Social Network Analysis
Random Networks

Robin Burke
DePaul University
Chicago, IL

Thanks to Professor Matthew Jackson for his notes on this topic.

Announcements

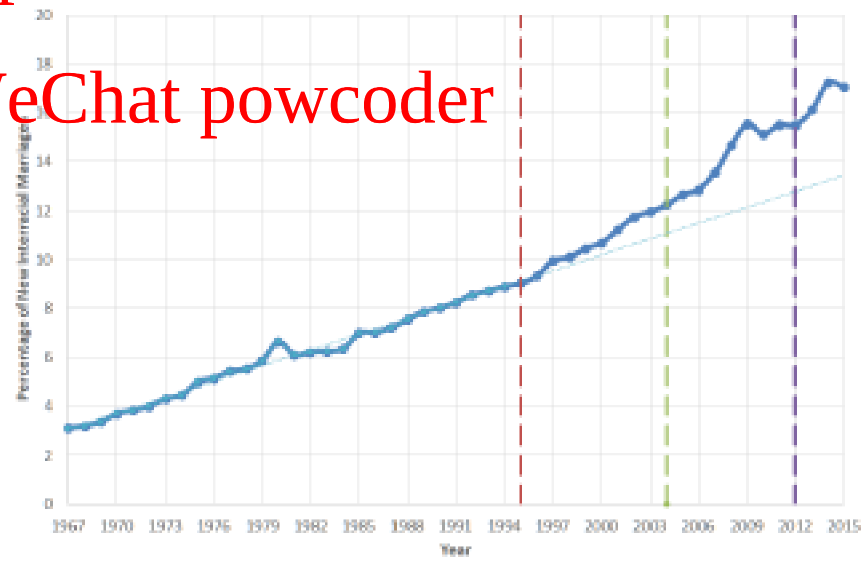
- Due tonight
  - Project data
- Next week
  - First visualization

First visualization

- A network visualization of your project data in Gephi
  - Could be a subset if the data is large
- Aim
  - Convey node importance
- Choose appropriate layout and mappings
- Two Powerpoint slides
  - visualization
  - one paragraph description
- Submit
  - to D2L

Cool result

- Online dating reduces homophily
- Ortega and Hergovich, 2017



Random networks

- Why do we study random networks?
  - the same reason that statisticians study random variables
- If we want to know if something is not happening by chance
  - "is significant"
- we have to know what chance looks like

Previous examples

- transitivity
  - “what is the transitivity of a random network of the same size?”
- reciprocity
  - “is a social network more reciprocal than a random network?”
- assortativity
  - “what if the edges were random?”

What is a random network?

- Static model
  - collection of nodes
  - edges randomly placed
- Dynamic model
  - nodes arrive
  - linked to existing network with random connections
- etc.

Erdos-Renyi Random Network

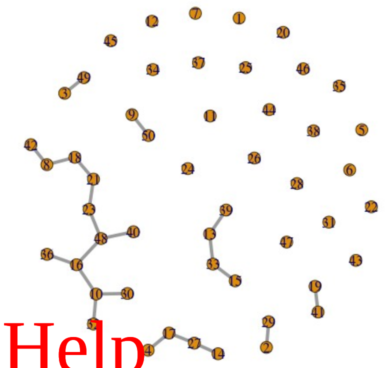
- Specify n, p
- Start with n nodes
- Go through all possible pairs of nodes
  - n(n-1)/2 for undirected network
- Create edges with probability p
- In R
  - erdos.renyi.game(n, p, mode="gnp")
  - also erdos.renyi.game(n, m, mode="gnm")
    - place m edges randomly

ER Random Network

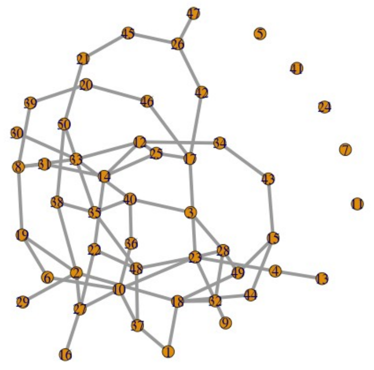
- Simplest network specification
  - only n and p
- Formed the basis for mathematical study of networks
  - for most of the 20th century

Examples

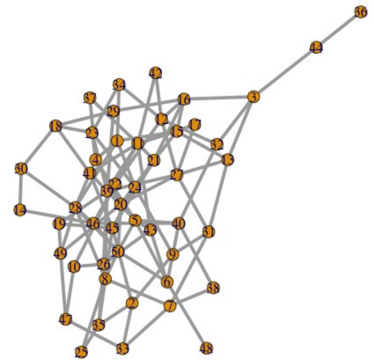
- ER n=50, p=0.02



- ER n=50, p=0.05



- ER n=50, p=0.1



Characteristics of random networks

- Sparse networks
  - somewhat stringy
- Dense networks
  - disordered looking
- Giant component "tipping point"
  - as p increases
  - probability of all nodes being one component → 1
  - tipping point: $p \approx 1/n$

Properties of social networks

- High clustering / transitivity
  - transitive closure
  - not in random networks
- Skewed degree distributions
  - A few highly connected individuals
  - not in random networks
- High closeness
  - “six degrees of separation”
- Are social networks “special” in this?

Small world hypothesis

- People in social networks
  - have much higher closeness than you might expect
- We don't expect "six degrees of separation"
  - but we see it
  - is this a significant effect?

Small world hypothesis

- Restatement
  - The average path length in a random network is large
    - O(n), maybe
  - As opposed to O(log(n))
- If this is true
  - then the shorter path lengths in social networks are “interesting”

Some preliminaries

- The network can't be too dense
  - then it is easy to achieve low average path length
  - p = 1.0?
- The network can't be too sparse
  - no connected component
- We will talk about large networks

Conditions

- $d(n) >= (1+\varepsilon)\log(n)$ for some $\varepsilon>0$
  - This makes the network dense enough for connectivity
- $d(n)/n \to 0$
  - This makes the network sparse enough
  - Sequences of networks

Theorem

- If the conditions are met
  - $d(n) >= (1+\varepsilon)\log(n)$ for some $\varepsilon>0$
  - $d(n)/n \to 0$
- Then
  - for large n, average path length is proportional to $\log(n) / \log(d)$

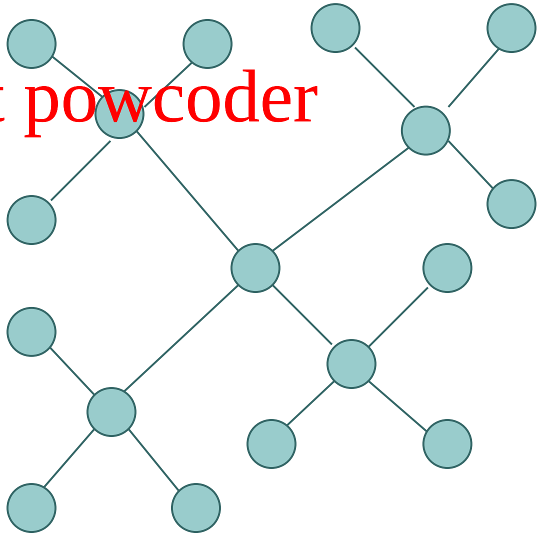
Restatement

- If the conditions are met
  - $d(n) >= (1+\varepsilon)\log(n)$ for some $\varepsilon>0$
  - $d(n)/n \rightarrow 0$
- Then
  - as $n \rightarrow \infty$

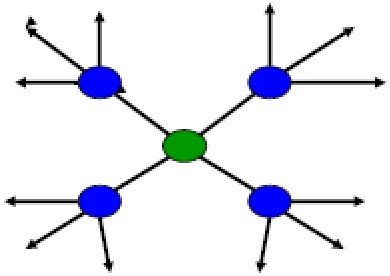
$$AveDist(n) \propto \frac{\log(n)}{\log(d(n))} \rightarrow 1$$

Simpler than a random graph

- A regular structure
- Each node has degree d
  - except leaves
- Cayley tree



Path lengths

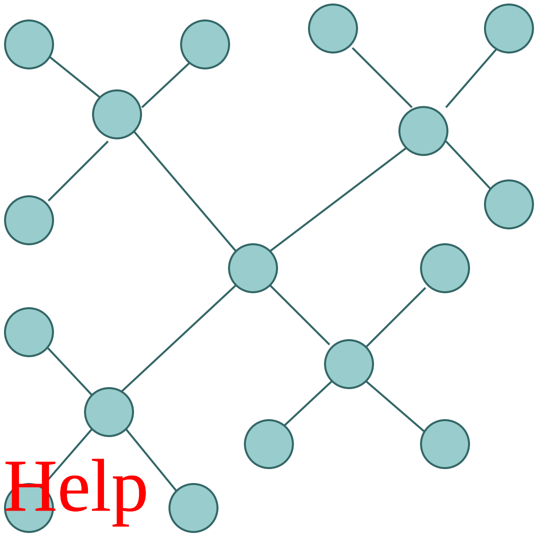


- 1 step: d nodes
- 2 step: d(d-1)
- 3 step: $d(d-1)^2$
- …
- k steps: roughly d^k

Diameter

- Diameter is 2k
- How many nodes?
  - $d+d(d-1)+\ldots d(d-1)^{k-1}$
  - $d((d-1)^k-1)/(d-2)$
  - roughly $(d-1)^k$
  - $(d-1)^k=n$
  - k on the order of log(n)/log(d)

Cayley Tree: Diameter



- Diameter is 2k
- How many nodes?
  - $d+d(d-1)+\dots d(d-1)^{k-1}$
  - $d((d-1)^k-1)/(d-2)$
  - roughly $(d-1)^k$
  - $(d-1)^k=n$
  - k on the order of log(n)/log(d)

But random graph?

- degree of nodes not identical
- BUT we can show that
  - the fraction of nodes with nearly average degree
  - approaches 1
- This is the reason for the bound
  - E[d] > (1+ ε)log(n)
- Also have to deal with edges pointing "backwards"

We can show

- Probability that a node has degree close to average
  - (Jackson's book has details)
  - $\Pr(d/3 \le d_i \le 3d) \ge 1 - e^{-d}$
- For nodes of all degrees
  - $\Pr(d/3 \le \text{all degrees} \le 3d) \ge (1 - e^{-d})^n$
- Use our substitution
  - $\Pr(d/3 \le \text{all degs} \le 3d) \ge (1 - 1/n^{1+\varepsilon})^n$
  - $\exp(-n^{-\varepsilon}) \to 1$

From Chebyshev inequality

n is large so $1/n^{\varepsilon}$ is small.

So

- If $d(n) > (1+\varepsilon)\log(n)$ then
  - $\Pr(d/3 \le \text{all degrees} \le 3d) \to 1$
- A corollary
  - $\log(n)/\log(3d) < k < \log(n)/\log(d/3)$
- k is on the order of $\log(n)/\log(d)$

What about doubling back?

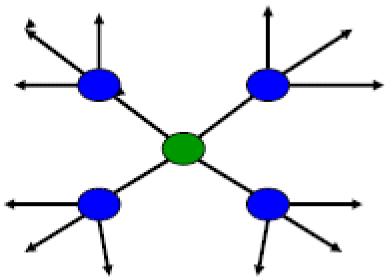
- Remember we are expanding outward
  - factor of d more nodes each step
- After k steps
  - d^k nodes reached
  - $n-d^k$ nodes unreached
- If k < log(n)/log(d) then
  - $n-d^k$ much bigger than d^k
  - most nodes are not reached until the last step
  - think about $n=10^6$, $d=10$

Diameter vs ave. distance

- Since most of the nodes are at maximum distance

  - average distance is the same order as diameter

  - (again, for large n)



Small world hypothesis

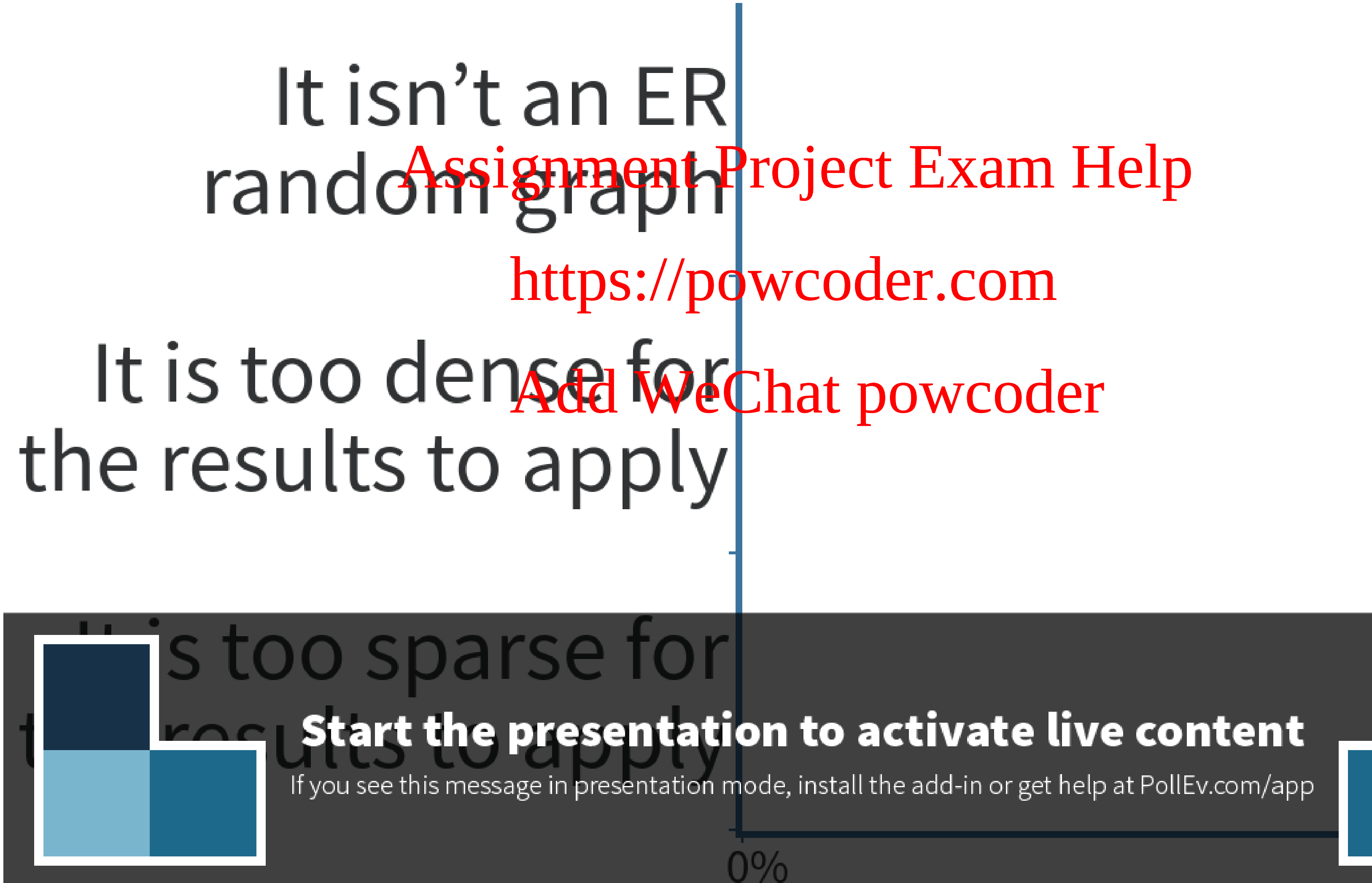
- Even in a random network, we should expect

  - a small diameter

  - a small average path length

- These properties are not special to social networks
  - properties of networks that are sparse, but not too sparse

…dom graph does not show the small world pr… …g that diameter grows as a function of N rath… log(N), it means



Start the presentation to activate live content

If you see this message in presentation mode, install the add-in or get help at PollEv.com/app

ER graph

- Model
  - n nodes
  - edges added with probability p
- What does this mean for the degree distribution?
  - we've seen that it has a different shape than many social netw

Degree distribution

- Binomial distribution
  - # of heads when we flip a coin k times
  - with probability of heads = p

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

  - k = n-1
    - other nodes for any given node
    - # of trials to add an edge
  - d = degree

$$\frac{(n-1)!}{d!(n-d-1)!} p^d (1-p)^{n-d-1}$$

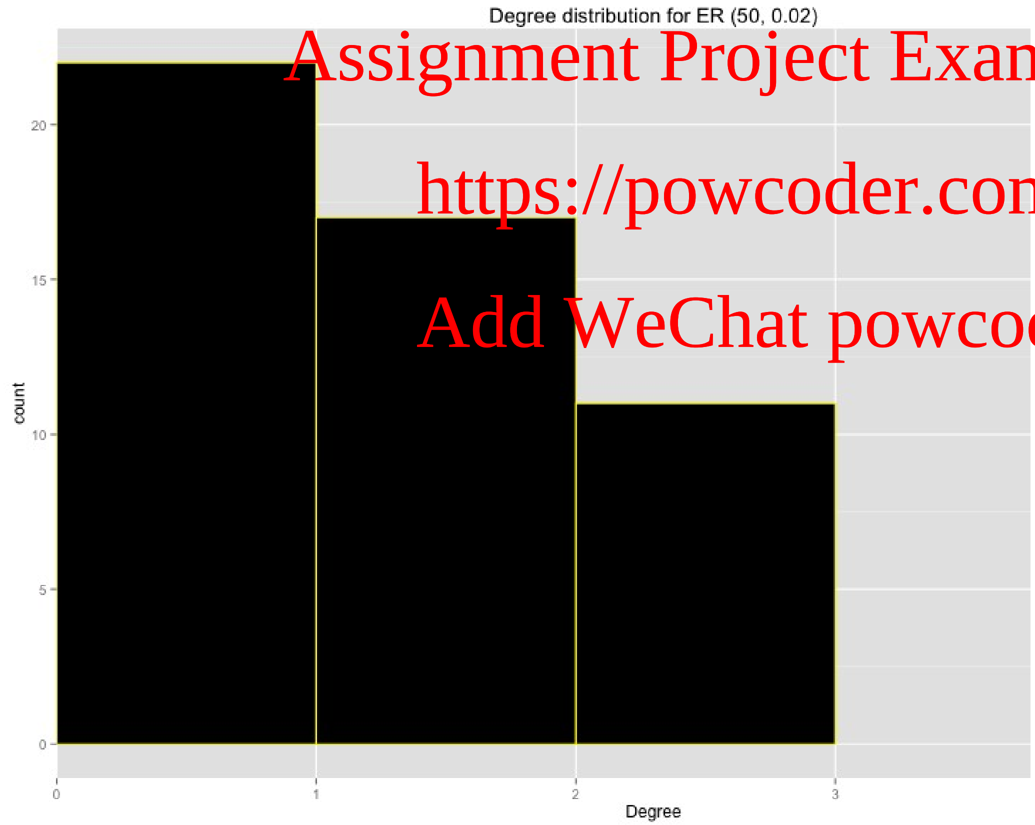
Poisson distribution

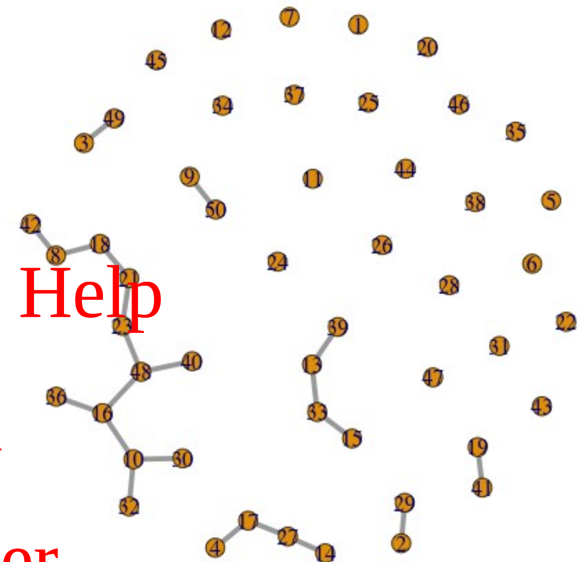
- For large n and small p
  - can be approximated with the Poisson distribution

$$\frac{(n-1)^d}{d!} p^d e^{-(n-1)p}$$

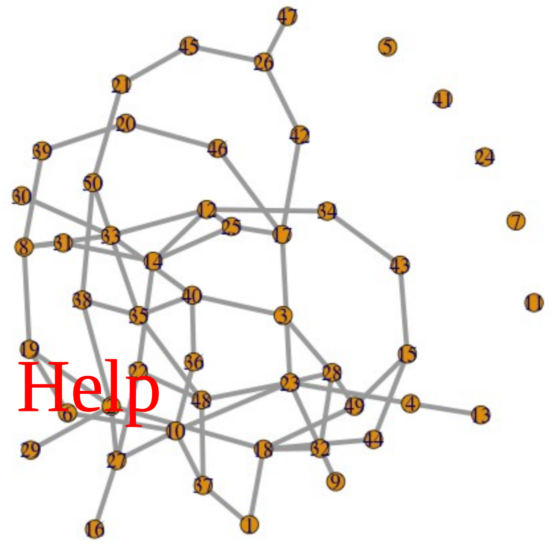
- ER random graphs
  - sometimes called “Poisson random graphs”

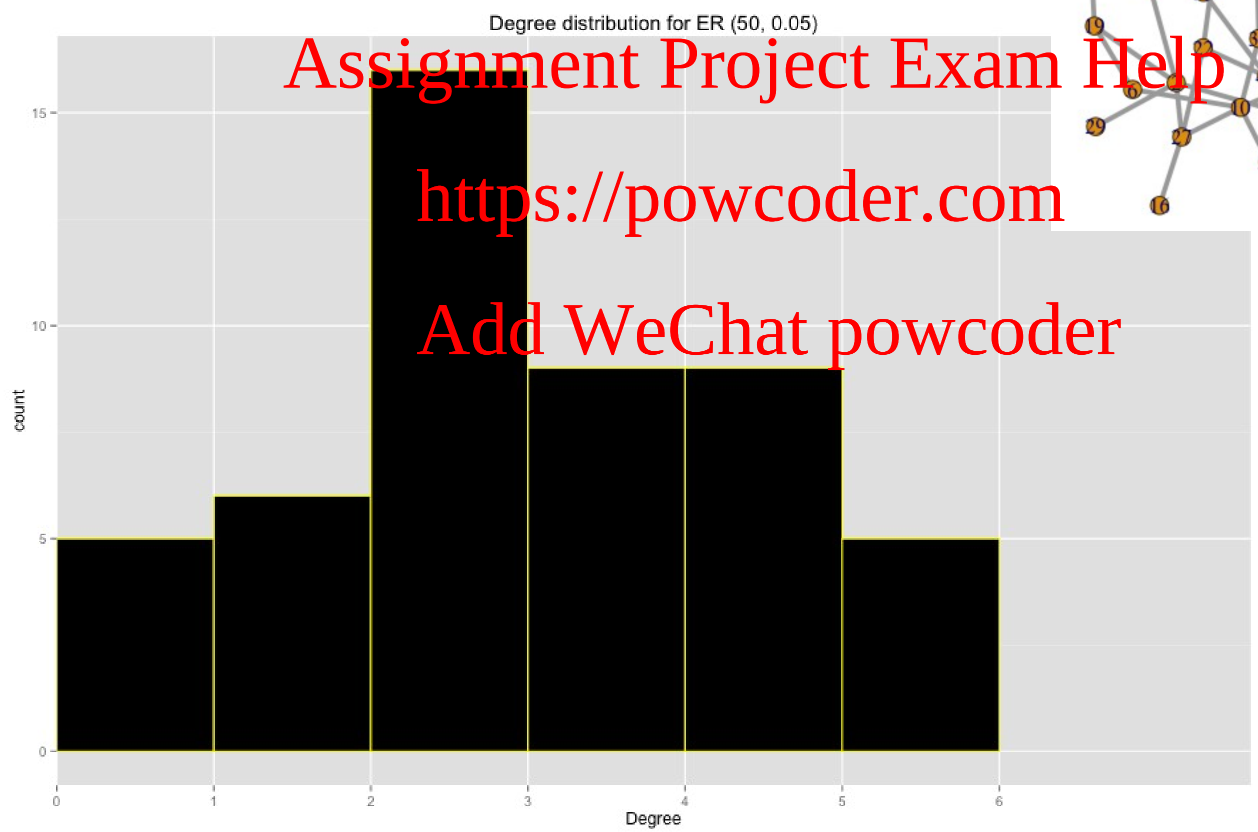
ER (50, 0.02)


Degree distribution for ER (50, 0.02)
count
Degree
0
5
10
15
20
0
1
2
3

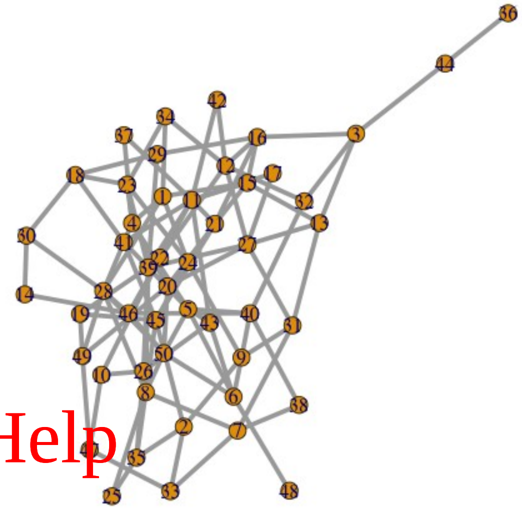


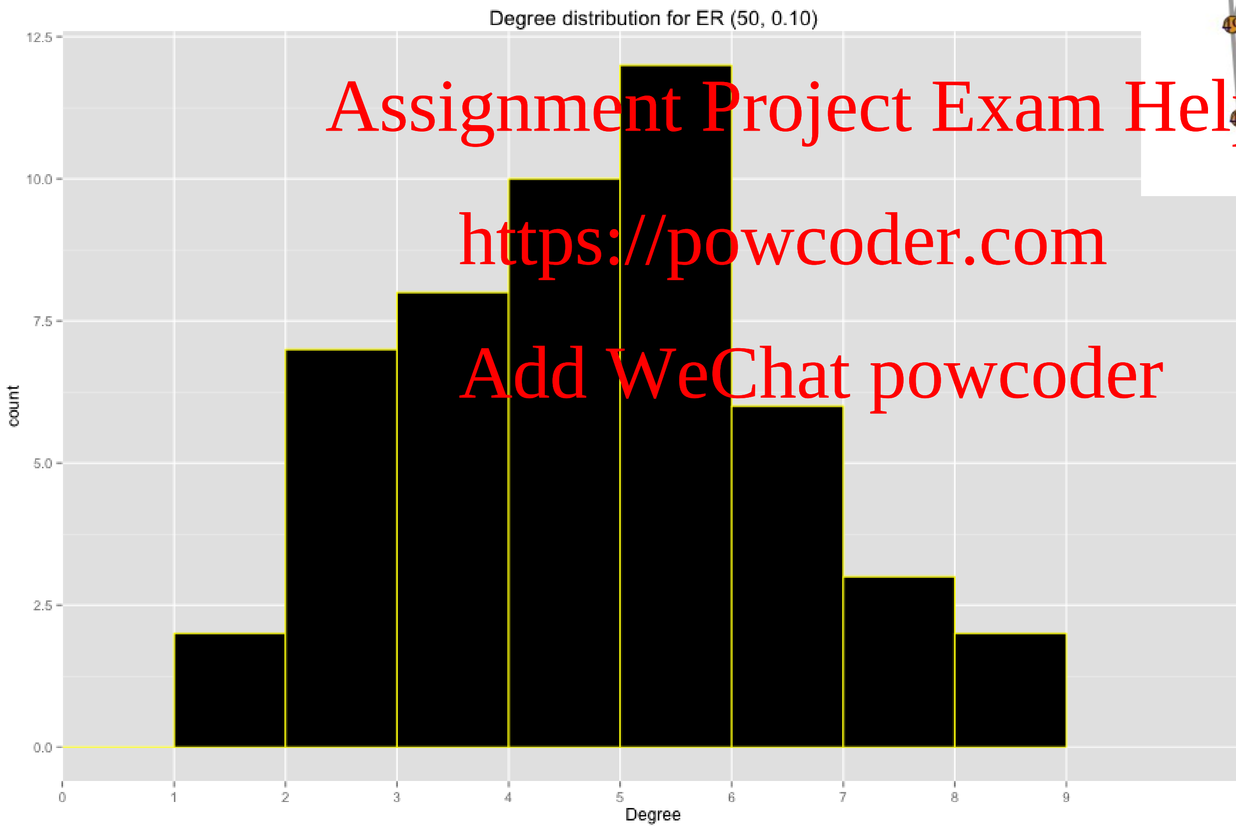
ER (50, 0.05)




Degree distribution for ER (50, 0.05)
count
Degree
0
5
10
15
0
1
2
3
4
5
6

ER (50, 0.1)



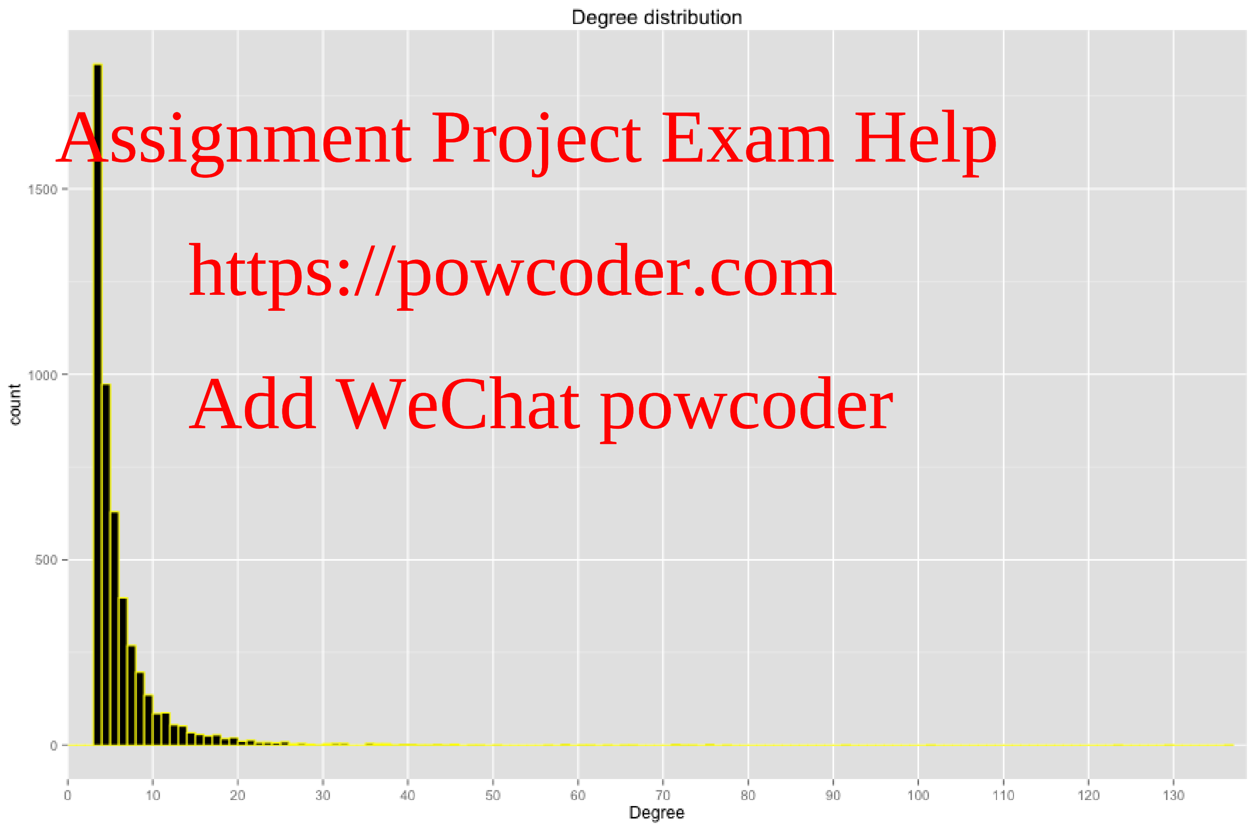


Degree distribution for ER (50, 0.10)
count
Degree
0.0
2.5
5.0
7.5
10.0
12.5
0
1
2
3
4
5
6
7
8
9

Prediction

- Networks will have a degree distribution centered around the average
  - few low degree nodes
  - few high degree nodes

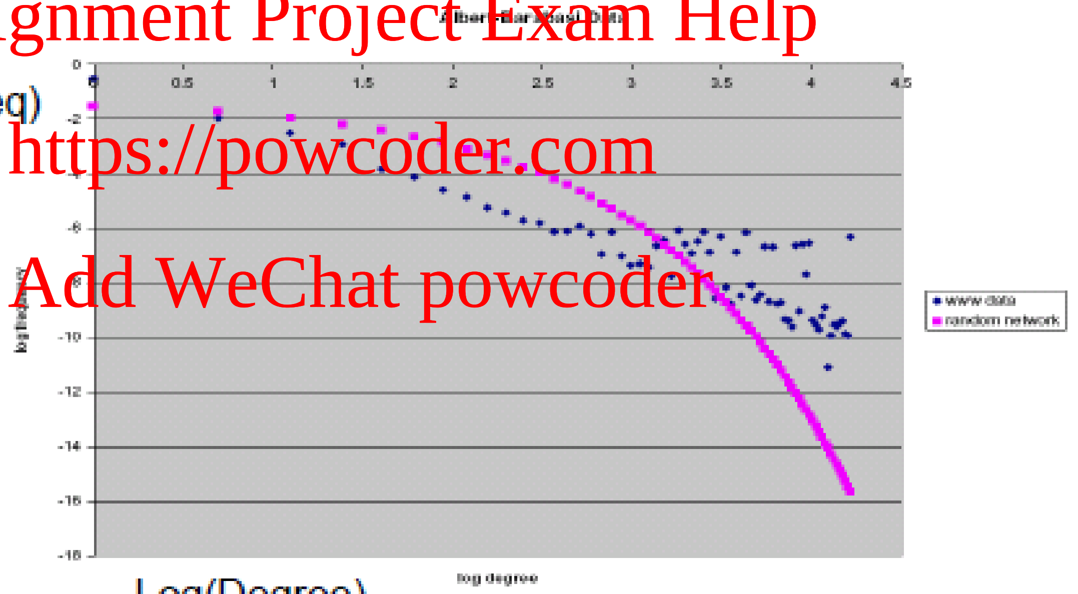
Doesn’t match



WWW degree distribution

- Albert, Jeong, Barabasi 1999

log(freq)



Log(Degree)

Not Poisson

- This finding is repeated in many other networks
  - bibliographic citation
  - email networks
  - online social networks
- We need another model

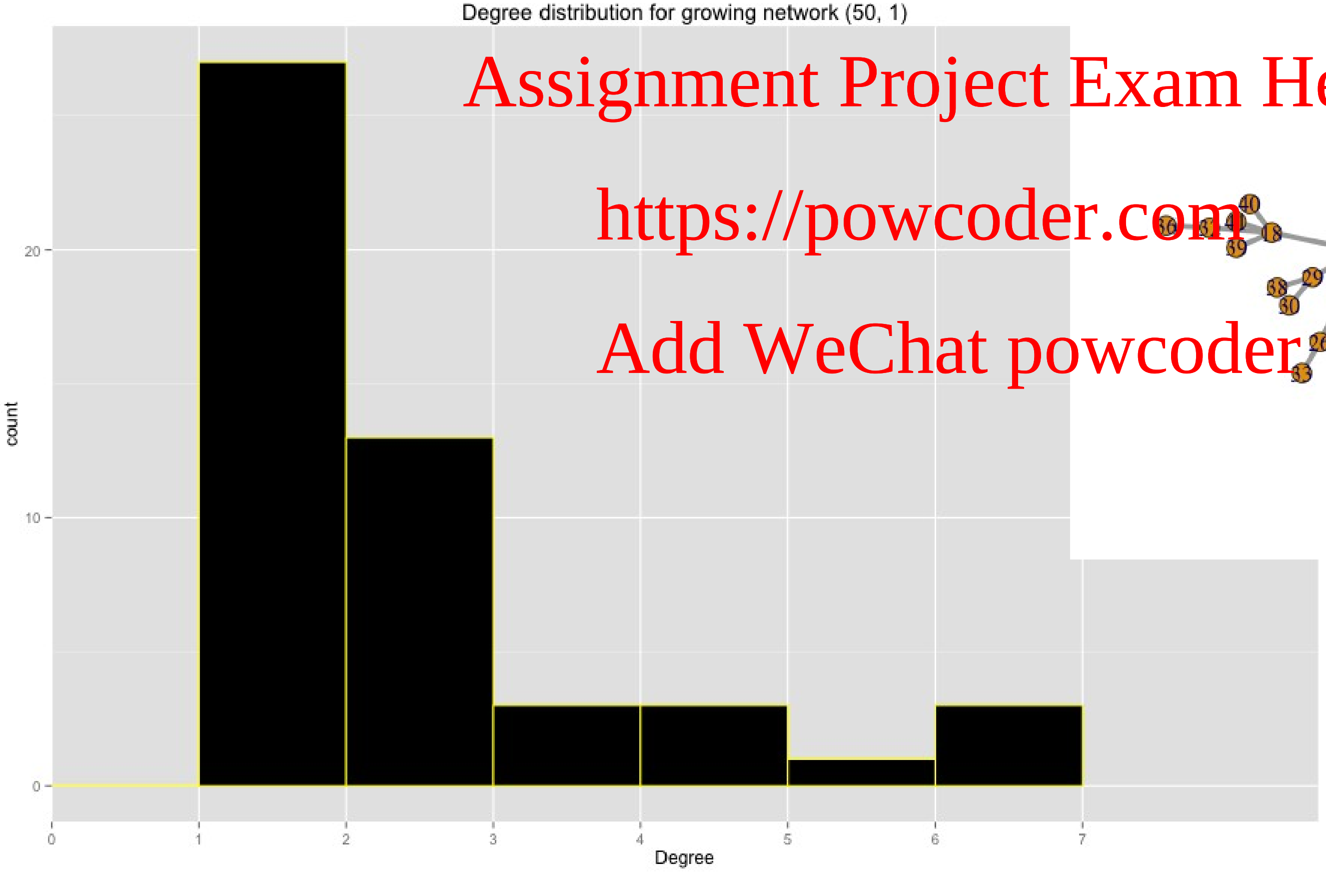
Growing random network

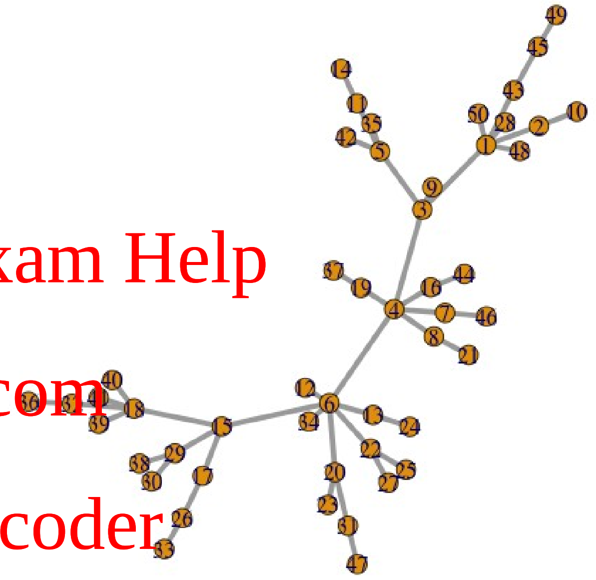
- Nodes arrive over time
- Natural heterogeneity
  - older nodes have more edges
- We can parameterize

Simple model

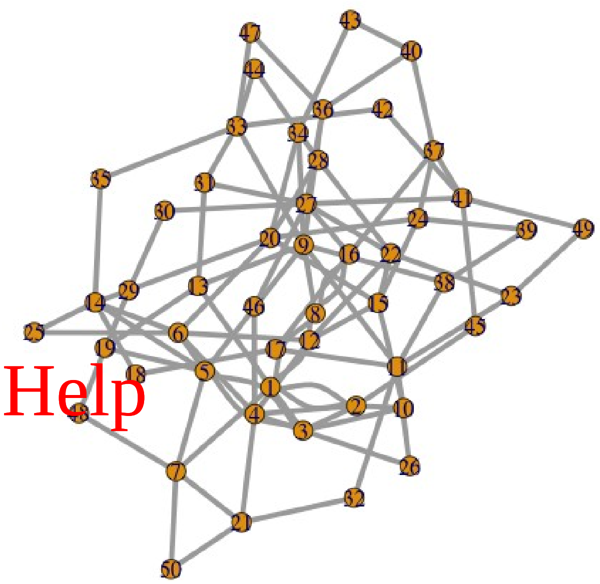
- One node is “born” at a time
- Forms m links to existing nodes
  - with equal probability
- Like ER
  - but we’re not dealing with all the nodes at once
- Start with an m node clique
  - (the math is easier!)
- In R
  - `sample_growing(n, m)`

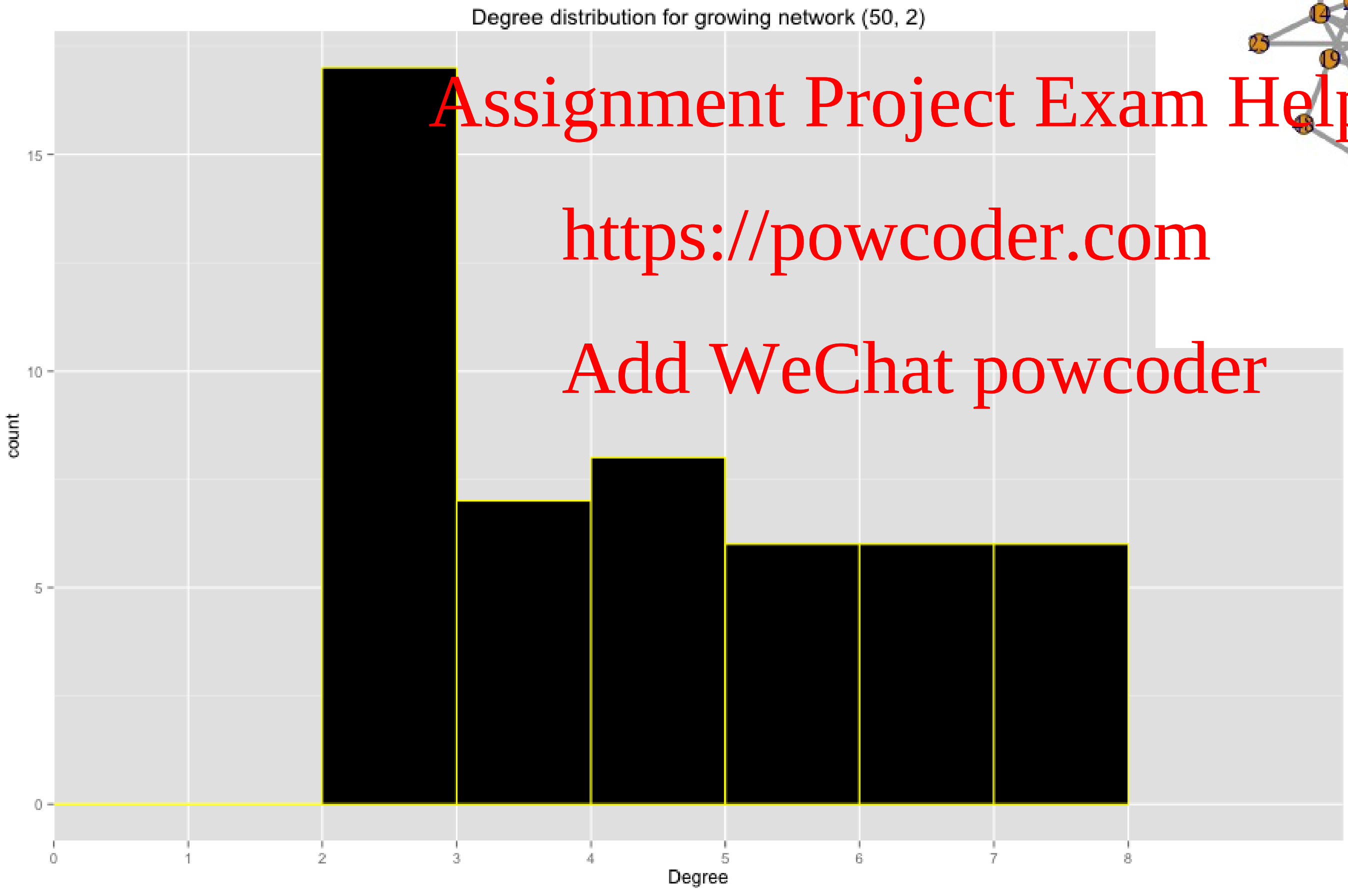
Growing (50, 1)


Degree distribution for growing network (50, 1)
count
20
10
0
0
1
2
3
4
5
6
7
Degree



Growing (50, 2)




Degree distribution for growing network (50, 2)
count
Degree
0
5
10
15
0
1
2
3
4
5
6
7
8

Degree distribution

- At time 0
  - first nodes n_0 each have degree m-1
- At time 1
  - node n_1 arrives makes m connections
  - n_0 have degree m
- At time 2
  - node n_2 arrives degree m
  - each other node has a m/m+1 chance of getting another edge
- At time k
  - node k arrives degree m
  - each other node has a m/(m+k) chance of getting an edge
- Probabilities vary over time

Degree distribution

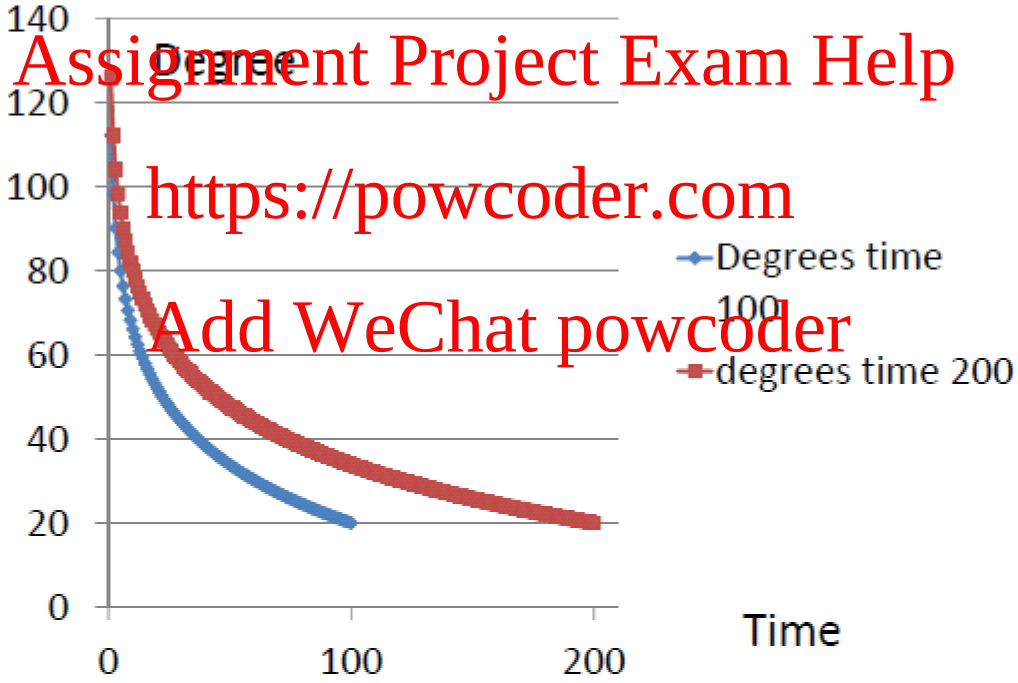
- Expected degree for node i born at $m < i < t$ is

  - $m + m/(i+1) + m/(i+2) + .. + m/t$

- Approximately

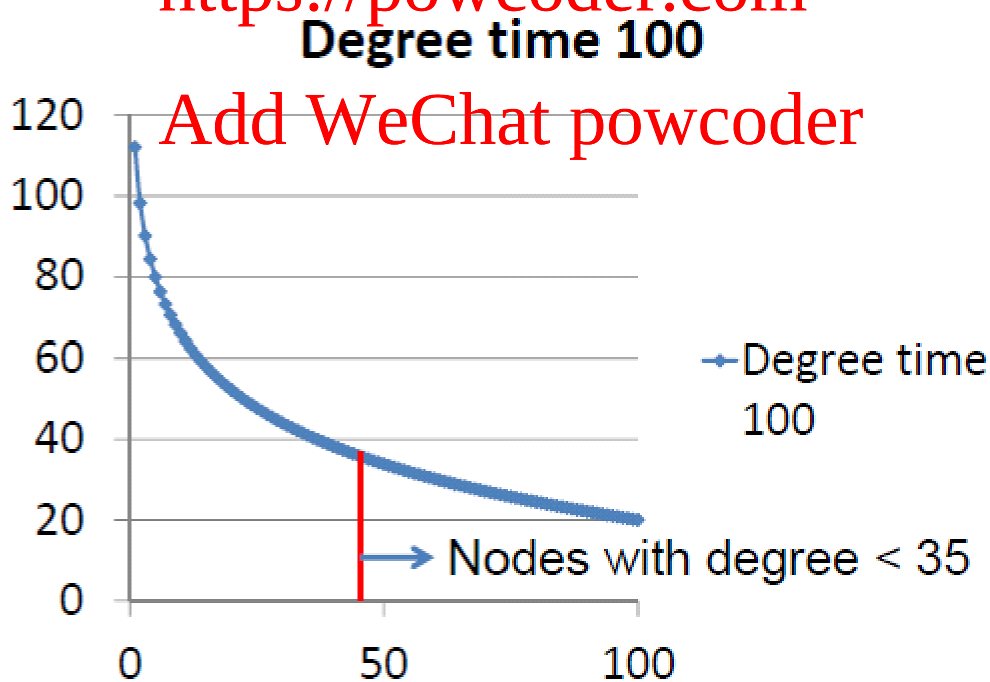
  - $m(1+\log(t/i))$ <- Harmonic numbers
- For any d
  - nodes that have degree less than d at time t
  - $m(1+\log(t/i)) < d$

Degree distribution by time



Degree time 100

- How many with degree < 35
- 20(1+log(100)/i) < 35
- $i > 100\ e^{-(35-20)/20} = 47.2$



Distribution

- Nodes with expected degree < d at t
  - Must be born recently
  - $i > t\, e^{-(d-m)/m}$
- We want the degree distribution at time t
  - how many nodes have $i > t\, e^{-(d-m)/m}$
  - $t - t\, e^{-(d-m)/m}$
  - divide by t to get a fraction
- $F_t(d) = 1 - e^{-(d-m)/m}$

Distribution of expected degree

- Not the same as the actual distribution
- Need to argue that for large t
  - we approximate the smooth curve
  - within some bounds
- See Jackson’s book

...ng random network has a skewed degree dist... because

The number of edges added is different at each time step

Older nodes have more chances to gain edges

A different probability distribution is used to compute edge probabilities

0%

Break