Social and Economic Networks

Static Random Network

Xiang Sun

September 27–October 11, 2017

Outline

1. Which networks form?
2. Static random network models
   - Poisson random networks
   - Small-world networks
   - Markov graphs/Exponential random graphs
   - Configuration model
   - An expected degree model
3. Properties of random networks
   - Large networks
   - Properties on limit
   - Threshold functions and phase transitions
   - Threshold for the existence of isolated points
   - Threshold function for connectivity
   - Threshold for giant component
   - Degree distribution of a neighboring node
   - Diameter estimation

Section 1

Which networks form?

Which networks form?

- Random network: How.
- Economic/game theoretic models: Why.

Section 2

Static random network models

Subsection 1

Poisson random networks

Poisson random networks

- Out of the all the possible networks on n nodes, one could simply pick one at random, with each network having an equal probability.
- In the $G(n, p)$ model, a graph is constructed by connecting nodes randomly. Each edge is included in the graph with probability p independent from every other edge.
- In the $G(n, M)$ model, a graph is chosen uniformly at random from the collection of all graphs which have n nodes and M edges.

The $G(n, p)$ model

- Every edge is formed with probability $p \in (0, 1)$ independently of every other edge.
- Let $I_{ij} \in \{0, 1\}$ be a Bernoulli random variable indicating the presence of edge $\{i, j\}$.
- For the Poisson random graph, random variables I_{ij} are independent and

$$I_{ij} = \begin{cases} 1, & \text{with probability } p, \\ 0, & \text{with probability } 1 - p. \end{cases}$$

The $G(n, M)$ model

- One could simply specify that the network should have M links, and then pick one of those networks at random with equal probability.
- That is, with each M-link network having probability $\binom{N}{M}^{-1}$, where $N = \binom{n}{2}$ is the number of potential links among n nodes.

The $G(n, p)$ model: Properties

- While these networks are static in the way that they are generated, much of the analysis of such random networks concerns what happens when n becomes large.
- $\mathbf{E}[\text{number of edges}] = \mathbf{E}[\sum I_{ij}] = \frac{n(n-1)}{2}p$.
- Using weak law of large numbers, we have for all $\alpha > 0$,

$$\lim_{n\to\infty} \mathbf{P}\left(\left|\sum I_{ij} - \frac{n(n-1)}{2}p\right| \geq \alpha\frac{n(n-1)}{2}\right) = 0.$$

$\Rightarrow$ The number of edges is random, but it is tightly concentrated around its mean for large n.

The $G(n, p)$ model: Degree distributions

- Let D be a random variable that represents the degree of a node.
- D is a binomial random variable with $\mathbf{E}[D] = (n-1)p$, *i.e.*,

$$\mathbf{P}(D = d) = \binom{n-1}{p} p^d (1-p)^{n-1-d}.$$

- Keeping the expected degree $(n-1)p$ constant as $n \to \infty$, D can be approximated with a Poisson random variable with $\lambda = (n-1)p$,

$$\mathbf{P}(D = d) = \frac{1}{d!}[(n-1)p]^d e^{-(n-1)p} = \frac{\lambda^d e^{-\lambda}}{d!}.$$

- This degree distribution falls off faster than an exponential in d, hence it is not a power-law distribution.

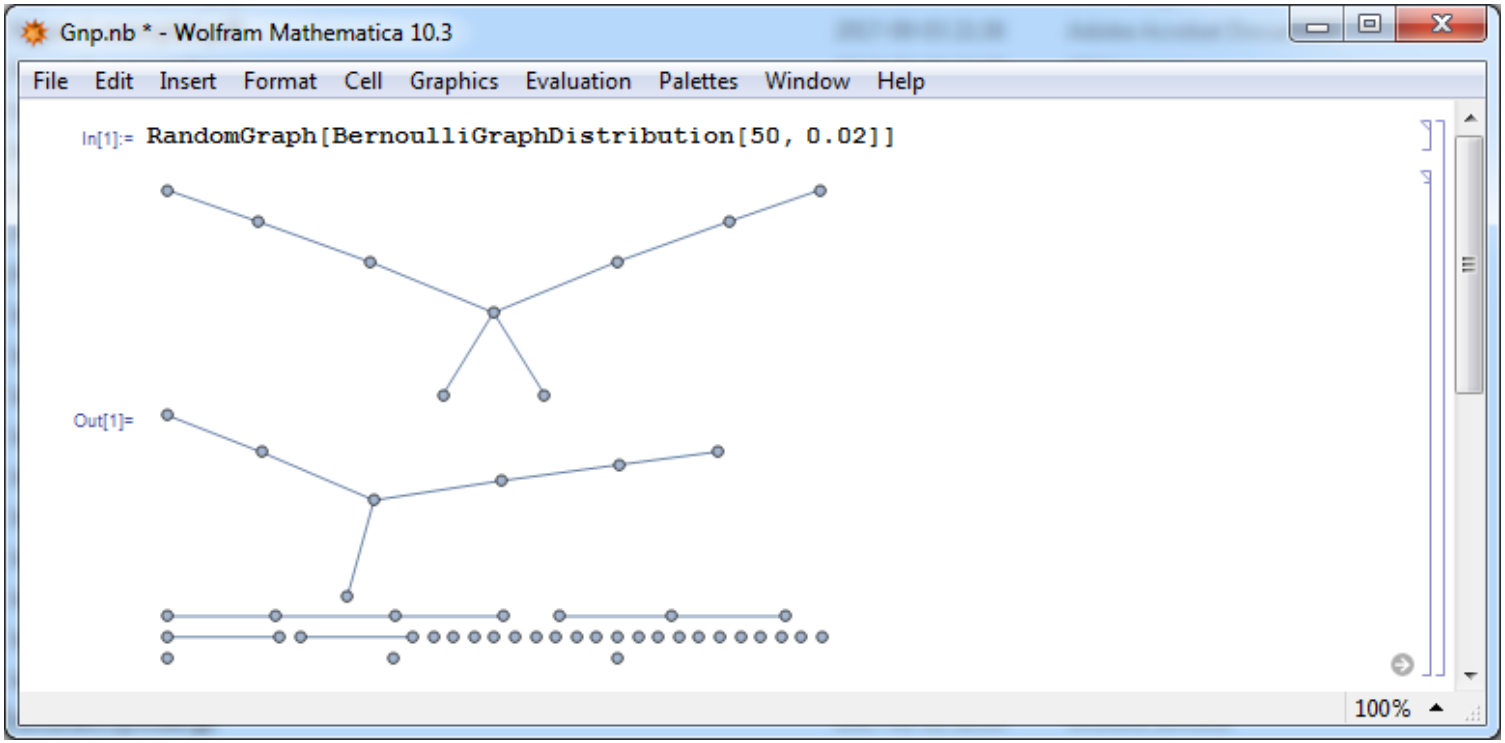
The $G(n, p)$ model: Clustering/diameter

- Individual clustering coefficient is p.
- Overall clustering coefficient is p.
- Clustering tends to 0, if max degree is bounded and network becomes large:

  If np is constant , then $p \to 0$ when $n \to \infty$.
- Diameter: small.

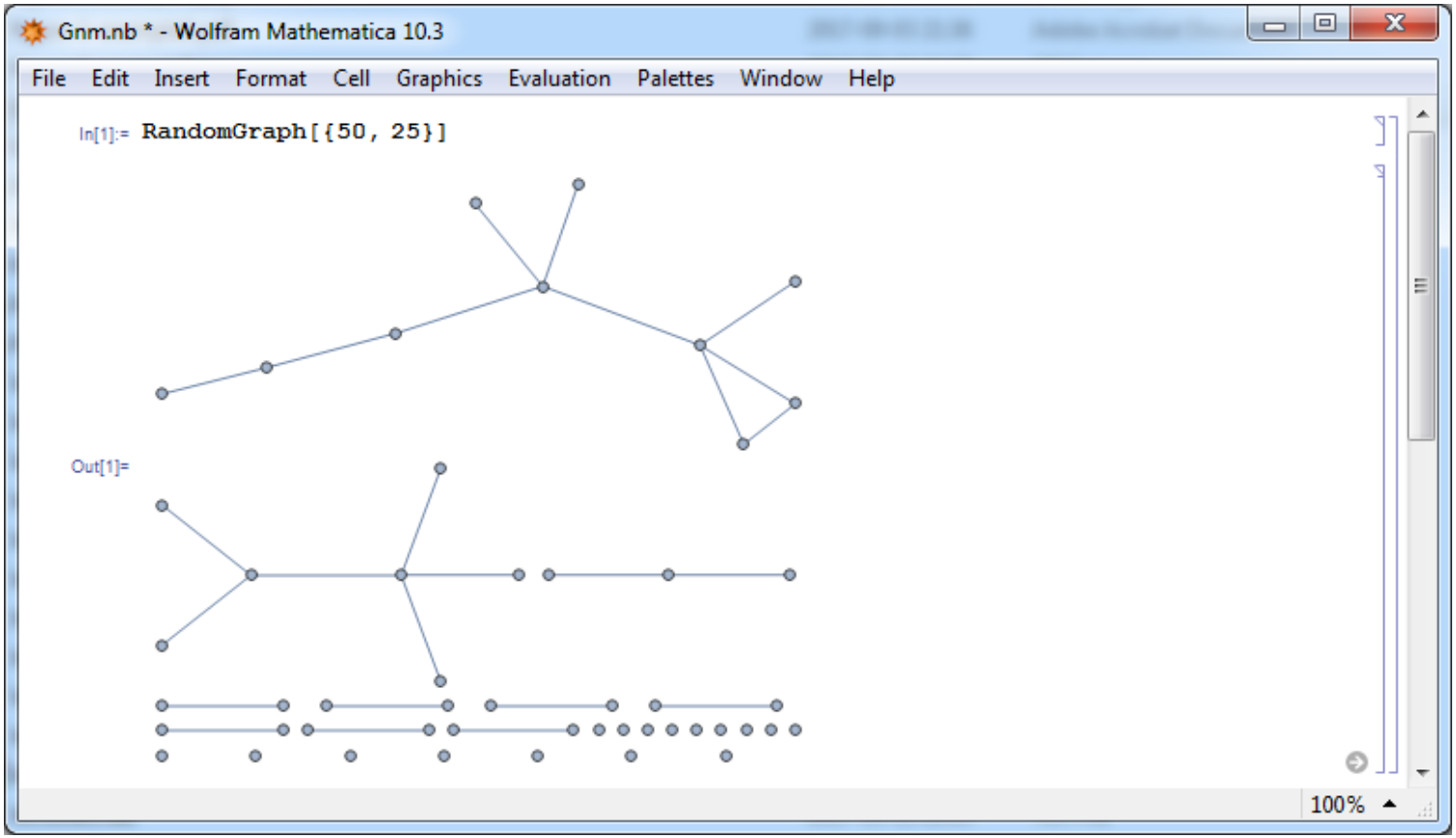
Poisson random graph: Algorithm

$G(n, p)$: `BernoulliGraphDistribution` in Mathematica.



Poisson random graph: Algorithm (Cont.)

$G(n, M)$: `RandomGraph` in Mathematica.



The $G(n, p)$ model: Issue

- While random graphs can exhibit some of the features of observed social networks, (*e.g.*, diameters that are small), it is clear that random graphs lack some of the features that are prevalent among social networks (such as the high clustering).
- Unrealistic in some sense.

Subsection 2

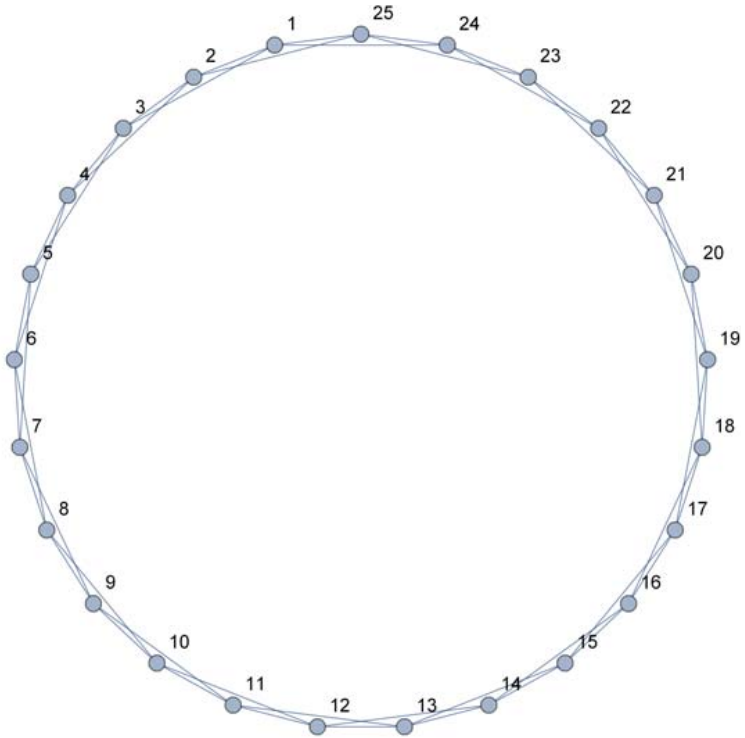
Small-world networks

Watts-Strogatz model

- To have a network with a small diameter and a high clustering, it is not convenience to use Poisson random graphs.
- The Watts-Strogatz model is a random graph generation model that produces graphs with small-world properties, including short average path lengths and high clustering.
- Watts-Strogatz model starts with a highly regular and clustered network (circle), and rewires a small number of edges to generate a small diameter.

Watts-Strogatz model: Illustration

We start with a very structured network that exhibits a high degree of clustering. For instance, let us construct a large circle, but then connect a given node to the nearest four nodes rather than just its nearest two neighbors.



Watts-Strogatz model: Illustration (Cont.)

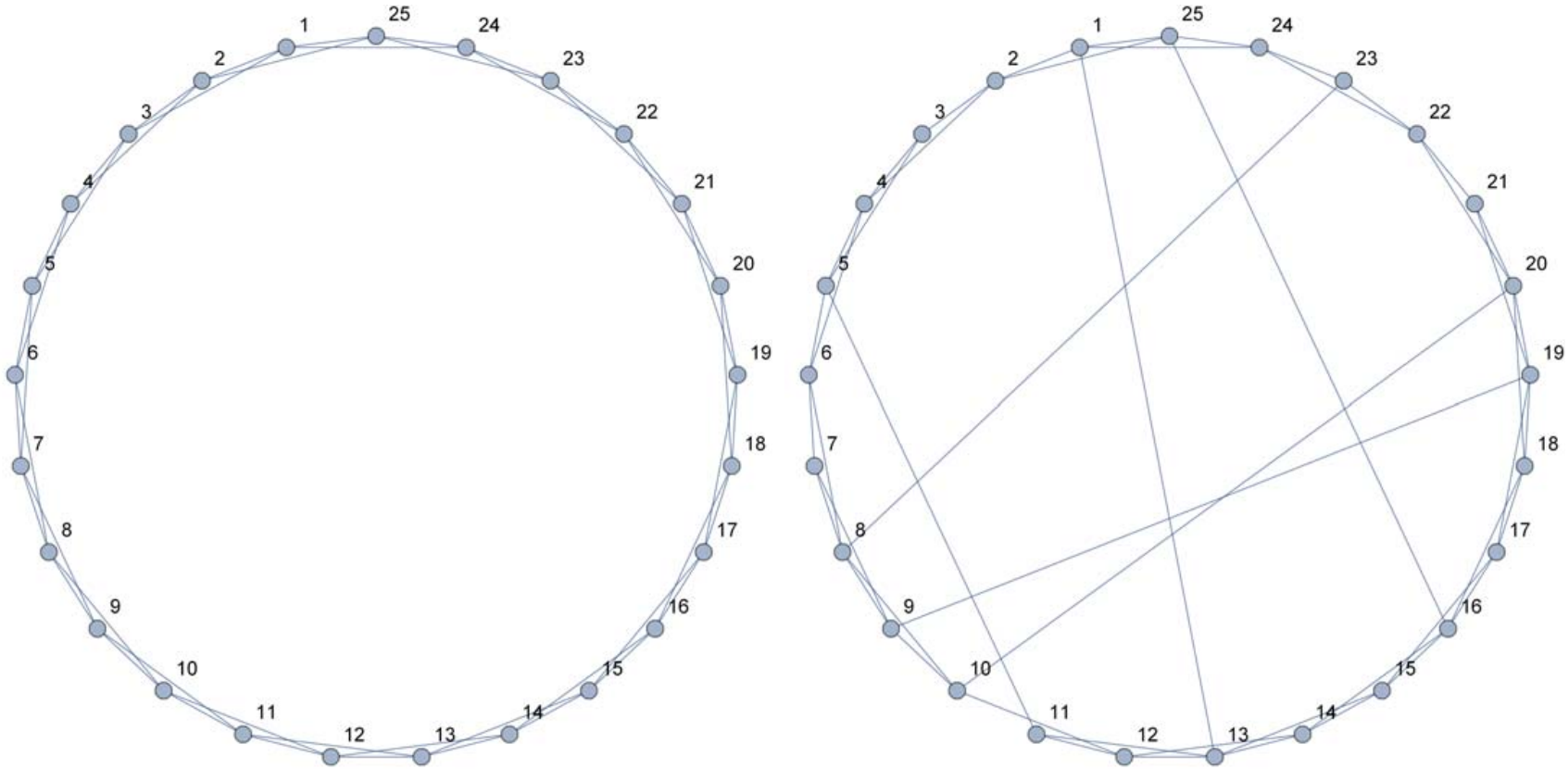
- Each node's individual clustering coefficient will be $\frac{1}{2}$.
- By adjusting the structure of the local connections we can also adjust the clustering.
- While this sort of regular network exhibits high clustering, it fails to exhibit some of the other features of many observed networks, such as a small diameter and at least some variance in the degree distribution.

  The diameter of such a network is on the order of $n = 4$.

Watts-Strogatz model: Illustration (Cont.)

- Randomly re-wiring relatively few links, we can end up with a network that has a much smaller diameter but still has substantial clustering.
- The re-wiring can be done by randomly selecting some link ij and disconnecting it and then randomly connecting i to another node k chosen uniformly at random from nodes which are not already neighbors of i.
- Of course, as more such re-wiring is done, the clustering will eventually vanish.
- The interesting region is where enough re-wiring has been done to substantially reduce (average and maximal) path length, but not so much that clustering vanishes.

Watts-Strogatz model: Illustration (Cont.)



Watts-Strogatz model: Illustration (Cont.)

- After having rewired just six links the diameter of the network has decreased from 6 (the left network) to 5 (the right network), with minimal impact on the clustering.
- There are 39 pairs of nodes at a distance of 6 from each other in the left network, which are all moved closer to each other by the rewiring.

Watts-Strogatz model: Algorithm

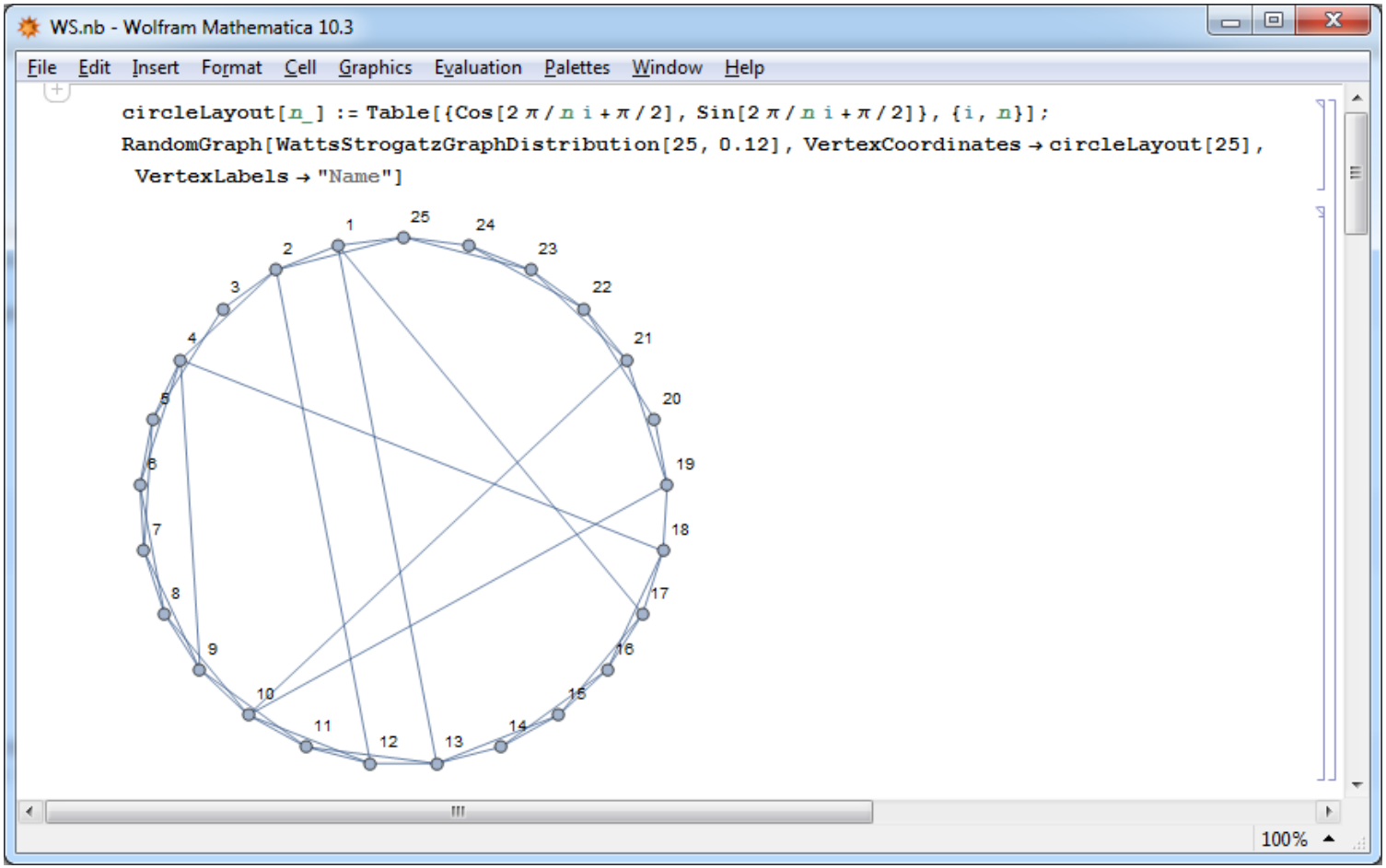
Given the desired number of nodes N, the mean degree $2K$, and a special parameter β, satisfying $0 \leq \beta \leq 1$ and $N \gg K \gg \ln N \gg 1$, the model constructs an undirected graph with N nodes and NK edges in the following way:

- Construct a regular ring lattice, a graph with N nodes each connected to $2K$ neighbors, K on each side.
- For every node $n_i = n_0, n_1, \ldots, n_{N-1}$, take every edge (n_i, n_j) with $i < j$, and rewire it with probability β.

  Rewiring is done by replacing (n_i, n_j) with (n_i, n_k) where k is chosen with uniform probability from all possible values that avoid self-loops ($k \neq i$) and link duplication (there is no edge $(n_i, n_{k'})$ with $k' = k$ at this point in the algorithm).

Watts-Strogatz model: Algorithm (Cont.)

`WattsStrogatzGraphDistribution` in Mathematica.



Watts-Strogatz model: Algorithm (Cont.)

- The underlying ring lattice structure of the model produces a locally clustered network, and the random links dramatically reduce the average path lengths.
- Varying β makes it possible to interpolate between a regular ring lattice ($\beta = 0$) and a random graph ($\beta = 1$) approaching $G(n, p)$ with $n = N$ and $p = \frac{NK}{\binom{N}{2}}$.

Watts-Strogatz model: Algorithm (Cont.)

Properties:

- The average path length falls very rapidly with increasing β.
- The clustering coefficient remains quite close to its value for the regular ring lattice, and only falls at relatively high β.
- The degree distribution in the case of the ring lattice is just a Dirac delta function centered at $2K$.

Subsection 3

Markov graphs/Exponential random graphs

Clustering

- Poisson random networks with average degrees growing more slowly than the number of nodes have clustering ratios tending to zero.

$\Rightarrow$ Too low to match many observed networks.

- Having dependencies in the model can produce nontrivial clustering.

Markov graphs/Exponential random graphs

- Frank and Strauss introduced the class of graphs "Markov graphs".
- Such random graph models were later imported to the social networks literature by Wasserman and Pattison under the name of p^* networks.
- The basic motivation is to provide a model that can be statistically analysed/estimated, and still allows for specific dependencies between the probabilities with which different links form.

Conditional dependencies

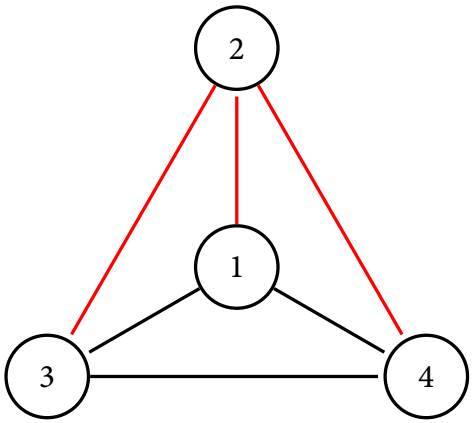
- Conditional dependencies can be introduced so that the probability of a link *ik* depends on whether *ij* and *jk* are present.
- The obvious challenge is that such dependencies will tend to interact with each other in ways that could make it impossible to specify the probability of different graphs in a tractable manner.
  - The conditional probability of a link *ik* depends on whether *ij* and *jk* are present, but also on any other adjacent pairs being present.
  - The conditional probability of *jk* depends on other adjacent pairs being present.
  - …
  - We end up with a complicated set of dependencies.

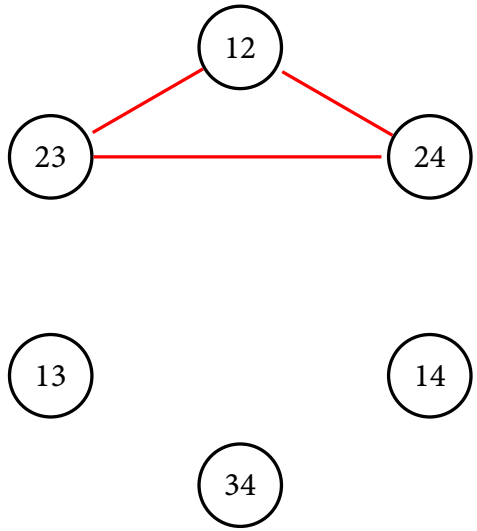
Dependence graph: Idea

- Consider n nodes.
- Keep track of the dependencies between links by another graph D (dependence graph), which is a graph among all of the $\binom{n}{2}$ possible links.

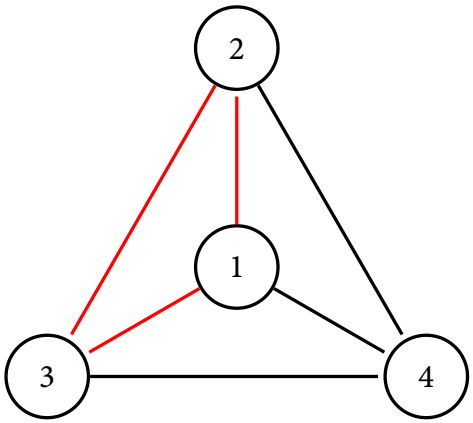
  D is not a graph on the original nodes, but a graph whose nodes are all the possible links.
- Idea: If ij and jk are neighbors in D, then there is some sort of conditional dependency between them, possibly in combination with other links.
- D captures which links are dependent on which others, possibly in quite complicated combinations.

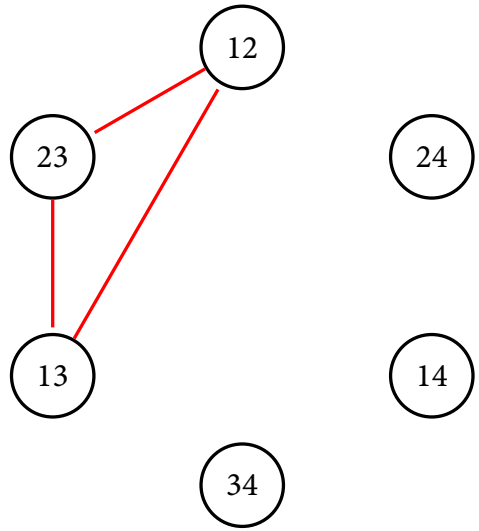
Dependence graph: Illustration





Dependence graph: Illustration (Cont.)





Dependence graph: Definition

- The original graph g with n nodes.
- Let M be the set of all the possible links in G.

$\Rightarrow$ $|M| = \binom{n}{2}$.

- M is served as the set of nodes for the dependence graph D.

Dependence graph: Illustration

- The Poisson random network is one where the set of links of D is empty, as all links are independent.
- If we wish to capture the idea that there might be clustering, then we would like the link ik to depend on the presence of ij and kj for each possible j. Thus, D would have ik connected to each other link that contains either i or k.

Dependence graph: Definition (Cont.)

- In the dependence graph D, a clique (派系) is a set of nodes of a completely connected subgraph of D.

  The singleton nodes are considered connected subgraphs.
- Let $C(D)$ be the set of all the cliques.
- In the case of a Poisson random network $C(D)$ would simply be the set of all links ij.

Dependence graph: Definition

- Given a generic element $A \in C(D)$, let $I_A(g)$

$$I_A(g) = \begin{cases} 1, & \text{if } A \subseteq g, \\ 0, & \text{otherwise.} \end{cases}$$

- If A is a triad $\{ij, jk, ik\}$, then $I_A(g) = 1$ if each of the links ij, jk and ik are in g, and $I_A(g) = 0$ otherwise.

Hammersley and Clifford's theorem

The probability of a given network g depends only on which cliques of D it contains, and that it can be written as

$$\log(\text{Prob}[g]) = \sum_{A \in C(D)} \alpha_A I_A(g) - c,$$

where c is a normalizing constant, and the α_A's are other free parameters.

Hammersley and Clifford's theorem: Illustration 1

- Let g be a Poisson random graph on n nodes.
- There are $\binom{n}{2}$ possible edges in g.
  These are the nodes of dependence graph D.
- $C(D) = \{ij \mid i, j \in g\}$
- Theorem implies that

$$\log(\text{Prob}[g]) = \sum_{ij \in C(D)} \alpha_{ij} - c.$$

Hammersley and Clifford's theorem: Illustration 1

- Theorem implies that

$$\log(\text{Prob}[g]) = \sum_{ij \in C(D)} \alpha_{ij} - c.$$

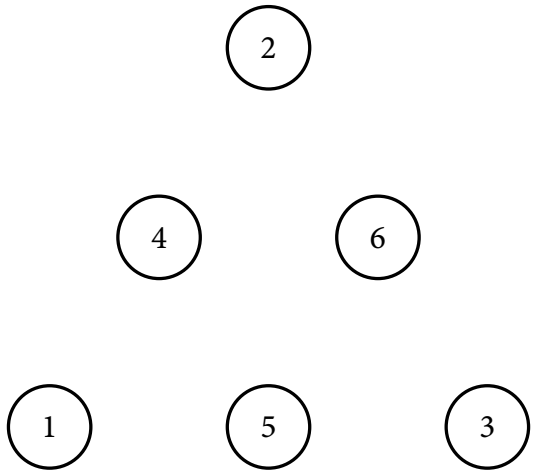
- To simplify things further, let us also suppose that there is a symmetry among nodes, so that the probability of any two networks that have the same architecture but possibly different labels on the nodes is identical.

$\Rightarrow$ $\log(\text{Prob}[g]) = \sum_{ij \in C(D)} \alpha - c = n_1(g)\alpha - c$, where $n_1(g)$ is the total number of links in g.

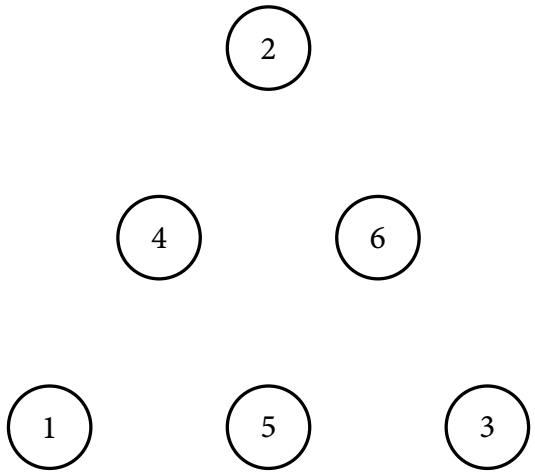
Hammersley and Clifford's theorem: Illustration 2

- Let g be a random graph on six nodes.
- There are 15 possible edges in g.

  These are the nodes of dependence graph D.

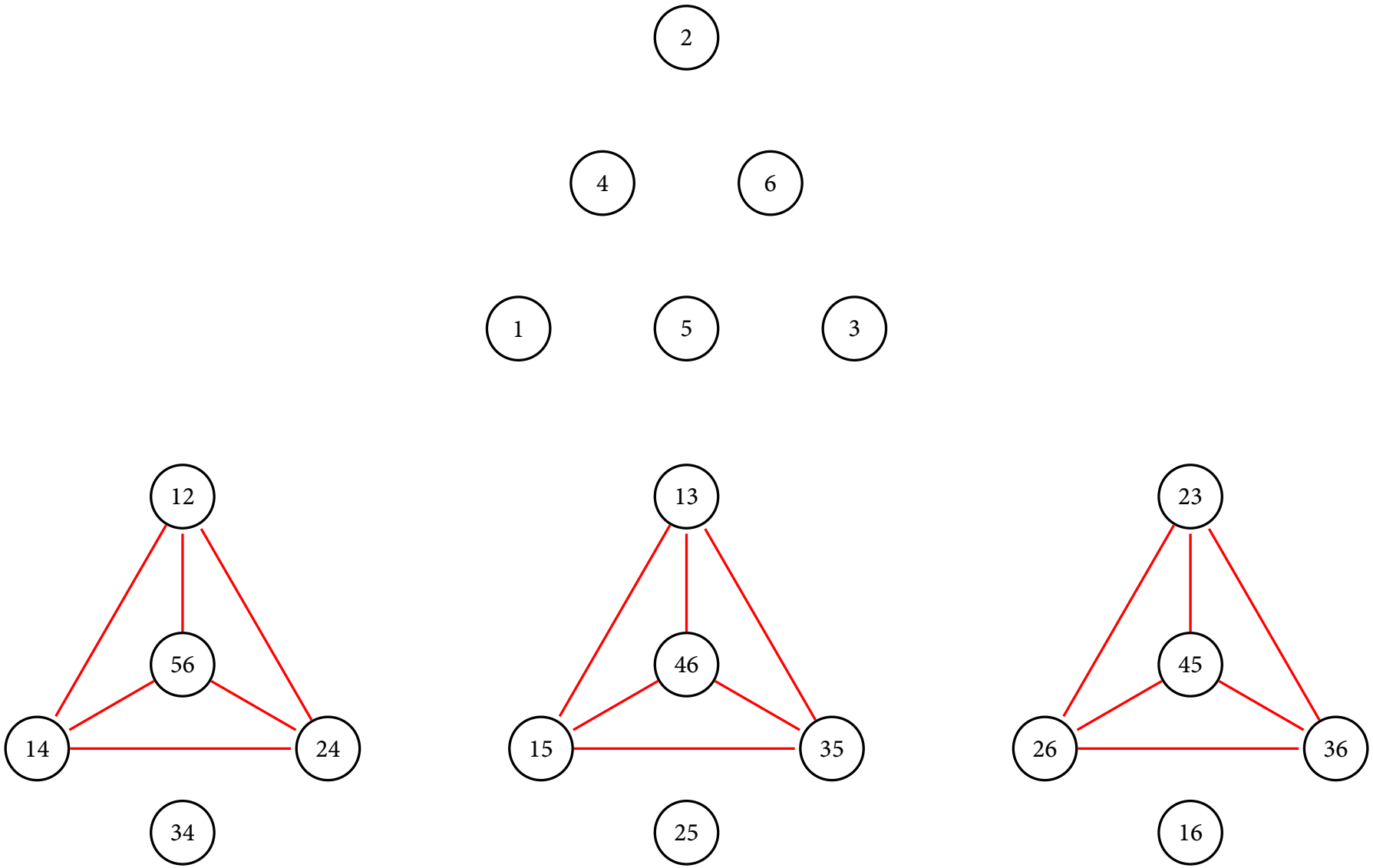


Hammersley and Clifford's theorem: Illustration 2

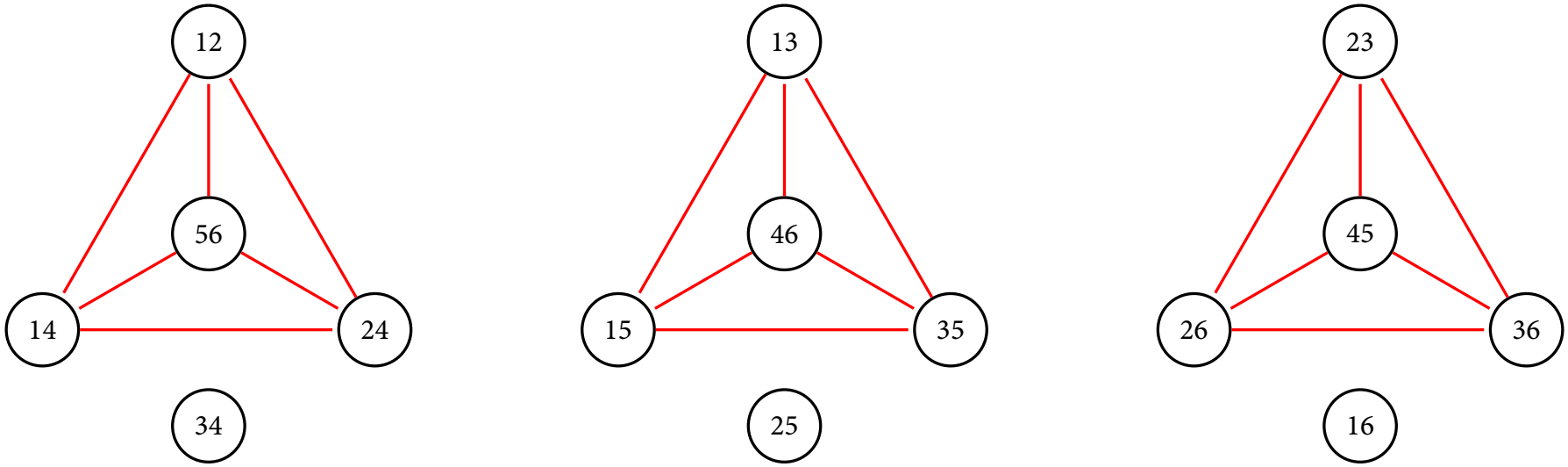


- Assumption of the dependence structure of g: two edges of g are conditional independent if they are not parallel.
- For example: 1-2 and 1-3 are independent, while 4-5 and 2-3 are dependent.

Hammersley and Clifford's theorem: Illustration 2



Hammersley and Clifford's theorem: Illustration 2



- There are six disjoint maximal cliques of D.
  Three of these have size 4 and three are singletons.
- Any nonempty subset of a maximal clique is a clique, so in total there are 48 cliques.
  15 of size 1, 18 of size 2, 12 of size 3, and 3 of size 4.
- Theorem yields a representation of Prob(g) with 48 parameters α_A.

Hammersley and Clifford's theorem: Comment

- In general, given that D can be very rich and that the α_A's can be chosen at will, this allows for an almost arbitrary probability specification.
- The difficulty and art in applying this type of model in practice is in specifying the dependencies sparingly and imposing restrictions on the α_A's so that the resulting probabilities are simple and practical.

Subsection 4

Configuration model

Configuration model

- While the Markov model of random networks allows for general forms of dependencies, it is hard to keep track of the degree distribution that it will generate, and to adjust that to match observed networks.
- In order to generate random networks with a given degree distribution, various methods have been proposed.
- One of the most widely used is what is referred to as the configuration model, as developed by Bender and Canfield.

Configuration model (Cont.)

- Work with degree sequences rather than degree distributions.
- Given a network on n nodes, we end up with a list of the degrees of different nodes: $(d_1, d_2, \ldots, d_n)$.
- The degree sequence directly tied to the degree distribution: the proportion of nodes that have degree d in this sequence is $P^n(d) = \frac{|\{i|d_i=d\}|}{n}$.

Configuration model: Algorithm

Suppose that we have an idea of the degree sequence $(d_1, d_2, \ldots, d_n)$ that we wish to generate in a network of n nodes.

1. Construct a sequence where node 1 is listed d_1 times, node 2 is listed d_2 times, *etc.*:

$$\underbrace{1, 1, \ldots, 1}_{d_1 \text{ entries}}, \underbrace{2, 2, \ldots, 2}_{d_2 \text{ entries}}, \ldots, \underbrace{n, n, \ldots, n}_{d_n \text{ entries}}.$$

2. Randomly pick two elements of the sequence.
3. Form a link between the two nodes corresponding to those entries.
4. Delete those entries from the sequence.
5. Repeat 2–4.

Configuration model: Issues

- It is possible to have more than one link between two nodes.

  As such, it generates what is called a multi-graph (allowing for multiple links) instead of a graph.
- Self links are possible and may even occur multiple times, while we have generally been ignoring self links in our discussion of networks up to this point.
- As a more minor point, the sum of the degrees needs to be even or else there will be a leftover entry at the end of the process.

Configuration model

- The process still has nice properties for large n.
- Generate a multi-graph, and then from it delete self-links and duplicate links between two nodes. This is then a graph.
- If the proportion of links we needed to delete is suitably small, then we end up with a graph with a degree distribution close to what we started with.
- Proposition: If $\frac{\max d_i}{n^{1/3}}$ tends to 0 ($n \to 0$), then the chance that any given node (including the largest ones) has a duplicate or self-link tends to 0.

Subsection 5

An expected degree model

An expected degree model

- Chung and Lu provide a different random model that also approximates a given desired degree sequence.
- The advantage of this process is that it forms a graph instead of a multi-graph, although it still allows for self loops and does not result in the exact degree sequence, even asymptotically.

Chung and Lu's process

- Start with n nodes and a desired degree sequence $(d_1, d_2, \ldots, d_n)$.
- Form a link between nodes i and j with probability $\frac{d_i d_j}{\sum_k d_k}$, where the degree sequence is such that that $(\max_i d_i)^2 < \sum_k d_k$, so that each of these probabilities is less than 1.
- It is clear that any node i's expected degree is indeed d_i, when a self-link ii is allowed to form with probability $d_i d_i / \sum_k d_k$.

Configuration model vs. the Chung-Lu' process

- Fix a degree sequence where all nodes have the same number of links $k = \langle d \rangle$.
- Consider the configuration model, where we delete self and duplicate links.
- The probability that any given node has no duplicate links or self links, and hence degree exactly k, converges to 1.
- From here it is not difficult to conclude that with a probability going to 1, the proportion of nodes with degree k will also converge to 1.

Configuration model vs. the Chung-Lu' process (Cont.)

- Fix a degree sequence where all nodes have the same number of links $k = \langle d \rangle$.
- The number of links to other nodes for any node follows a binomial distribution on $n - 1$ draws with a probability of $\frac{k}{n}$.
- As the probability of self links vanishes, the probability that the degree is the same as the number of links excluding self links approaches 1.

Configuration model vs. the Chung-Lu' process (Cont.)

- As n becomes large, a binomial distribution of $n-1$ draws with probability $\frac{k}{n}$ places a probability bounded away from 1 on having exactly k links.
- The probability of having exactly k links can be approximated from a Poisson approximation, and we find a probability on the order of

$$\frac{e^{-k}k^k}{k!},$$

  which is maximized at $k=1$ and always less than $\frac{1}{2}$.
- Under the Chung-Lu process, although the expected degree of any given node is k, the chance that it ends up with exactly k links is bounded away from 1, regardless of whether we allow self links.

Configuration model vs. the Chung-Lu' process (Cont.)

- This tells us that the realized degree distribution will differ significantly from the distribution of the expected degree sequence, which places full weight on degree k.
- While the configuration process (under suitable conditions) leads to a degree distribution more closely tied to the starting one, the Chung-Lu expected degree process is still of interest and more naturally relates to the Poisson random networks. Both are useful.

Section 3

Properties of random networks

Subsection 1

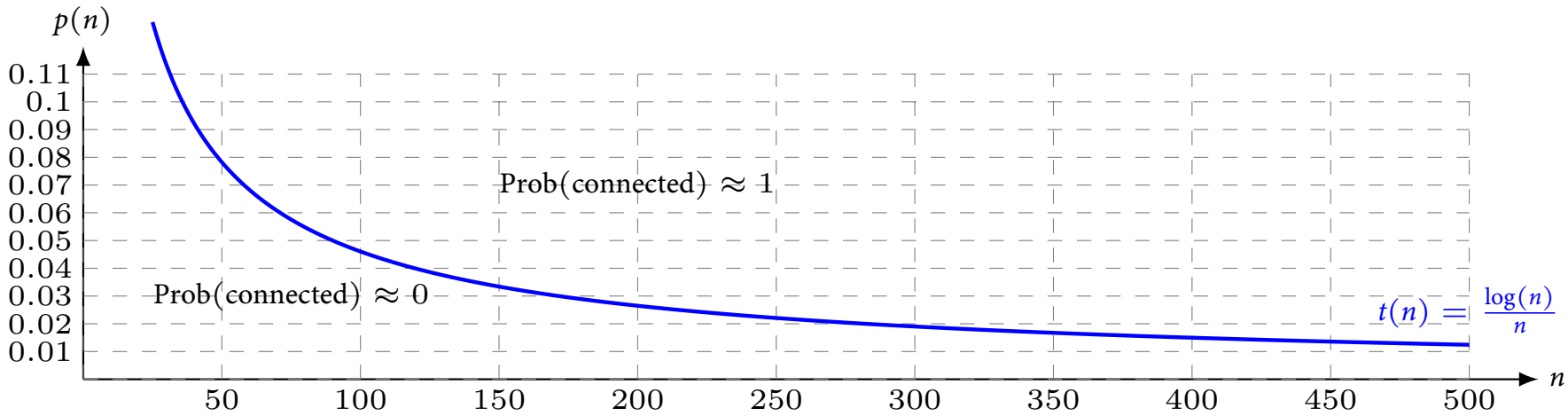
Large networks

Properties on limit

- Other questions of interest:
  - Does the graph have isolated nodes? cycles? Is it connected?
- For random graph models, we are interested in computing the probabilities of these events, which may be intractable for a fixed n.
  - Every network has some probability of forming.
  - How to make sense of that?
- Therefore, most of the time, we resort to an asymptotic analysis, where we compute (or bound) these probabilities as $n \to \infty$.

Phase transition

- Interestingly, often properties hold with either a probability approaching 1 or a probability approaching 0 in the limit.
- Consider a Possion random network with link formation probability $p(n)$ (again interest in $p(n) \to 0$ as $n \to \infty$).



- The graph experiences a phase transition (相变) as a function of graph parameters (also true for many other properties).

Subsection 2

Properties on limit

Specifying properties

- N: the set of nodes.
- $G(N)$: the set of all the undirected networks on the set of nodes N.
- A property is a subset $A(N) \subseteq G(N)$.
  - A specification of which networks have that property.

Examples of properties

- $A(N) = \{g \in G(N) \mid N_i(g) \text{ is nonempty for all } i \in N\}$.
  - property of no isolated nodes.
- $A(N) = \{g \in G(N) \mid \ell(i,j) \text{ is finite for all } i,j \in N\}$.
  - network is connected.
- $A(N) = \{g \in G(N) \mid \ell(i,j) < \log(n) \text{ for all } i,j \in N\}$.
  - diameter is less than $\log(n)$.

Monotone properties

- A property $A(\cdot)$ is monotone if

$$\left.\begin{array}{c} g \in A(N) \\ g \subseteq g' \in G(N) \end{array}\right\} \Longrightarrow g' \in A(N).$$

- All three of the previous properties are monotone.

Limiting properties

- In order to deduce things about random networks, we often look at "large" networks, by examining limits.
- Examine a sequence of Poisson random networks, with probability $p(n)$.
- Deduce things about properties as $n \to \infty$.

Subsection 3

Threshold functions and phase transitions

Threshold functions and phase transitions

- A threshold function for some given (monotone) property $A(\cdot)$ is a function $t(n)$ such that

$$\text{Prob}[A(N)] \to \begin{cases} 1, & \text{if } \frac{p(n)}{t(n)} \to \infty, \\ 0, & \text{if } \frac{p(n)}{t(n)} \to 0, \end{cases}$$

  where $n = |N|$.
- When such a threshold function $t(n)$ exists, it is said that a phase transition occurs at that threshold.

Phase transition: Example

- Define property A as $A = \{\text{number of edges} > 0\}$.
- We are looking for a threshold for the emergence of the first edge.
- Recall that

$$\mathbf{E}[\text{number of edges}] = \binom{n}{2} p(n) = \frac{n(n-1)}{2} p(n) \approx \frac{n^2}{2} p(n).$$

- Let $\hat{n} = \binom{n}{2}$ and $\lambda(n) = \hat{n} p(n) \approx \frac{n^2}{2} p(n)$.

Phase transition: Example

- Assume that $\lim_{n\to\infty} \frac{p(n)}{2/n^2} = 0$.
- Then $\mathbf{E}[\text{number of edges}] \to 0$.
- Thus, $\text{Prob}(\text{number of edges} > 0) \to 0$.

Phase transition: Example

- Assume that $\lim_{n\to\infty} \frac{p(n)}{2/n^2} = \infty$.
- The number of edges can be approximated by a Poisson distribution (just like the degree distribution):

$$\text{Prob}(\text{number of edges} = k) = \binom{\hat{n}}{k} p^k (1-p)^{\hat{n}-k} \approx \frac{e^{-\lambda(n)}\lambda^k}{k!}.$$

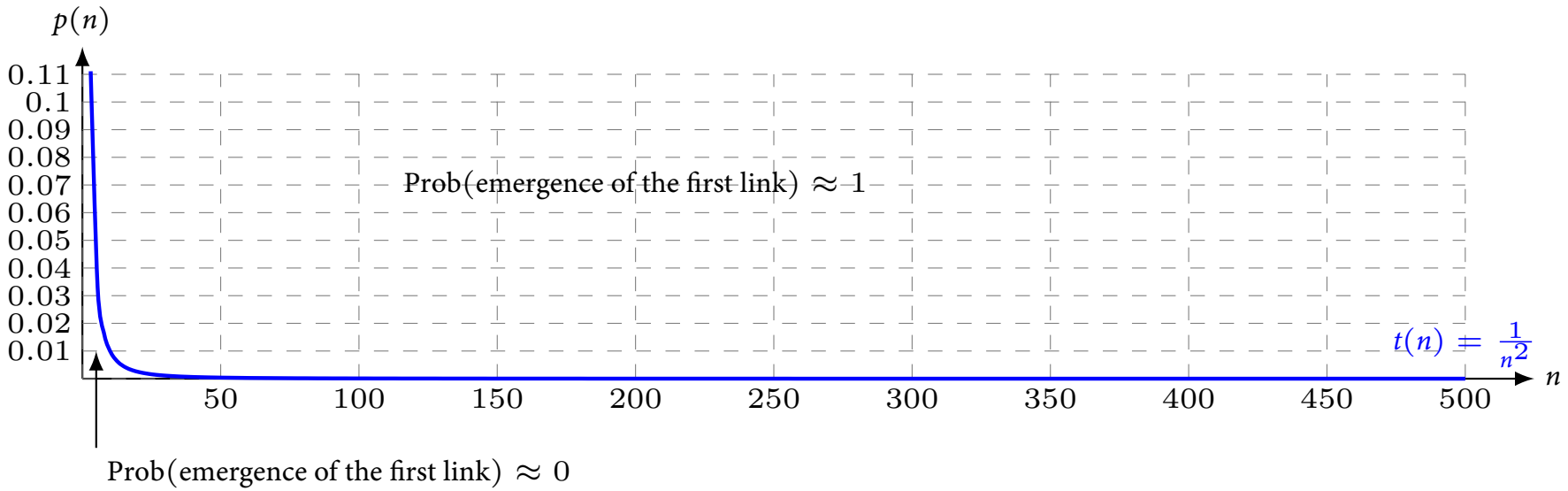
- Thus, $\text{Prob}(\text{number of edges} = 0) \approx e^{-\lambda(n)}$.
- Since $\lim_{n\to\infty} \lambda(n) = \lim_{n\to\infty} \frac{p(n)}{2/n^2} = \infty$, we have

$$\lim_{n\to\infty} \text{Prob}(\text{number of edges} = 0) \approx \lim_{n\to\infty} e^{-\lambda(n)} = 0.$$

Phase transition: Example

- Hence, the function $t(n) = \frac{1}{n^2}$ is a threshold function for the emergence of the first link:
  - When $p(n) \ll \frac{1}{n^2}$, the network is likely to have no edges in the limit;
  - When $p(n) \gg \frac{1}{n^2}$, the network has at least one edge with probability going to 1.

Phase transition: Example



Phase transition: Example

- How large should $p(n)$ be to start observing triples in the network?
  - We have $\mathbf{E}[\text{number of triples}] = n^3 p^2$, using a similar analysis we can show that $t(n) = \frac{1}{n^{3/2}}$ is a threshold function.
- How large should $p(n)$ be to start observing a tree with k nodes (and $k-1$ arcs)?
  - We have $\mathbf{E}[\text{number of trees}] = n^k p^{k-1}$, and the function $t(n) = \frac{1}{n^{k/(k-1)}}$ is a threshold function.

Phase transition: Example

- The threshold function for observing a cycle with k nodes is $t(n) = \frac{1}{n}$.
- Below the threshold of $\frac{1}{n}$, largest component of the graph includes no more than a factor times $\log(n)$ of the nodes.
- Above the threshold of $\frac{1}{n}$, a giant component emerges, which is the largest component that contains a nontrivial fraction of all nodes, i.e., at least cn for some constant c.
- The giant component grows in size until the threshold of $\frac{\log(n)}{n}$, at which point the network becomes connected.

Phase transition: Example

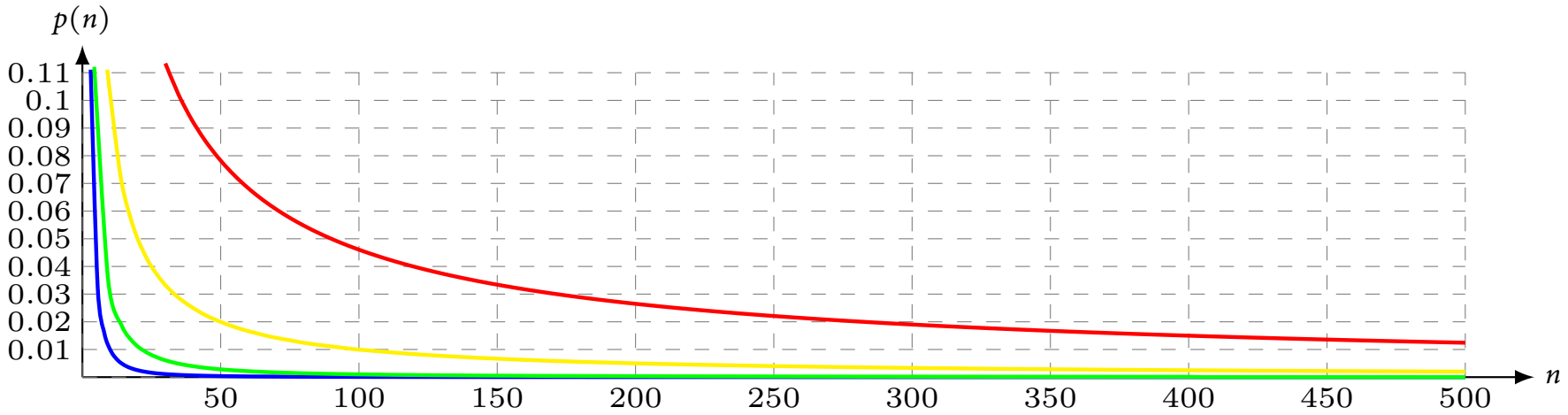


Figure: Blue: $\frac{1}{n^2}$; Green: $\frac{1}{n^{3/2}}$; Yellow: $\frac{1}{n}$; Red: $\frac{\log(n)}{n}$

Phase transition: Illustration

n	$\frac{1}{n^2}$	$\frac{1}{n^{3/2}}$	$\frac{1}{n}$	$\frac{\log(n)}{n}$
	first link	triples	cycle	connected
50	0.0004	0.0028	0.02	0.0782

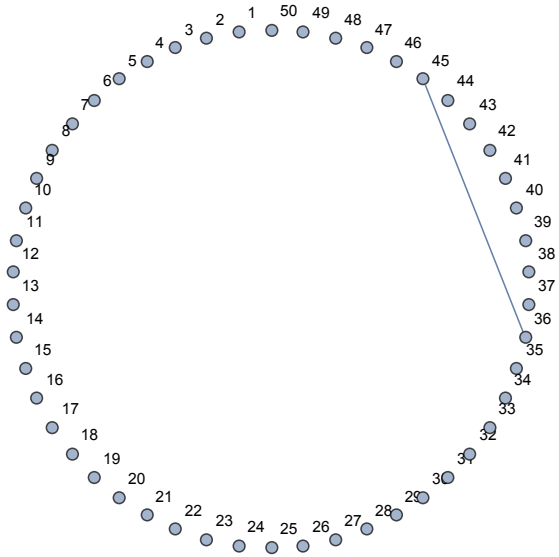


Figure: The emergence of the first link: $G(50, p = 0.001)$.

Phase transition: Illustration

n	$\frac{1}{n^2}$	$\frac{1}{n^{3/2}}$	$\frac{1}{n}$	$\frac{\log(n)}{n}$
	first link	triples	cycle	connected
50	0.0004	0.0028	0.02	0.0782

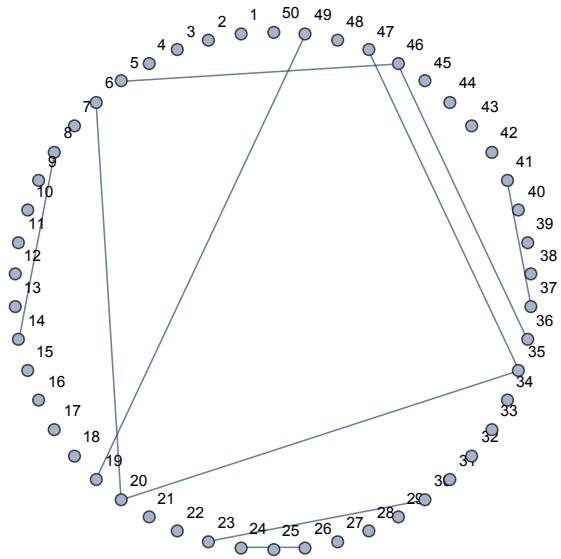


Figure: A first component with more than two nodes: $G(50, p = 0.01)$.

Phase transition: Illustration

n	$\frac{1}{n^2}$	$\frac{1}{n^{3/2}}$	$\frac{1}{n}$	$\frac{\log(n)}{n}$
	first link	triples	cycle	connected
50	0.0004	0.0028	0.02	0.0782

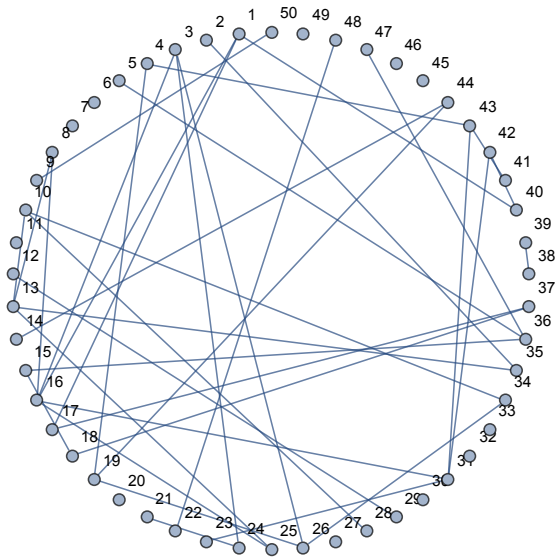


Figure: Emergence of cycles: $G(50, p = 0.03)$.

Phase transition: Illustration

n	$\frac{1}{n^2}$	$\frac{1}{n^{3/2}}$	$\frac{1}{n}$	$\frac{\log(n)}{n}$
	first link	triples	cycle	connected
50	0.0004	0.0028	0.02	0.0782

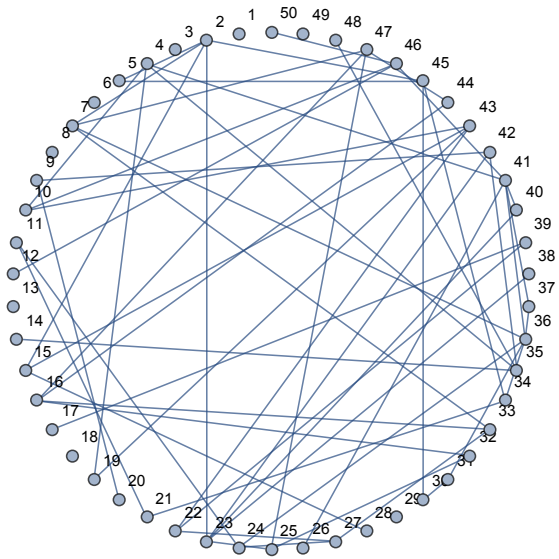


Figure: Emergence of a giant component: $G(50, p = 0.05)$.

Phase transition: Illustration

n	$\frac{1}{n^2}$	$\frac{1}{n^{3/2}}$	$\frac{1}{n}$	$\frac{\log(n)}{n}$
	first link	triples	cycle	connected
50	0.0004	0.0028	0.02	0.0782

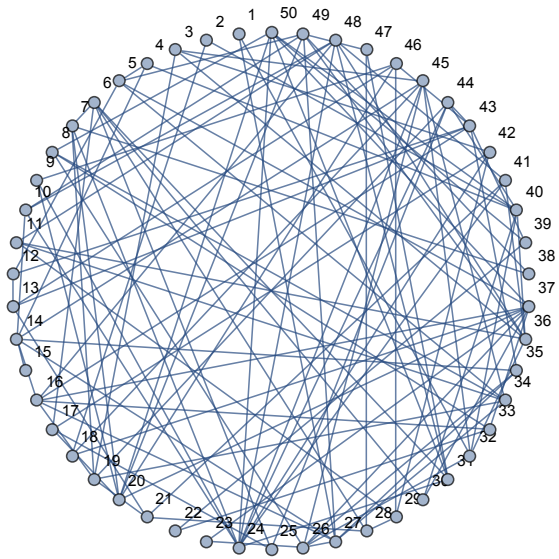


Figure: Emergence of connectedness: $G(50, p = 0.1)$.

Subsection 4

Threshold for the existence of isolated points

Threshold for the existence of isolated points

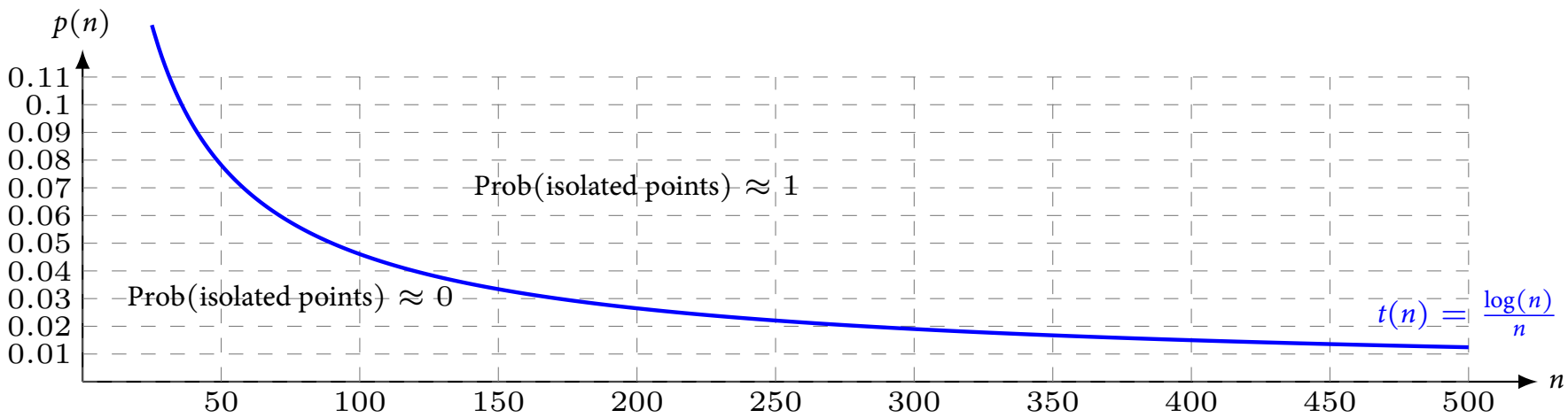
- Theorem: The threshold for the (non)existence of isolated vertices is $t(n) = \frac{\log n}{n}$.
- It is sufficient to show that

$$\text{Prob}(\text{nonexistence of isolated points}) \to \begin{cases} 0, & \text{if } \lambda(n) = \frac{p(n)}{t(n)} \to 0, \\ 1, & \text{if } \lambda(n) = \frac{p(n)}{t(n)} \to \infty. \end{cases}$$

- We will show a stronger result: Let $p(n) = \lambda \frac{\log(n)}{n}$.

$$\text{Prob}(\text{nonexistence of isolated points}) \to \begin{cases} 0, & \text{if } \lambda < 1, \\ 1, & \text{if } \lambda > 1. \end{cases}$$

Threshold for the existence of isolated points (Cont.)



Proof

- Write $p = p(n) = \lambda \frac{\log(n)}{n}$. Clearly, $p \to 0$.
- Let I_i be a Bernoulli random variable defined as

$$I_i = \begin{cases} 1, & \text{if node } i \text{ is isolated,} \\ 0, & \text{otherwise.} \end{cases}$$

- We can write the probability that an individual node is isolated as

$$q = \text{Prob}(I_i = 1) = \left(1 - p\right)^{n-1} \approx e^{-pn} = e^{-\lambda \log(n)} = n^{-\lambda},$$

where the third equality is due to $p \to 0$.

This also implies that $q \to 0$.

Proof

- Let $X = \sum_{i=1}^{n} I_i$ denote the total number of isolated nodes.
- It suffices to show that

$$\text{Prob}(X = 0) \to \begin{cases} 0, & \text{if } \lambda < 1, \\ 1, & \text{if } \lambda > 1. \end{cases}$$

- We have

$$\mathrm{E}[X] = n \times q \approx n \times n^{-\lambda} = n^{1-\lambda}.$$

Proof: The case $\lambda < 1$

- Assume that $\lambda < 1$.
- For $\lambda < 1$, we have $\mathbf{E}[X] \approx n^{1-\lambda} \to \infty$. We want to show that this implies $\text{Prob}(X = 0) \to 0$.
  - In general, this is not true.
    For example, $\text{Prob}(X = 0) = \frac{1}{n^2}$ and $\text{Prob}(X = n) = 1 - \frac{1}{n^2}$.
  - Can we use a Poisson approximation (as in the previous example)?
    No, since the random variables here are dependent.
  - We show that the variance of X is of the same order as its mean.

Proof: The case $\lambda < 1$ (Cont.)

- We compute the variance of X:

$$\begin{aligned}\operatorname{var}(X) &= \sum_i \operatorname{var}(I_i) + \sum_i \sum_{j \neq i} \operatorname{cov}(I_i, I_j) \\ &= n \operatorname{var}(I_1) + n(n-1) \operatorname{cov}(I_1, I_2) \\ &= nq(1-q) + n(n-1)(\mathbf{E}[I_1 I_2] - \mathbf{E}[I_1]\mathbf{E}[I_2]),\end{aligned}$$

  where the second and third equalities follow since the I_i are identically distributed Bernoulli random variables with parameter q (dependent).
- We have

$$\begin{aligned}\mathbf{E}[I_1 I_2] &= \text{Prob}(I_1 = I_2 = 1) = \text{Prob}(\text{both 1 and 2 are isolated}) \\ &= (1-p)^{n-2}(1-p)^{n-2}(1-p) = \frac{q^2}{1-p}.\end{aligned}$$

Proof: The case $\lambda < 1$ (Cont.)

- Combining the preceding two relations, we obtain

$$\begin{aligned}\mathrm{var}(X) &= nq(1-q) + n(n-1)\left[\frac{q^2}{1-p} - q^2\right] \\ &= nq(1-q) + n(n-1)\frac{q^2 p}{1-p}\end{aligned}$$

- Since $p \to 0$ and $q \to 0$, we have

$$\begin{aligned}\mathrm{var}(X) &\sim nq + n^2 q^2 \frac{p}{1-p} \sim nq + n^2 q^2 p \\ &= nn^{-\lambda} + n^2 n^{-2\lambda} \frac{\lambda \log(n)}{n} \\ &\sim nn^{-\lambda} = \mathbf{E}[X],\end{aligned}$$

where $a(n) \sim b(n)$ denotes $\frac{a(n)}{b(n)} \to 1$ as $n \to \infty$.

Proof: The case $\lambda < 1$ (Cont.)

- Second moment method: Let X be a non-negative integer valued random variable. Then

$$\text{Prob}(X = 0)(\mathbf{E}[X])^2 \leq \text{var}(X).$$

- By the second moment method, we have

$$\text{Prob}(X = 0)(\mathbf{E}[X])^2 \leq \text{var}(X) \sim \mathbf{E}[X].$$

- Therefore,

$$\text{Prob}(X = 0) \leq \kappa \frac{\mathbf{E}[X]}{(\mathbf{E}[X])^2} = \kappa \frac{1}{\mathbf{E}[X]} \to 0.$$

Proof: The case $\lambda > 1$

- Assume that $\lambda > 1$.
  We want to show that $\text{Prob}(X = 0) \to 1$.
- We have $\mathbf{E}[X] = n(1-p)^{n-1}$.
- Then,

$$\mathbf{E}[X] = n(1-p)^{n-1} \leq ne^{-p(n-1)} = o(ne^{-\log(n)}) = o(1),$$

  where the second inequality is due to $e^x \geq x + 1$ for any x.
- First moment method: Let X be a non-negative random variable. Then

$$\text{Prob}(X > 0) \leq \mathbf{E}[X].$$

- By the first moment method, we have

$$\text{Prob}(X > 0) = o(1) \Rightarrow \text{Prob}(X = 0) \to 1.$$

Subsection 5

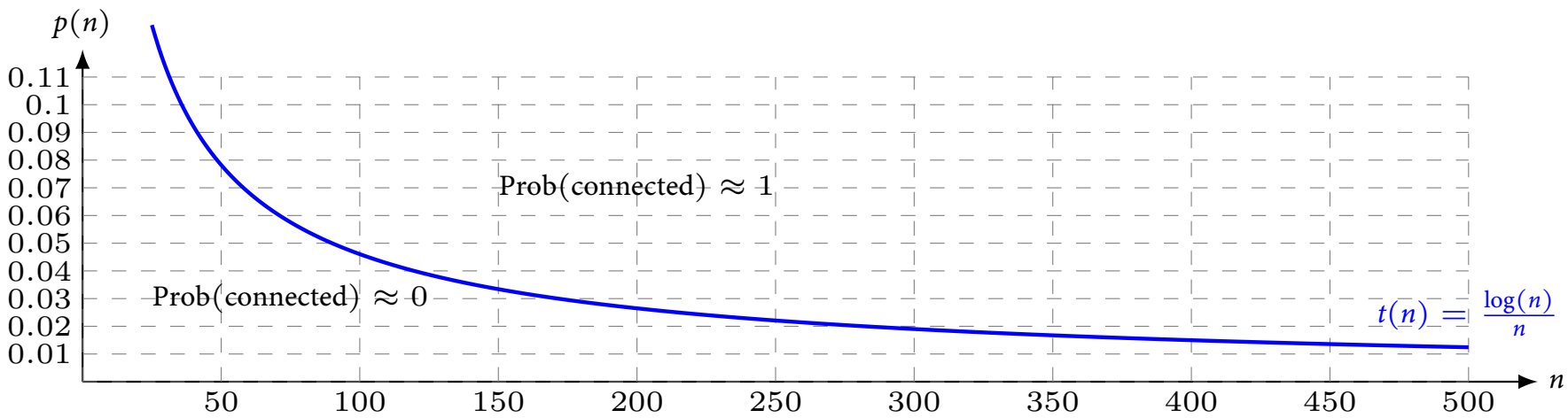
Threshold function for connectivity

Threshold function for connectivity

- Theorem (Erdos and Renyi 1959): A threshold function for the connectedness of a Poisson random network is $t(n) = \frac{\log(n)}{n}$.
- We will show a stronger result: Let $p(n) = \lambda \frac{\log(n)}{n}$.

$$\text{Prob}(\text{connectivity}) \to \begin{cases} 0, & \text{if } \lambda < 1, \\ 1, & \text{if } \lambda > 1. \end{cases}$$

Threshold function for connectivity (Cont.)



Proof

- Write $p = p(n) = \lambda \frac{\log(n)}{n}$. Clearly, $p \to 0$.
- Let I_i be a Bernoulli random variable defined as

$$I_i = \begin{cases} 1, & \text{if node } i \text{ is isolated,} \\ 0, & \text{otherwise.} \end{cases}$$

- We can write the probability that an individual node is isolated as

$$q = \text{Prob}(I_i = 1) = \left(1 - p\right)^{n-1} \approx e^{-pn} = e^{-\lambda \log(n)} = n^{-\lambda},$$

where the third equality is due to $p \to 0$.

This also implies that $q \to 0$.

Proof

- Let $X = \sum_{i=1}^{n} I_i$ denote the total number of isolated nodes.
- Then, we have

$$\mathbf{E}[X] \approx n \times n^{-\lambda} = n^{1-\lambda}.$$

Proof: The case $\lambda < 1$

- Assume that $\lambda < 1$.
- We have shown that

$$\text{Prob}(\text{at least one isolated node}) \to 1.$$

- Therefore, $\text{Prob}(\text{disconnected}) \to 1$ as $n \to \infty$, completing the proof.

Proof: The case $\lambda > 1$

- Assume that $\lambda > 1$.

  We want to show that $\text{Prob}(\text{connectivity}) \to 1$.
- We have shown that

$$\text{Prob}(X > 0) \to 0.$$

- However, we need more to establish connectivity.

Proof: The case $\lambda > 1$ (Cont.)

- The event "graph is disconnected" is equivalent to the existence of k nodes without an edge to the remaining nodes, for some $k \leq \frac{n}{2}$.
- We have

$$\text{Prob}(\{1, 2, \ldots, k\} \text{ not connected to the rest}) = (1-p)^{k(n-k)}.$$

- Therefore,

$$\text{Prob}(\exists\, k \text{ nodes not connected to the rest}) = \binom{n}{k}(1-p)^{k(n-k)}.$$

Proof: The case $\lambda > 1$ (Cont.)

- We have

$$\text{Prob}(\text{disconnected}) \leq \sum_{k=1}^{n/2} \binom{n}{k} (1-p)^{k(n-k)}.$$

- By some (ugly) calculations, we have

$$\text{Prob}(\text{disconnected}) = o(1).$$

Subsection 6

Threshold for giant component

Giant component

- We have shown that when $p(n) \ll \frac{\log n}{n}$, $G(n,p)$ is disconnected with high probability.
- In cases for which the network is not connected, the component structure is of interest.
- We have argued that in this regime the expected number of isolated nodes goes to infinity.

  This suggests that $G(n,p)$ should have an arbitrarily large number of components.

Giant component (Cont.)

- We will next argue that the threshold $t(n) = \frac{1}{n}$ plays an important role in the component structure of the graph.
- Let $\lambda = \frac{p(n)}{t(n)}$.
- For $\lambda < 1$, all components of the graph are "small".
- For $\lambda > 1$, the graph has a unique giant component, i.e., a component that contains a constant fraction of the nodes.
- We will analyze the component structure in the vicinity of $p(n) = \frac{\lambda}{n}$ using a branching process approximation.

Giant component: Intuition

- Form a Poisson random network on $n-1$ nodes with a probability of any given link being $p > \frac{1}{n}$.
- Now let us add a last node, and again connect this node to each other node with an independent probability p.
- Let q be the fraction of nodes in the largest component of the $n-1$-node network.
- As a fairly accurate approximation for large n, this will also be the fraction of nodes in the largest component of the n node network.

  The only possible exception to this is if the added node ends up connecting two large components that were not connected before. As argued above, the chance of having two components with large numbers of nodes that are not connected to each other goes to 0 in n, given that $p > \frac{1}{n}$.

Giant component: Intuition (Cont.)

- Now, the chance that this added node ends up outside of the giant component is the probability that none of its neighbors are in the giant component.
- If the new node has degree d_i this probability is converging to $(1-q)^{d_i}$, as we let n become large.
- As we can think of any node as having been added in this way, in a large network the expected frequency of nodes of degree d_i that end up outside of the giant component is approximately $(1-q)^{d_i}$.
- So, the overall fraction of nodes outside of the giant component, $1-q$, can then be found by averaging $(1-q)^{d_i}$ across nodes.

Giant component: Intuition (Cont.)

- This leads to

$$1 - q = \sum_d (1 - q)^d P(d).$$

- $P(\cdot)$ follows the Poisson distribution:

$$\begin{aligned} 1 - q &= \sum_d (1 - q)^d \frac{e^{-(n-1)p}((n-1)p)^d}{d!} \\ &= e^{-(n-1)p} \sum_d \frac{[(n-1)p(1-q)]^d}{d!} \\ &= e^{-(n-1)p} e^{(n-1)p(1-q)} = e^{-(n-1)pq}. \end{aligned}$$

Giant component: Intuition (Cont.)

- There is always a solution of $q = 0$ to this equation. In the case where the average degree is larger than 1 (i.e., $p > 1/(n-1)$), and only then, there is also a solution for q that lies between 0 and 1. (why?)
- This corresponds to phase transition, in that the appearance of such a giant component comes above the threshold of $(n-1)p = 1$.
- That is, there is a marked difference in the structure of the resulting network depending on whether average degree is bigger or smaller than one.

Subsection 7

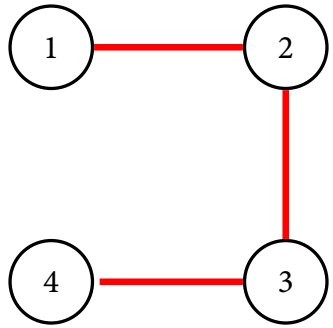
Degree distribution of a neighboring node

Degree of a neighboring node

- Start at some node i with degree d_i.
- Consider a neighbor j. How many neighbors do we expect j to have?
- This is important in estimating the size of i's expanding neighborhoods, in keeping track of contagion and transmission of beliefs, in estimating diameters, and many other calculations.
- Basically, any time that we consider some process that moves through the network and we wish to keep track of how many lines it expects to have to be able to follow at a next step, this is an important sort of calculation.

Example

This degree distribution of a neighboring node is different from the degree distribution. Consider the following network:



- Degree distribution: $P(1) = P(2) = \frac{1}{2}$.
- If we randomly pick a link and then randomly pick an end of it, there is a $\frac{2}{3}$ chance that we find a node of degree 2 and a $\frac{1}{3}$ chance that we find a node of degree 1.

Degree distribution of a neighboring node

- Higher degree nodes are involved in a proportionately higher percentage of the links.
- If we randomly pick a link and a node at the end of it, and we consider two nodes of degrees d_j and d_k, then node k is relatively $\frac{d_k}{d_j}$ times more likely to be the one we find than node j.
- The distribution of degrees of a node found by choosing a link uniformly at random from a network that has degree distribution P and then picking either one of the end nodes with equal probability is

$$\tilde{P}(d) = \frac{P(d)d}{\langle d \rangle}.$$

Degree distribution of a neighboring node (Cont.)

Details:

- Prob(a degree-d node) $= \frac{d}{d'}$ Prob(a degree-d' node).
- We have

$$\begin{aligned}\tilde{P}(d) &= \text{Prob(degree-}d\text{ nodes)} = P(d)\,\text{Prob(a degree-}d\text{ node)} \\ &= P(d)\frac{d}{d'}\,\text{Prob(a degree-}d'\text{ node)} \\ &= \frac{P(d)d}{p(d')d'}P(d')\,\text{Prob(a degree-}d'\text{ node)} = \frac{P(d)d}{p(d')d'}\tilde{P}(d')\end{aligned}$$

- Since $\sum_d \tilde{P}(d) = 1, \sum_d \frac{P(d)d}{P(d')d'}\tilde{P}(d') = 1$, and hence

$$\tilde{P}(d') = \frac{P(d')d'}{\sum_d P(d)d} = \frac{P(d')d'}{\langle d\rangle}.$$

Subsection 8

Diameter estimation

Diameter estimation

- let us start by calculating the diameter of a network which makes such calculations relatively easy.
- Suppose that we examine a component that we know to be a tree so that there are no cycles.
- A method of obtaining an upper bound on diameter is to pick some node and then successively expanding its neighborhood by following paths of length ℓ, where we increase ℓ until the paths are long enough so that we reach all nodes.

Cayley tree

- Consider a tree such that every node either has degree k or degree 1 (the leaves), and such that there is a root node that is equidistant from all of the leaves.
- Start from that root node.
- If we then move out by a path of 1, we have reached k nodes.
- By traveling on all paths of length 2, we will have reached all of the nodes in the immediate neighborhoods of the nodes in the original node's neighborhood. We will have reached $k + k(k-1)$ or k^2 nodes.

Diameter estimation: Cayley tree

- Extending this reasoning, by traveling on all paths of length ℓ, we will have reached

$$k + k(k-1) + k(k-1)^2 + \cdots + k(k-1)^{\ell-1} = k\frac{(k-1)^\ell - 1}{k-2},$$

  roughly $(k-1)^\ell$.
- To reach $n-1$ nodes, we need roughly $(k-1)^\ell = n$, or ℓ is on the order of $\frac{\log n}{\log d}$.
- Diameter is roughly $2\frac{\log n}{\log d}$.

Diameter estimation: Poisson random networks

- A randomly picked node i has an expected number of neighbors of $\langle d \rangle$.
- If we presume that nodes' degrees are approximately independent, then each of these neighbor nodes has a degree described by the distribution $\tilde{P}(d)$.
- Thus, each of these neighbor nodes has an expected number of neighbors (besides i) of

$$\sum_d (d-1)\tilde{P}(d) = \sum_d (d-1)\frac{P(d)d}{\langle d \rangle} = \frac{\langle d^2 \rangle - \langle d \rangle}{\langle d \rangle}.$$

- The expected number of i's second neighbors is roughly $\langle d \rangle \frac{\langle d^2 \rangle - \langle d \rangle}{\langle d \rangle}$.

Diameter estimation: Poisson random networks (Cont.)

- Iterating, the expected number of k-th neighbors is estimated by

$$\langle d \rangle \left(\frac{\langle d^2 \rangle - \langle d \rangle}{\langle d \rangle} \right)^{k-1}.$$

- This means that expanding out to a ℓ-th neighborhood reaches

$$\sum_{k=1}^{\ell} \langle d \rangle \left(\frac{\langle d^2 \rangle - \langle d \rangle}{\langle d \rangle} \right)^{k-1} \text{ nodes.}$$

- Let the above expression be $n - 1$.
-

$$\ell = \frac{\log\left((n-1)\frac{\langle d \rangle - 1}{\langle d \rangle} + 1 \right)}{\log \langle d \rangle} \approx \frac{\log n}{\log \langle d \rangle}.$$