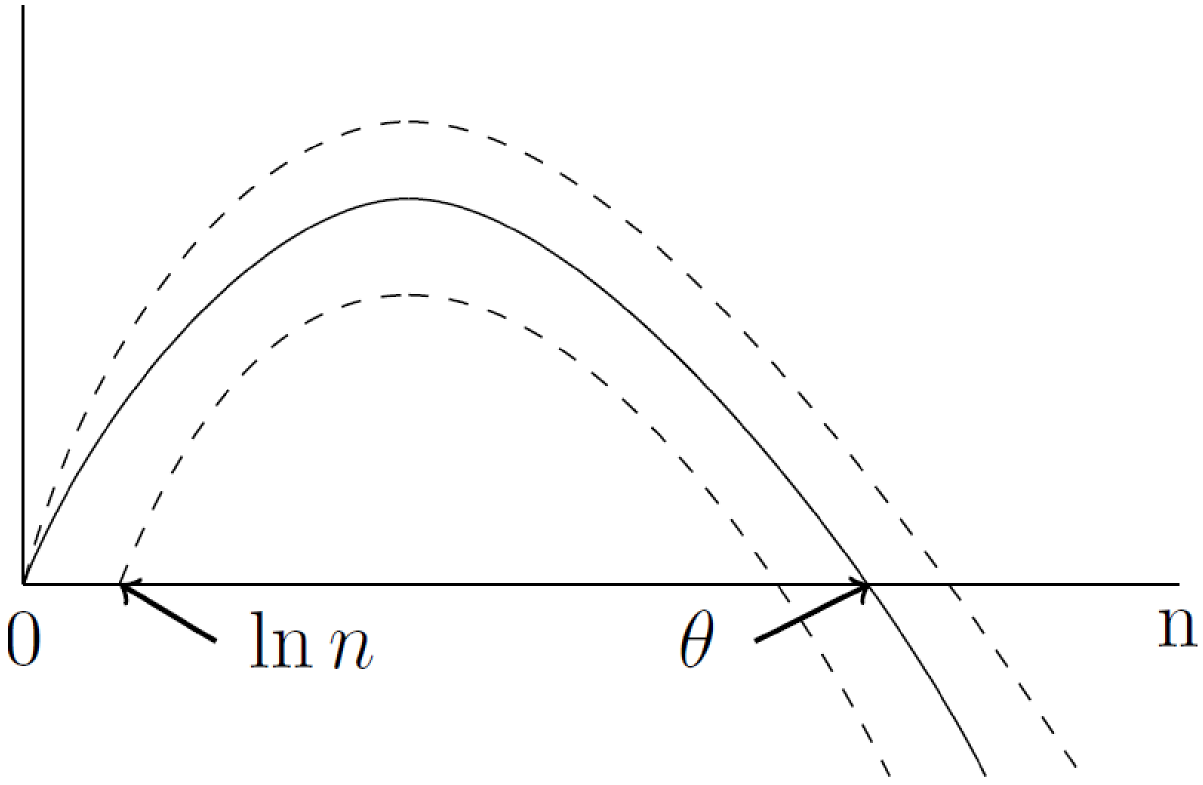
Expected size of frontier = 0

when $n(1\text{-}d/n)^{\theta} = n - \theta$.

In other words

$\exp(-d\ (\theta/n)\ ) = 1 - \theta/n$.

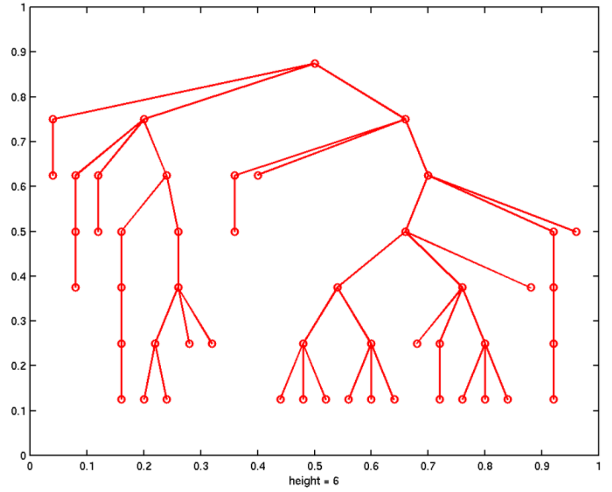
(Without giving the proof) the expected size of the giant component is approximately this θ.



Solid curve = expected value of the frontier
Dashed curve = probable range for the frontier

Branching Processes

What do we study about branching processes?



We will derive the exact value of

- the extinction probability
- the expected size of the tree conditioned on extinction

In particular, when the expected number of children is not 1, the conditional expected size is finite.

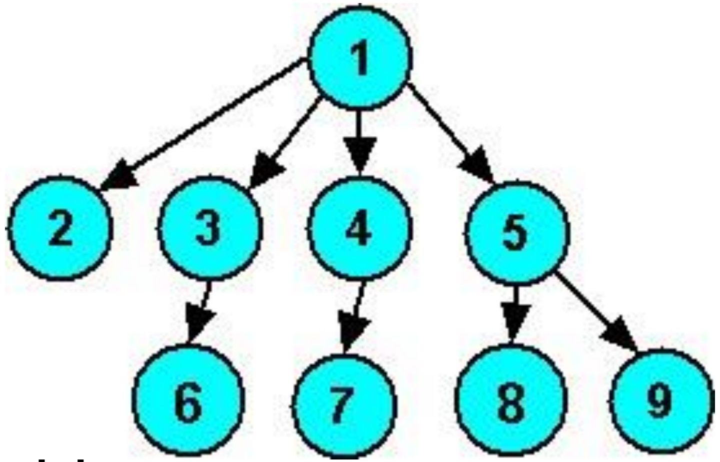
We know that *G(n, d/n),* when *d > 1,* consists of a giant component of size Ω(n) and small components of size O(lg n). This suggests that the expected size of the small components is constant.

Generating function

- Let *Y* be the random variable equal to the number of children of a node.
- Let $p_i = Pr[Y = i]$.
- The generating function for *Y* is the function $f(x) = \sum_{i=0}^{\infty} p_i x^i$.

“A generating function is a clothesline on which we hang up a sequence of numbers for display.” Herbert Wilf, Generatingfunctionology

Composition of generating functions



If *f(x)* is probability generating function for # children
for every node in 1st generation and
g(x) is probability generating function for # children
for every node in 2nd generation,
f(g(x)) is probability generating function for # grandchildren.

Proof. If *g(x)* is p.g.f. for *Y* and *h(x)* is p.g.f. for *Z*,
and *Y, Z* are independent, then
g(x)h(x) is the p.g.f. for *Y + Z*.

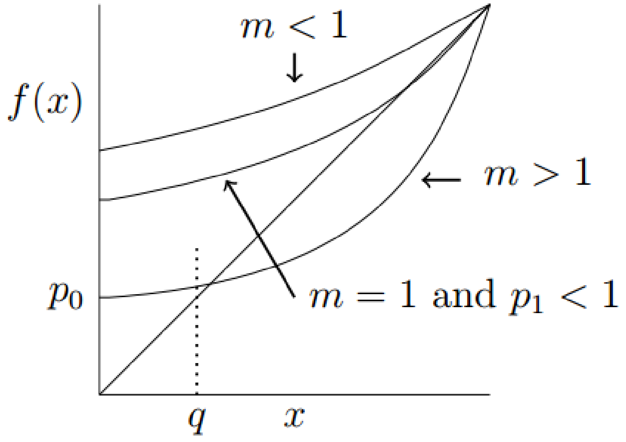
children in *j*th generation

- The generating function for total #children in *j*th generation is $f_j(x)$,

  where $f_{j+1}(x) = f(f_j(x))$ and $f_1(x) = f(x)$.

- The functions $f_j(x)$ are power series with non-negative coefficients. Therefore, they are non-decreasing and convex on [0, 1].

- If $p_0 < 1$, they are also strictly increasing.

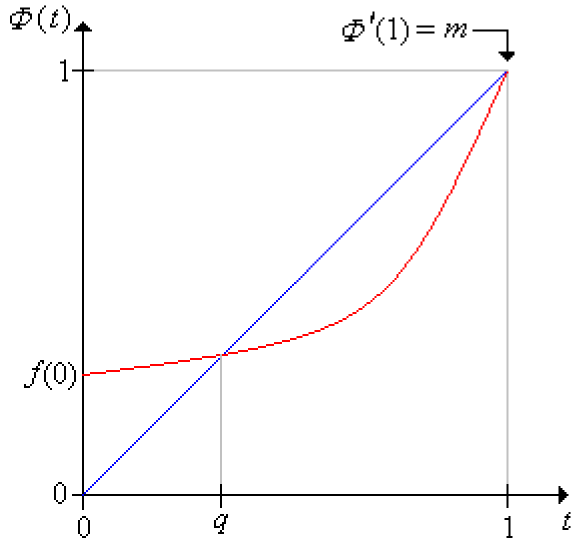


Probability of extinction

If q is the probability of extinction, we have $q = \sum_{i=0}^{\infty} p_i q^i$.
In other words, q is a root of $f(x) = x$.
1 is always a root of $f(x) = x$.

- If ($E[Y] < 1$) or ($E[Y] = 1$, $p_1 < 1$), then the only root is $q = 1$
  because $f'(1) \leq 1$ and f is strictly convex.
- If $Y = 1$, then q = 0.
- If E[Y] > 1, there is only one root < 1
  since f'(1) > 1 and *f* is convex.

Since q is not 1, q is this other root.



Another way of deriving extinction probability

If q is the smallest root of f(x) = x,

then $f_j(0)$ tends to q as j gets larger.

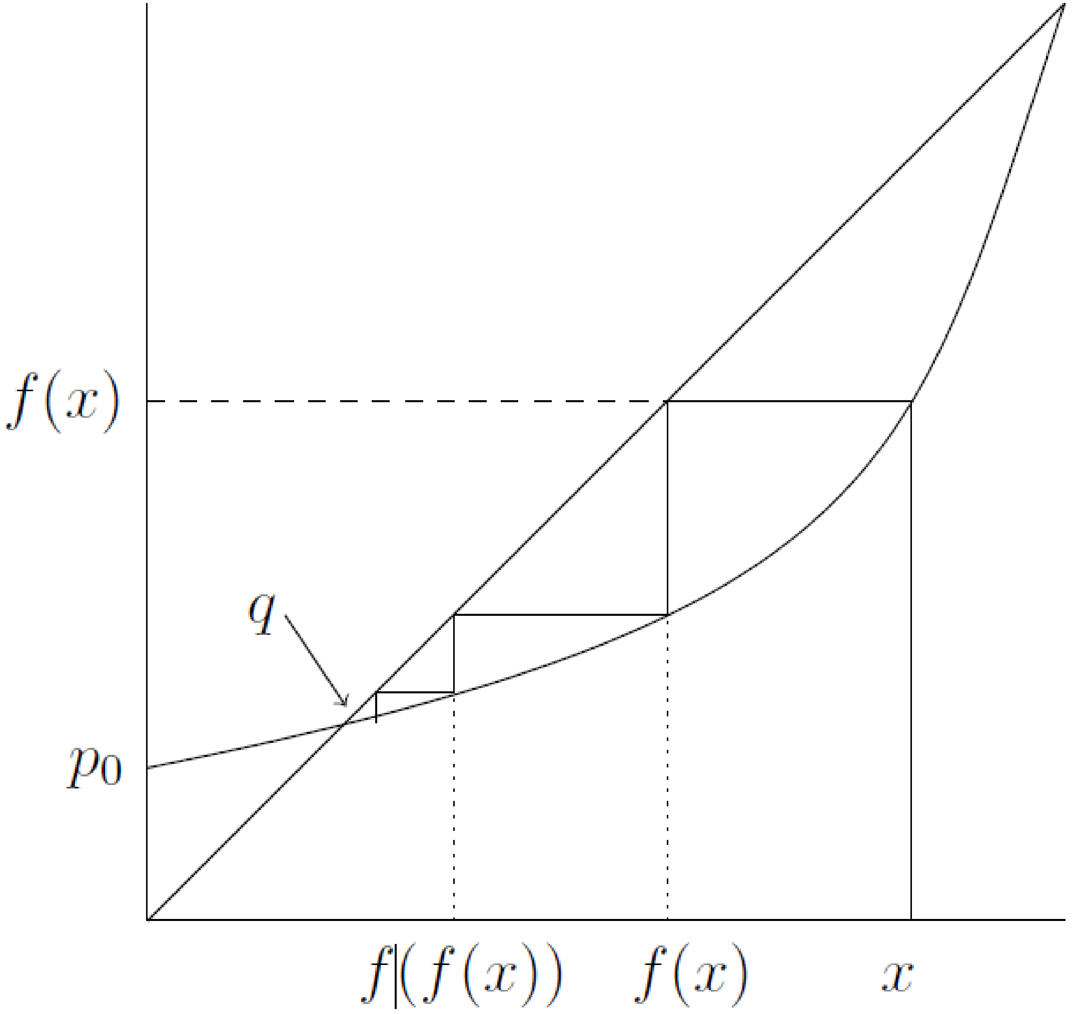
Therefore, extinction probability is q.

Also for any x, $f_j(x)$ tends to q

as j gets larger.

Thus, coefficients of non-constant terms

in $f_j(x)$ tends to zero.



Real biological systems

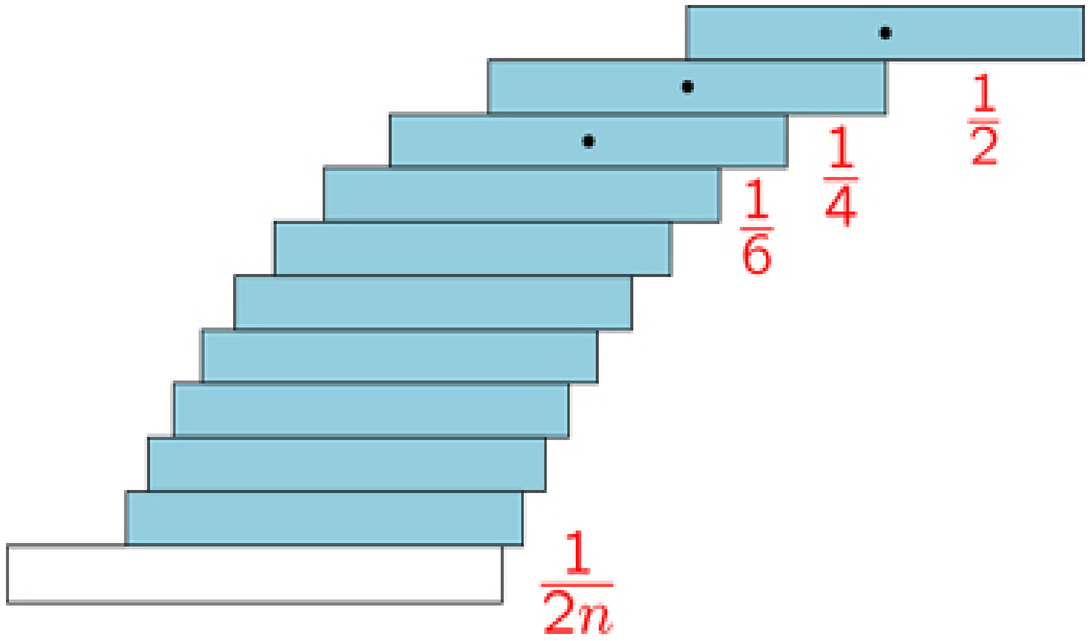
- In the branching processes we analyzed, the population either dies out or the population size goes to infinity.
- In real world, processes often go to *stable* populations.
- This is due to other factors, like the distribution of # children depends on the size of whole population.

Expected size of extinct families

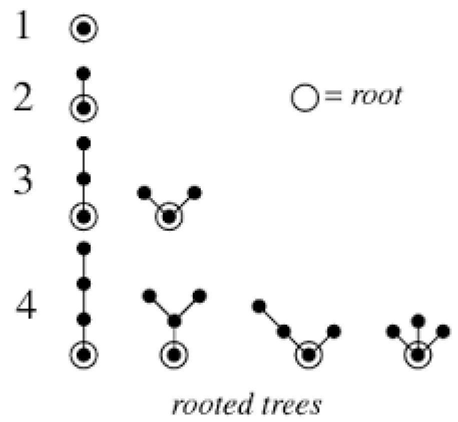
Finite random variable may have infinite expected value

Let X be a positive integer random variable with $p_i = 6/(i^2 \pi^2)$.

$$EX = \sum_{i=1}^{\infty} i \cdot 6/(i^2\pi^2) = 6/\pi^2 \sum_{i=1}^{\infty} 1/i = \infty$$



Expected size of extinct families (easy cases)



- *E[Y] < 1*: It dies out with probability 1.

  Expected size of level *l* is $E[Y]^l$.

  Expected tree size = $1/(1 - E[Y])$.

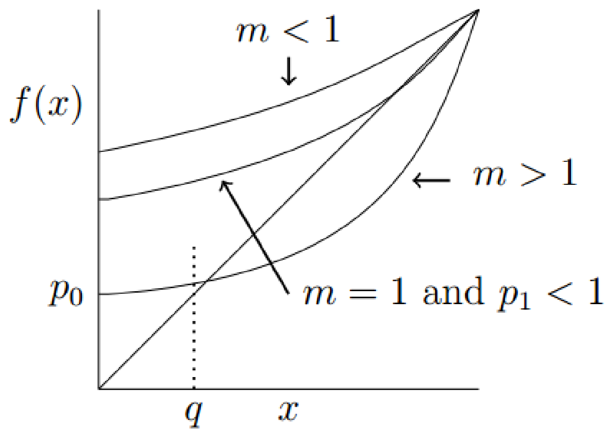
- *E[Y] = 1, Pr[Y = 1] = 1*: The tree never dies.
- *E[Y] = 1, Pr[Y = 1] < 1*: The tree dies out with probability 1.

  Expected size at level *l* is 1.

  Expected tree size is infinity.

Expected size of extinct families: case E[Y] > 1

Note $f'(q) < 1$



- Let the root have *i* children.
- Pr[tree finite | *i*] = q^i.

- By Bayes rule, Pr[*i* | tree finite] = Pr[tree finite | *i*] p_i / Pr[tree finite]

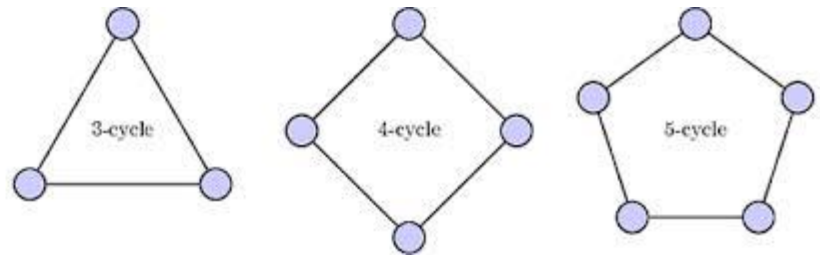
  $= q^i p_i / q = p_i q^{i-1}$

- We now have a new branching process with probabilities $p_i q^{i-1}$.
- Expected number of children in this branching process is $f'(q)$.
- Expected size of extinct families = $1/(1 - f'(q))$.

Emergence of cycles

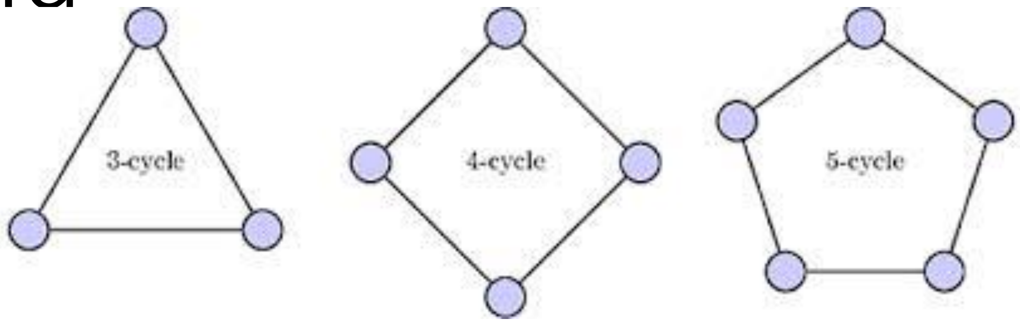
Theorem. Threshold for emergence of cycles is *p = 1/n.*

- Expected # cycles = $\sum_{k=3}^{n} \frac{n(n-1)\ldots(n-k+1)}{k!} \frac{(k-1)!}{2} (d/n)^k$.
- The above sum is at most $\sum_{k=3}^{\infty} d^k$.



- When d = o(1), the expected # cycles is o(1), so by 1st moment method, there is a cycle with probability only o(1).
- When d = ω(1), we already showed there is a triangle almost surely.

cycles around the threshold



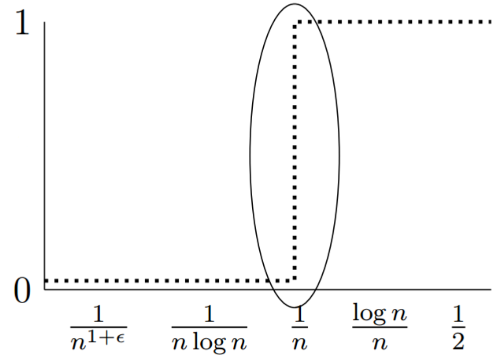
- Suppose *d* is constant.

- If *d < 1*, expected # cycles ≤ $\sum_{k=3}^{\infty} d^k = O(1)$.

- If *d ≥ 1,* expected # cycles is at least
$$\sum_{k=3}^{\lg n} \frac{n(n-1)\ldots(n-k+1)}{2kn^k} = \sum_{k=3}^{\lg n} \frac{1-o(1)}{2k} = \omega(1).$$

Threshold for emergence of cycles is not sharp.

- When *d = 1 + Ω(1)*, there is a giant component in *G(n, (1+d)/(2n))*.
- *G(n, d/n)* has a lot more edges than *G(n, (1+d)/(2n))*, and each extra edge forms a cycle in the giant component with constant probability.
- Therefore, there are ω(1) cycles in *G(n, d/n)* almost surely.



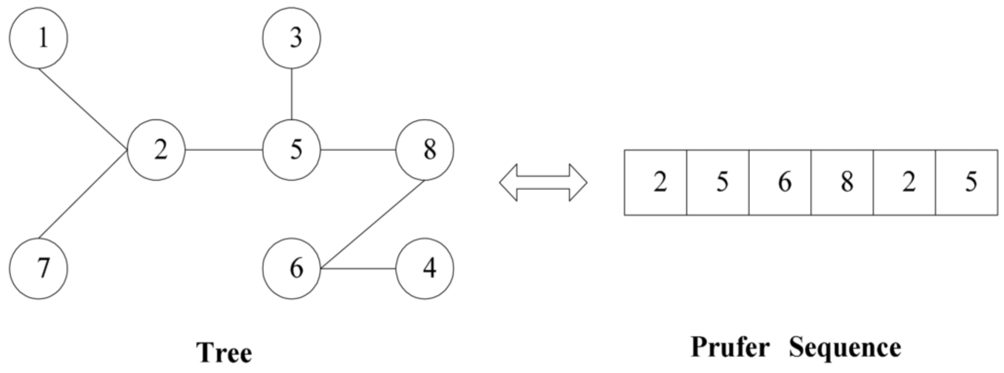
- When *d = 1 – Ω(1)*, do BFS over the whole graph.
- In each connected component, other than the BFS tree we have not finalized existence of other edges.
- There are on average O(n) non-finalized edges (since expected size of components is O(1) by branching processes).
- Therefore, with at least positive constant probability, there is no cycle.
- Also, with at least positive constant probability, there is a cycle.

Full connectivity

connected components of size *k*

The expected # connected components of size *k* is at most

$$\binom{n}{k} k^{k-2} p^{k-1} (1-p)^{k(n-k)}.$$

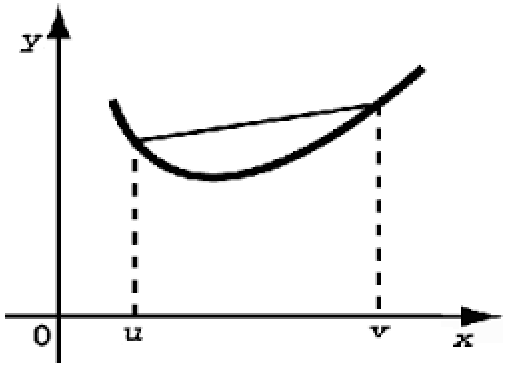


- # trees on *k* vertices is k^{k-2}.
- # tree edges = $k - 1$
- # edges crossing the component = $k(n-k)$.

When $p = c \ln n / n$ for constant $c > ½$, there is no component of size between 2 and $n/2$.

$$\binom{n}{k} k^{k-2} p^{k-1} (1-p)^{k(n-k)} \leq e^{f(k)}$$

for $f(k) = \ln n + k + k \ln \ln n - 2 \ln k + k \ln c - ck \ln n + ck^2 \ln n / n$.
using $\binom{n}{k} \leq (ne/k)^k$ and $1 - p \leq e^{-p}$.



$f''(x) > 0$ so $f(k)$ attains maximum over $k \in [2, n/2]$ at the endpoints.
$f(k) \approx (1 - kc) \ln n$ for constant k and $f(n/2) \approx -cn \ln n / 4$.
Thus, expected number of components of size in $[2, n/2]$ is $O(n^{(1-2c)(1+o(1))})$.
Now use 1st moment method.

Thm. $p = \ln n / n$ is sharp threshold for connectivity.

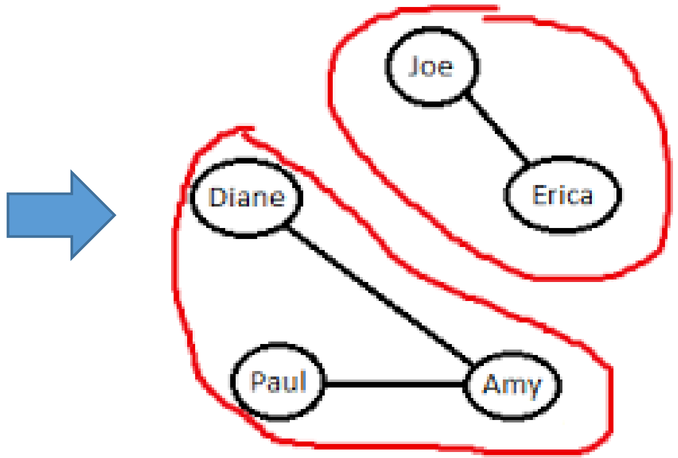
- Let $p = c \ln n / n$.

- For $c < 1$, we already showed there is an isolated vertex.

- For $c > 1$, there is no isolated vertex.

  So almost surely all components are of size $> n/2$.

  But there cannot be two components of size $> n/2$.



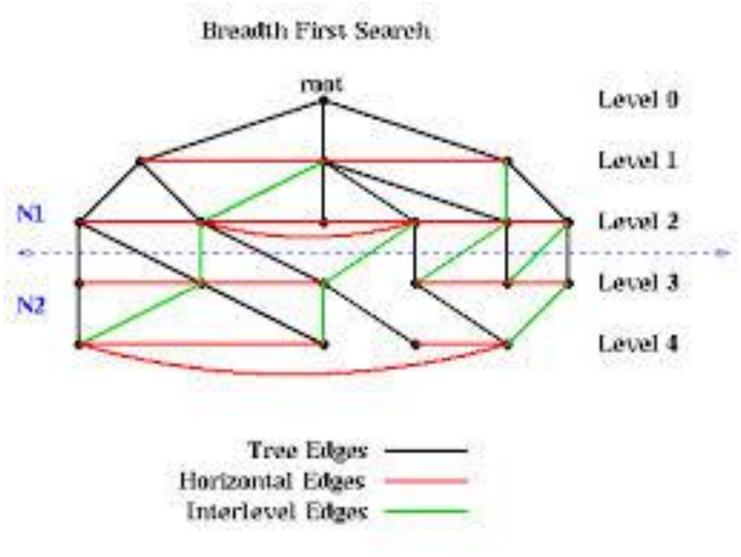
Threshold for logarithmic diameter

When $p = c \ln n / n$ for large constant c, the graph has diameter $O(\log n)$.

If you run BFS from a vertex,

the first level has $\geq c\,(1 - \varepsilon) \ln n$ vertices for large c.

(We proved concentration for degrees at the beginning of course.)



If S_l is nodes at level l, while $|S_1|+\ldots+|S_i| \leq n/1000$,

by Chernoff w.p. $1 - \exp(-\Omega(|S_i|))$, $|S_{i+1}| \geq 2\,|S_i|$.

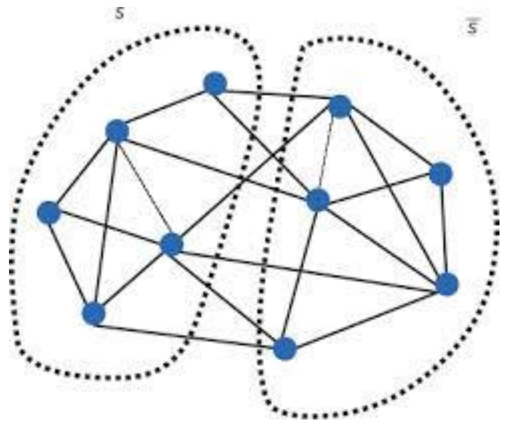
(The expected size of S_{i+1} is at least $200\,|S_i|$.)

By union bound, the neighborhood of each vertex at distance O(lg n) is of size $\geq n/1000$.

Almost surely, there is an edge between any two disjoint sets of vertices of size *n/1000.*

The probability that there is no edge between sets *S* and *T* is

$$(1-p)^{|S||T|} \le e^{-p|S||T|} \le e^{-c(\ln n)n/10^6}.$$



There are only 2^{2n} such pairs of sets.

By union bound, almost surely

all such sets *S* and *T* are connected.

In particular, neighborhoods of logarithmic depth for any two vertices are connected.

Summary of phase transitions we proved

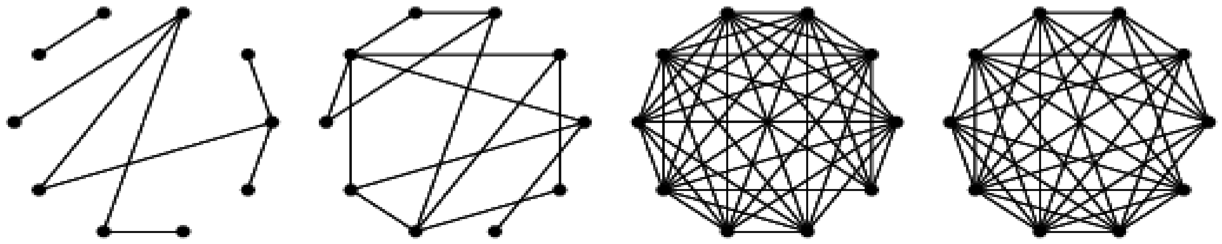
Property	Threshold
cycles	$1/n$
giant component	$1/n$
giant component + isolated vertices	$\frac{1}{2}\frac{\ln n}{n}$
connectivity, disappearance of isolated vertices	$\frac{\ln n}{n}$
diameter two	$\sqrt{\frac{2\ln n}{n}}$

Phase transitions for increasing properties

Do all graph properties have thresholds for Erdos-Renyi graphs?

- All increasing properties have a threshold.

- A property is increasing if when *G* has the property, adding edges it still has the property.

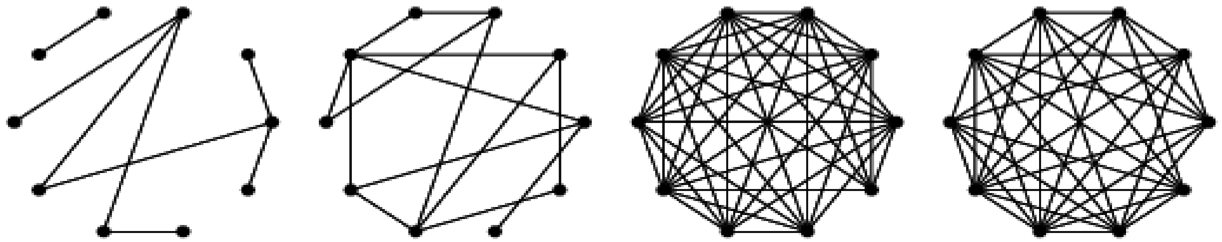


- Examples of increasing properties: having cycle, connectivity, no isolated vertices, having giant component, Hamiltonicity, …

For increasing property Q, and $0 \le p \le q \le 1$, $\Pr[G(n, p) \text{ has } Q] \le \Pr[G(n, q) \text{ has } Q]$

Proof. Generate *G(n, q)* as follows:

- first sample *G(n, p)*
- Between every pair of nodes that is not an edge in *G(n, p),* add an edge with probability $(q - p) / (1 - p)$.

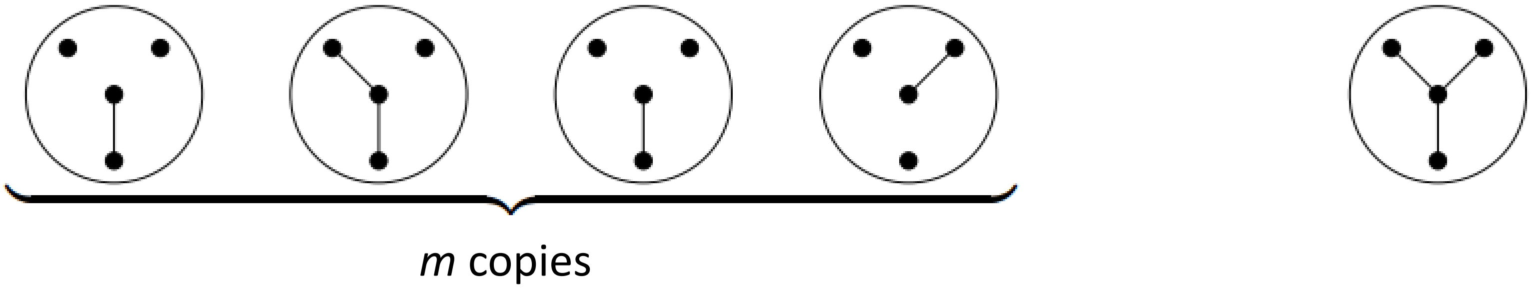


With the above sampling, if *G(n, p)* has property *Q,* so does *G(n, q).*

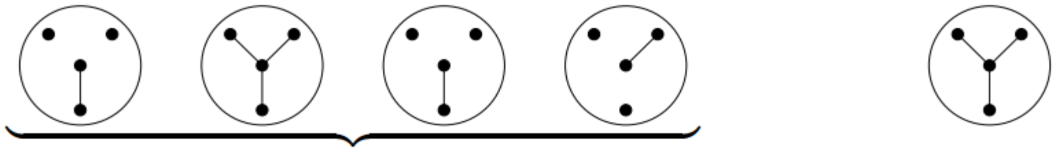
m-fold replication of $G(n, p)$

is a new graph with n vertices whose edges are the union of m independent copies of $G(n, p)$.

It is equivalent to $G(n, q)$ for $q = 1 - (1 - p)^m \leq mp$.

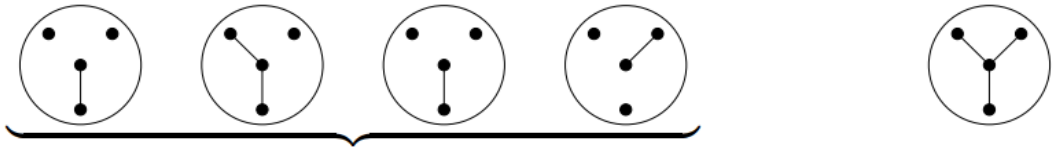


Relation of *m*-fold replication with *G(n,p)*



If any graph has three or more edges, then the m-fold replication has three or more edges.

- Pr[G(n,mp) has Q] ≥ Pr[G(n, q) has Q]



Even if no graph has three or more edges, the m-fold replication might have three or more edges.

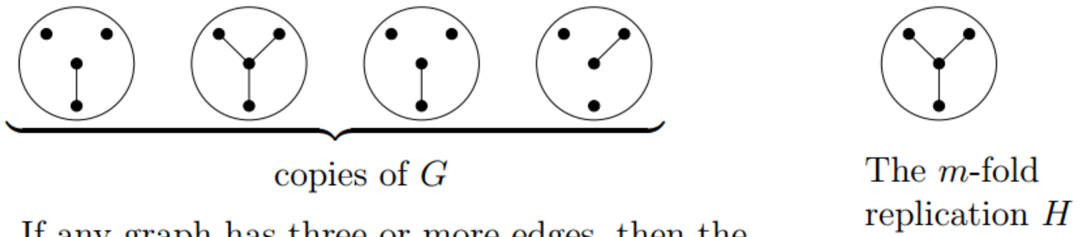
- Pr[G(n, q) has Q] ≥ 1 – (1 – Pr[G(n, p) has Q])^m.

Thm. Increasing properties have thresholds.

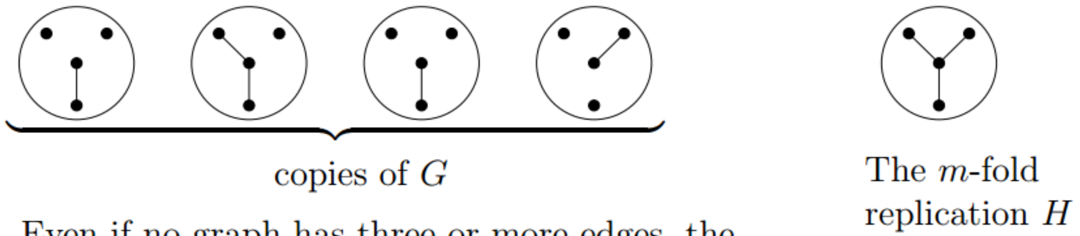
Let p be such that Pr[$G(n, p)$ has Q] = ½.

- If $p' = mp$, Pr[$G(n, p')$ has Q] $\geq 1 - (1 - \Pr[G(n,p) \text{ has } Q])^m = 1 - 2^{-m}$.
- If $p' = p/m$, $1/2 = \Pr[G(n, p) \text{ has } Q] \geq 1 - (1 - \Pr[G(n,p') \text{ has } Q])^m$.

Thus, p is a threshold.



If any graph has three or more edges, then the m-fold replication has three or more edges.



Even if no graph has three or more edges, the m-fold replication might have three or more edges.