Network Science: Principles and Applications[2]

CS 695 - Fall 2016

Amarda Shehu, Fei Li

Department of Computer Science
George Mason University

[2]Part of the material is from Alistair Sinclair's lecture notes. Part of the material is from the book "Foundations of Data Science" by Avrim Blum, John Hopcroft, and Ravindran Kannan

Introduction

- Computer science as an academic discipline began in the 60's.
  Emphasis was on *programming languages*, *compilers*, *operating systems*, and the *mathematical theory* that supported these areas.
  Courses in theoretical computer science covered *finite automata*, *regular expressions*, *context free languages*, and *computability*.
- In the 70's, *algorithms* were added. The emphasis was on making computers useful.
- Today, a fundamental change is taking place and the focus is more on *applications*.
  The merging of computing and communications has played an important role.

Topics – domain-independent techniques

Use computers to *understand* and make usable data arising in applications, not just how to make computers useful on specific well-defined problems.

One of the major changes is the switch from discrete mathematics to more of an emphasis on **probability**, **statistics**, and **numerical methods**.

From *local choices* to *global phenomena*:

1. random graphs
2. random walks
3. Markov chains
4. generative models
5. etc.

Note

What is different about the modern study of large graphs from traditional graph theory and graph algorithms is: One seeks statistical properties of these very large graphs rather than an exact answer to questions.

Note

One formulates abstract models of graphs that are not completely realistic in every situation, but admit a nice mathematical development that can guide what happens in practical situations.

Start — law of disorder, law of statistical behavior

Problem (Drunkard's walk)

Suppose we watch a drunkard who has been leaning against a lamp post in the middle of a large paved city square and then has decided to go nowhere in particular. How far will be our drunkard from the lamp post after he has executed, say, a hundred phases of his irregular zigzag journal?

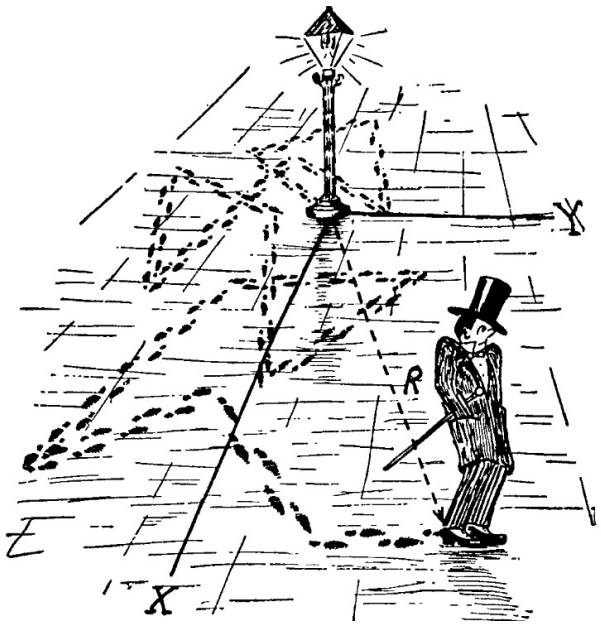


Figure : from George Gamow, "One two three . . . infinity, facts and speculations of science", 1974

Draw on the pavement two coordinate axes with the origin in the lamp post. The X-axis coming toward us and the Y-axis to the right.
Let R be the distance of the drunkard form the lamp post after the total of n zigzags. If now X_n and Y_n are the projections of the n-th lege of the track on the corresponding axis, we have

$$\begin{aligned} R^2 &= (X_1 + X_2 + \cdots + X_n)^2 + (Y_1 + Y_2 + \cdots + Y_n)^2 \\ &= \sum_{i=1}^{n} X_i^2 + \sum_{i=1}^{n} Y_i^2 + 2\left(\sum_{i \neq j} (X_i \cdot X_j) + \sum_{i \neq j} (Y_i \cdot Y_j)\right) \\ E[R^2] &= \sum_{i=1}^{n} E[X_i^2] + \sum_{i=1}^{n} E[Y_i^2] + 2\left(\sum_{i \neq j} E[X_i \cdot X_j] + \sum_{i \neq j} E[Y_i \cdot Y_j]\right) \\ &= \sum_{i=1}^{n} X_i^2 + \sum_{i=1}^{n} Y_i^2 \end{aligned}$$

Applications of the drunkard problem

1. *diffusion speed* (dye molecule movement):
2. *heat propagation speed* (hot electron gas and cold electron gas):
3. *particles of light*:
   - sun's radius $\approx 700,000$km
   - free pass of a light quantum is solar matter $\approx$ 1cm
     require $\left(7 \cdot 10^{10}\right)^2 \approx \left(5 \cdot 10^{21}\right)$ drunkard's steps to reach the surface
   - light speed $\approx 300,000$km
   - require $(1/3) \cdot 10^{-10} \approx 3 \cdot 10^{-11}$ second to take a step
   - entire travel time is about $1.5 \cdot 10^{11}$ seconds, which is about 5000 years
4. ...

Random networks (graphs)

Large graphs appear in many contexts such as the **World Wide Web**, **the Internet**, **social networks**, **journal citations**, and other places.

large graphs statistical properties (statistical mechanics)	traditional graph theory and algorithms an exact answer to questions (mechanics)

- The subject began in 1959/1960 with the monumental paper *On the Evolution of Random Graphs* by Paul Erdos and Alfred Renyi.
- The book *Random Graphs* by Bollobs (2001) is the standard source.
- Another book, also entitled *Random Graphs* by Janson, Luczak and Rucinski (2000) is excellent.

The philosophy behind a random network is simple: Placing the links randomly between the vertices. A random network consists of n labeled vertices where each vertex pair is connected with the same probability p.

Definition (Random network model)

Let n be a positive integer, $0 < p < 1$. The random graph $G(n, p)$ is a probability space over the set of graphs on the vertex set $\{1, \ldots, n\}$ determined by

$$\Pr[\{i, j\} \in G] = p,$$

with these events mutually independent.

There are two *equivalent* (?) ways of defining a random network:

- $G(n, l)$ model: n labeled vertices are connected with exact l randomly placed links. (Erdos & Renyi, 1959)
- $G(n, p)$ model: Each pair of n labeled vertices is connected with probability p. (Gilbert, 1959).

Problem

Are these two models **really** *equivalent to other other?*

Exercise

Given a $G(n, l)$, generate an *equivalent* $G(n, p)$.

Proof.

$$p = \frac{2l}{n(n-1)}$$

□

Exercise

Given a $G(n, p)$, generate an equivalent $G(n, l)$.

Proof.

$$l = \frac{n(n-1)}{2} \cdot p$$

□

Algorithm 1 Construct a random network $G(n, p)$

1: Start with n isolated vertices.
2: **for** $i = 1$ to $n(n-1)/2$ **do**
3: select a vertex pair, and generate a random number $\alpha \in [0, 1]$;
4: **if** $\alpha \geq p$ **then**
5: connect the selected vertex pair with a link.
6: **end if**
7: **end for**

Each random network we generate with the same parameters n, p will look slightly different.

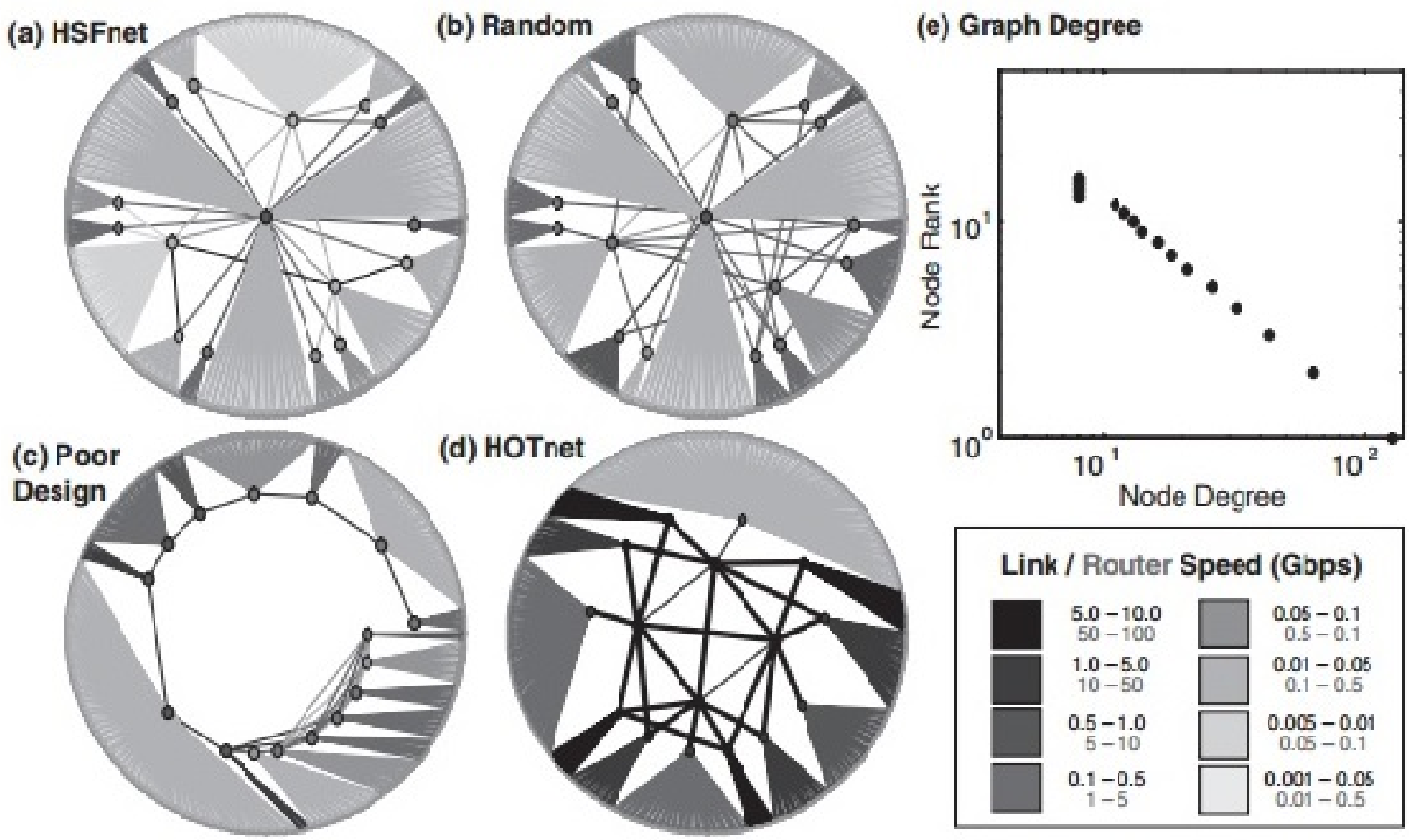


Figure : from Li et al's "Towards a Theory of Scale-Free Graphs: Definition, Properties, and Implications", Internet Mathematics, 2005

Algorithm 2 Procedure 1

1: Start with a complete graph with n vertices.
2: **for** $i = 1$ to $n(n-1)/2$ **do**
3: select a vertex pair, and generate a random number $\alpha \in [0, 1]$;
4: **if** $\alpha < p$ **then**
5: disconnect them.
6: **end if**
7: **end for**

Questions to answer

Given a network $G(n, p)$, some questions to answer:

1. **networks**:
   $|G(n, p)|$
   observe $\rightarrow$ interpret $\rightarrow$ model $\rightarrow$ validate $\rightarrow$ control (Michael Mitzenmacher, Harvard University)
   evolution of a random network
2. **vertex**:
   (average/expected) degree for a single vertex
3. **vertices**:
   distribution of the number of links
   power law distribution or log-nominal
4. **vertex** + **vertex**:
   connected component; distance (If we increase the average degree of a network, we will observe an abrupt transition from a collection of disconnected vertices to a state in which the graph contains a giant component.)
   small-world property
5. ...

Example (Phase transition at the threshold $d = 1$, from $o(n)$ to $\Omega(n)$)

Suppose the vertices represent people and an edge means the two people it connects know each other.
Given a chain of connections, such as A knows B, B knows C, C knows D, ..., and Y knows Z, we say that A indirectly knows Z. Thus, all people belonging to a connected component of the graph indirectly know each other.
Suppose each pair of people, independent of other pairs, tosses a coin that comes up heads with probability $p = d/n$. If it is heads, they know each other; if it comes up tails, they don't. The value of d can be interpreted as the expected number of people a single person directly knows.
The question arises as to how large are sets of people who indirectly know each other?

Definition (Threshold function)

A function $r(n)$ is a *threshold function* for some property P, if whenever $p = p(n) \ll r(n)$ then $G(n, p)$ does not satisfy P almost always, and whenever $p \gg r(n)$ then $G(n, p)$ satisfies P almost always.

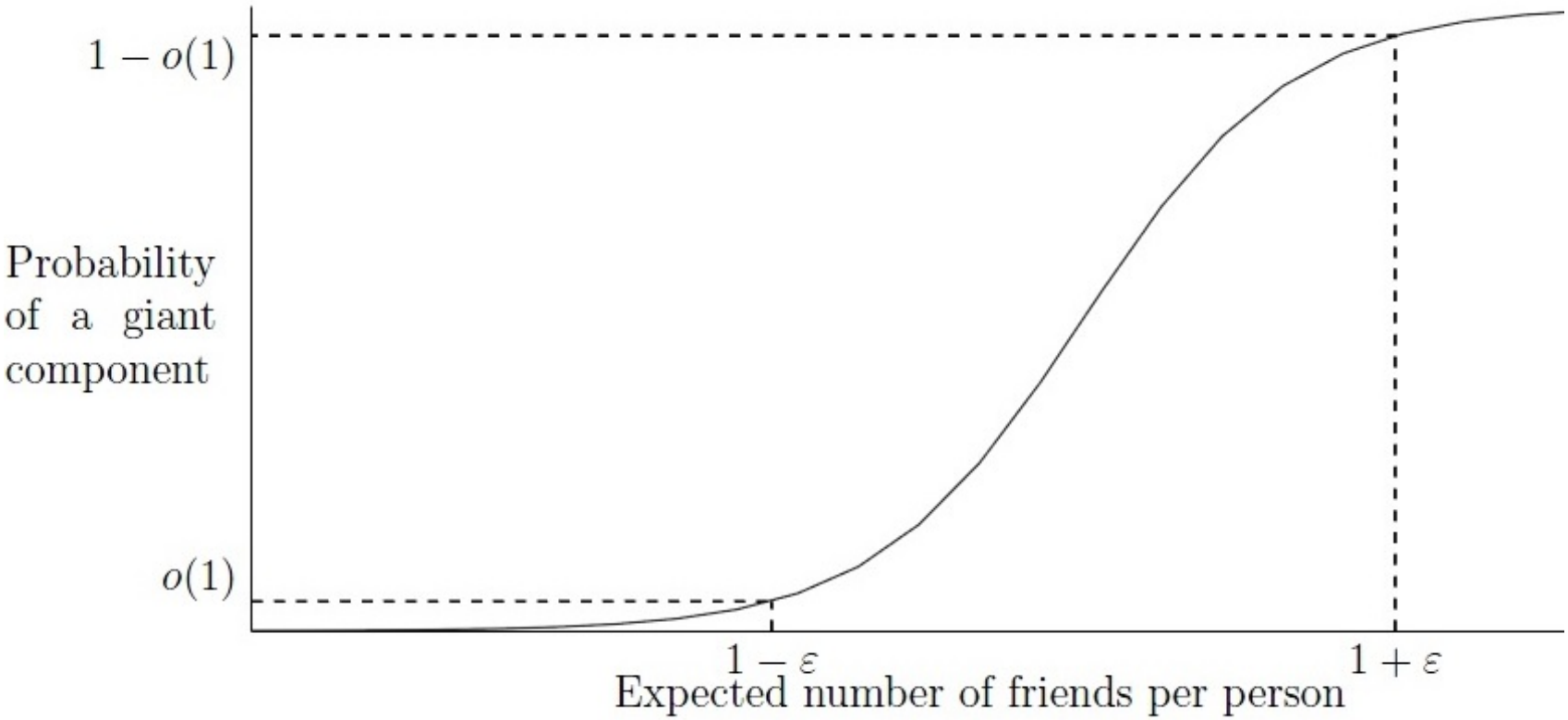


Figure : from Avrim Blum, John Hopcroft, and Ravindran Kannan's book "Foundations of Data Science"

Example (Melting points)

Think of a solid as a three-dimensional grid of molecules, with neighboring molecules joined by bonds. Adding energy excites molecules and breaks bonds. We assume that bonds break at random as we raise the temperature (energy level). Each temperature corresponds to some fraction of bonds broken. While the graph remains largely connected, the material seems solid. Breaking off small pieces does not change this, but when all'the components are small the global nature of the material changes. Small components of.molecules float freely, like a liquid or gas.

Problem

How many edges are needed to make a random graph on n vertices connected?

Algorithm 3 Experiment 1 (m)

1: Start with n vertices $V = \{1, 2, \ldots, n\}$ and $E = \emptyset$.
2: **for** $i = 1$ to m **do**
3: select an edge $e = \{i, j\} \notin E$ uniformly at random;
4: add e to E;
5: **end for**
6: **return** $G = (V, E)$.

Sample space: All $\binom{\binom{n}{2}}{m}$ labeled graph G with n vertices and m edges, each equally likely.

Exercise

Enumerate the entire sample space for $n = 4$ and $m = 3$.

Let G be a random graph (sample point) as Experiment 1. We want to find the *smallest value* of m such that Pr[G is connected] is close to 1.

Algorithm 4 Experiment 2

1: Start with n vertices $V = \{1, 2, \ldots, n\}$ and $E = \emptyset$.
2: **while** $G = (V, E)$ is not connected **do**
3: select an edge $e = \{i, j\} \notin E$ uniformly at random;
4: add e to E;
5: **end while**
6: **return** $G = (V, E)$.

Sample space: All connected graphs G obtained from Experiment 2, with probabilities ???

Exercise

Enumerate the entire sample space (including the probabilities) for $n = 4$.

Consider the random variable $X = |E|$, i.e., the number of edges in the graph G at the end of the experiment. Let us compute $E(X)$.

1. $X_k =$ number of edges added to reduce number of components from k to $k-1$.
2. $X = \sum_{k=2}^{n} X_k$ and $E(X) = \sum_{k=2}^{n} E(X_k)$.

What does X_k look like?

1. $X_n = 1$ always with probability 1.
2. $X_{n-1} = 1$ with probability 1.
3. The distribution of X_{n-2} depends on the first two edges; but presumedly its expectation is not much bigger than 1.
4. Similarly, for $k < n-2$, the distribution of the X_k become rather complicated, but maybe we can compute an **upper bound** on $E(X_k)$.

Lemma

For all k, we have $E(X_k) \leq \frac{n-1}{k-1}$.

Proof.

Suppose G has exactly $k > 1$ components. Consider *any* vertex i.
Our experiment is equally likely to pick any of the edges $\{i, j\}$ that is not in E.
There are at most $n - 1$ such edges, of course. How many of them reduce the number of components? *At least* $k - 1$.
Therefore, the probability that any such edge reduces the number of components is at least $\frac{k-1}{n-1}$. Since this holds for every vertex i, it holds in general.
But now we see that $X_k \leq Y_k$, where $Y_k =$ coin flips up to and including first head for a coin with $\Pr[\text{heads}] = p = \frac{k-1}{n-1}$. And we know $E(Y_k) = \frac{1}{p} = \frac{n-1}{k-1}$.
Hence $E(X_k) \leq \frac{n-1}{k-1}$. □

We are done:

$$\begin{aligned} E(X) &= \sum_{k=2}^{n} E(X_k) \\ &\leq (n-1)\left(1+\frac{1}{2}+\frac{1}{3}+\cdots+\frac{1}{n-1}\right) \\ &\approx (n-1)(\ln(n-1)+\gamma) \end{aligned}$$

So the expected number of edges at the end of Experiment 2 is at most $(n-1)(\ln(n-1)+\gamma)$ as $n \to \infty$.

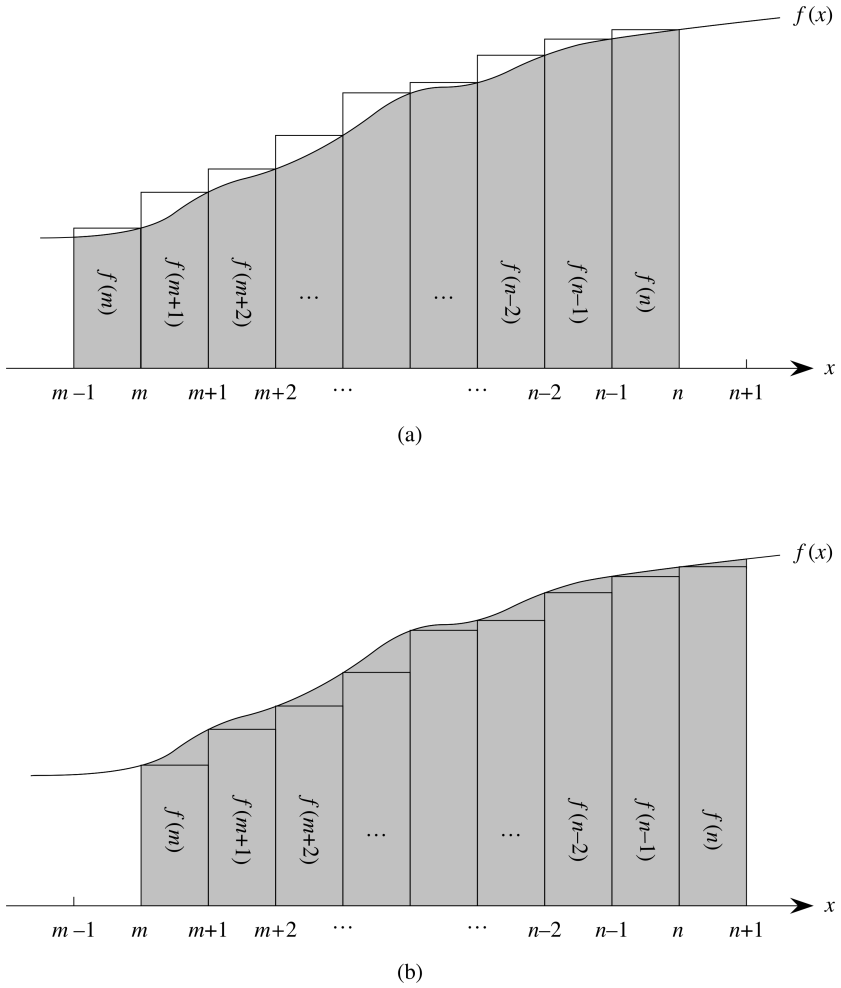


Figure : Approximation of $\sum_{k=m}^{n} f(k)$ by integrals. The area of each rectangle is shown within the rectangle, and the total rectangle area represents the value of the summation. The integral is represented by the shaded area under the curve. By comparing areas in (a), we get $\int_{m-1}^{n} f(x)dx \leq \sum_{k=m}^{n} f(k)$, and then by shifting the rectangles one unit to the right, we get $\sum_{k=m}^{n} f(k) \leq \int_{m}^{n+1} f(x)dx$ in (b).

Note

This is only an upper bound on $E(X)$. The exact answer (which requires more effort) is $E(X) \approx \frac{n}{2} \ln n$. So we are off by only a factor of 2.

Exercise

In the above proof, we said $X_k \leq Y_k$. Note that X_k and Y_k are not numbers but random variables (over different sample spaces). What it means, precisely, is the $\Pr[X_k \geq z] \leq \Pr[Y_k \geq z]$ for all z.
Show that $X_k \leq Y_k$ implies that $E(X_k) \leq E(Y_k)$, as we used in the proof.

We have seen that the expected number of edges required to make the graph connected is at most $M(n) = (n-1)(\ln(n-1) + \gamma)$.

Problem

What is the probability that we need much more than this?

Theorem (Markov's inequality)

Let X be a random variable taking non-negative values, and let $\mu = E(X)$. Then $\Pr[X \geq c \cdot \mu] \leq \frac{1}{c}$, $\forall c \geq 1$.

Proof.

$$\begin{aligned}
\mu &= E(X) \\
&= \sum_k \Pr[X = k] \cdot k \\
&\geq \sum_{k \geq c \cdot \mu} \Pr[X = k] \cdot k \\
&\geq c \cdot \mu \sum_{k \geq c \cdot \mu} \Pr[X = k] \\
&= c \cdot \mu \cdot \Pr[X \geq c \cdot \mu]
\end{aligned}$$

Exercise

The above proof is not formally valid when $\mu = 0$, since in the last step we cancel μ. Is the theorem still true when $\mu = 0$?

Exercise

Give a simple counterexample which shows that Markov's inequality is definitely false if we drop the assumption that X is non-negative.

Applying Markov's inequality to the random variable X, we get

$$\Pr[X \geq c \cdot M(n)] \leq \frac{1}{c} \qquad \forall c \geq 1$$

For example, if we add $10M(n)$ edges, the probability that G is connected is at least $9/10$.

Suppose we modify Experiment 2 slightly so that, if G becomes connected before m edges have been added, we still continue to add random edges until G has exactly m edges. View each point of the sample space of this modified Experiment 2 as $G = G' + G''$, where G' is the graph consisting of the first m edges and G'' is the remainder. (Note that G'' will be empty if G' is connected.) Then it should be clear that

$$\Pr[X \leq m] = \Pr[G' \text{ is connected}]$$

What is the relationship with Experiment 1?
The sample space of graph G' is *exactly* the same as the sample space of Experiment 1. So, we get

$$\Pr_1[G \text{ is connected}] = \Pr_2[G' \text{ is connected}] = \Pr_2[X \leq m],$$

where $\Pr_1$ and $\Pr_2$ denote probabilities in Experiments 1 and 2 respectively.

Now, putting $m = c \cdot M(n)$ in Experiment 1, we get from $\Pr[X \geq c \cdot M(n)] \leq \frac{1}{c}$ that

$$\Pr_1[G \text{ is connected}] = \Pr_2[X \leq c \cdot M(n)] \geq 1 - \frac{1}{c},$$

which gives a good answer to our original question about Experiment 1; i.e., a random graph with n vertices and $m = c(n-1)(\ln(n-1) + \gamma)$ edges is connected with probability at least $1 - \frac{1}{c}$.

On the total number of links/degrees

Theorem

Consider $G(n, p)$. The expected number of links in a random graph can be calculated as

$$p\frac{n(n-1)}{2}$$

The average degree of a random network is $p(n-1)$.

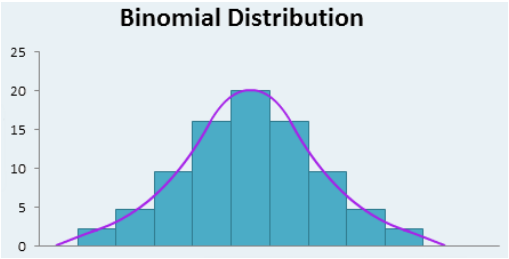
Proof.

Use indicator variables. □

Theorem

The degree distribution p_k providing the probability that a randomly chosen node has degree k follows

$$p_k = \binom{n-1}{k} p^k \cdot (1-p)^{n-1-k}.$$



Most real networks are sparse, hence $\bar{k} \leq n$. In this limit the degree distribution is well *approximated* by the Poisson distribution

$$p_k = e^{\bar{k}} \frac{\bar{k}^k}{k!}$$

Assume $k \ll n$.

$$
\begin{aligned}
p_k &= \binom{n-1}{k} p^k \cdot (1-p)^{n-1-k} \\
&= \frac{(n-1)(n-1-1)(n-1-2)\cdots(n-1-k)!}{k!(n-1-k)!} p^k (1-p)^{n-1-k} \\
&\approx \frac{(n-1)^k}{k!} p^k (1-p)^{n-1-k} \\
&= ((n-1)p)^k \frac{(1-p)^{n-1-k}}{k!} = \bar{k}^k \frac{(1-p)^{n-1-k}}{k!} \\
(1-p)^{n-1-k} &= e^{\ln(1-p)^{n-1-k}} \\
&= e^{(n-1-k)\ln(1-p)} \\
&= e^{(n-1-k)\ln(1+(-p))} \\
&= e^{(n-1-k)\ln \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n}(-p)^n} \\
&= e^{(n-1-k)\ln\left((-p)-\frac{(-p)^2}{2}+\frac{(-p)^3}{3}-\ldots\right)} \\
&\approx e^{(n-1-k)\frac{\bar{k}}{n-1}}
\end{aligned}
$$

Definition (Poisson distribution)

An event can occur 0, 1, 2, ... times in an interval. The average number of events in an interval is designated λ. λ is the event rate, also called the *rate parameter*. The probability of observing k events in an interval is given by the equation

$$\Pr[k \text{ events in interval}] = \frac{\lambda^k e^{-\lambda}}{k!},$$

where

- λ is the average number of events per interval
- e is the number 2.71828... (Euler's number) the base of the natural logarithms
- k takes values $0, 1, 2, \ldots$
- $k! = k \times (k-1) \times (k-2) \times \ldots \times 2 \times 1$ is the factorial of k.

This equation is the probability mass function (PMF) for a Poisson distribution.

Example

Existence of triangles in $G(n, d/n)$ What is the expected number of triangles in $G(n, d/n)$, when d is a constant?

Proof.

□

As the number of vertices increases one might expect the number of triangles to increase, but this is not the case.

Example (Triangle-free graphs)

Given n and p, whether does the random graph $G(n,p)$ have triangles?

Let X be the number of triangles contained in $G(n,p)$.

$$\begin{aligned} E[X] &= \binom{n}{3} p^3 \\ \lim_{n\to\infty} E[X] &= \lim_{n\to\infty} \binom{n}{3} p^3 \\ &= \frac{n(n-1)(n-2)}{6} p^3 \\ &\approx \frac{n^3}{6} p^3 \end{aligned}$$

Suggest parameterizations $p = c/n$. We have

$$\begin{aligned} \lim_{n\to\infty} E[X] &= \frac{c^3}{6} \\ \lim_{n\to\infty} \Pr[X=0] &= \frac{\lambda^0 e^{-\lambda}}{0!} \\ &= e^{-c^3/6} \end{aligned}$$

Definition (Upper cutoff)

To determine the expected degree of the *largest node* in a random network, called the *network's upper cutoff*, we define the degree $k_{\max}$ such that in a network of n nodes we have at most one node with degree higher than $k_{\max}$. The network's largest node satisfies:

$$\begin{aligned} n(1 - P(k_{\max})) &\approx 1 \\ n(1 - e^{-\bar{k}} \sum_{k=0}^{k_{\max}} \frac{\bar{k}^k}{k!}) &\approx 1 \end{aligned}$$

Example

For $n = 10^9$, and $\bar{k} = 1,000$ corresponding to roughly the size and average degree of the globe's social network, we obtain $k_{\max} = 1,185$, indicating that a random network lacks extremely popular individuals, or hubs. $k_{\min} = 816$.

Theorem (Rapoport's observation on network density)

Changes in the structure of a random graph with increasing p, illustrating the absence of a giant component for small p and its sudden emergence once p exceeds a critical value. We have a giant component if and only if when each node has on average one link.

Proof.

□

Theorem (Gilbert, 1959)

When p is constant, almost every $G(n,p)$ is connected.

Proof.

We can make $G(n,p)$ disconnected by picking a vertex partition into two sets and forbidding edges between the two sets. Occurrence of edges within the sets is irrelevant. We bound the probability q_n that $G(n,p)$ is disconnected by summing $\Pr([S,\bar{S}]=\emptyset)$ over all bipartitions S, $\bar{S}$. Graphs with many components are counted many times. When $|S|=k$, there are $k(n-k)$ possible edges in $[S,\bar{S}]$. Each has probability $1-p$ of not appearing, independently, so $\Pr([S,\bar{S}]=\emptyset)=(1-p)^{k(n-k)}$. Considering all S generates each partition from each side, so

$$q_n \leq (1/2)\sum_{k=1}^{n-1}\binom{n}{k}(1-p)^{k(n-k)}$$

□

Considering all S generates each partition from each side, so

$$q_n \leq (1/2) \sum_{k=1}^{n-1} \binom{n}{k} (1-p)^{k(n-k)}$$

This formula is symmetric in k and $n-k$; hence q_n is bounded by $\sum_{k=1}^{n/2} \binom{n}{k} (1-p)^{k(n-k)}$. We loosen the bound to simplify it. Using $\binom{n}{k} < n^k$ and $(1-p)^{n-k} \leq (1-p)^{n/2}$ (for $k \leq n/2$) yields $q_n < \sum_{k=1}^{n/2} (n(1-p)^{n/2})^k$. For large enough n, we have $n(1-p)^{n/2} < 1$. This makes our bound the initial portion of a convergent geometric series. We obtain $q_n < x/(1-x)$, where $x = n(1-p)^{n/2}$. Since $n(1-p)^{n/2} \to 0$ when p is constant, our bound on q_n approaches 0 as $n \to \infty$.

Markov's inequality ($\Pr[X \geq t] \leq \frac{E(X)}{t}$) implies that if X is integer-valued, then $E(X) \to 0$ implies $\Pr(X = 0) \to 1$.

Theorem

If p is constant, then almost every $G(n, p)$ has diameter 2 (and hence is connected).

Proof.

Let $X(G(n, p))$ be the number of unordered vertex pairs with no common neighbor. If there are none, then $G(n, p)$ is connected and has diameter 2. By Markov's Inequality, we need only show $E(X) \to 0$. We express X as the sum of $\binom{n}{2}$ indicator variables $X_{i,j}$, one for each vertex pair $\{v_i, v_j\}$, where $X_{i,j} = 1$ if and only if v_i, v_j have no common neighbor.

When $X_{i,j} = 1$, the $n - 2$ other vertices fail to have edges to both of these, so $\Pr(X_{i,j} = 1) = (1 - p^2)^{n-2}$ and $E(X) = \binom{n}{2}(1 - p^2)^{n-2}$. When p is fixed, $E(X) \to 0$, and hence almost every $G(n, p)$ has diameter 2. □

Roughly speaking, random graphs with constant edge probability are connected because they have many more edges than needed to be connected. To improve the results, we want to make p as small as possible to have almost every $G(n, p)$ connected.

Theorem (Threshold function for single component)

Consider $G(n, p)$. $\ln n/n$ is a threshold probability function for the disappearance of isolated vertices (that is, minimum degree of $G(n, p) > 1$).

Small-world phenomenon

Social networks are so rich in short paths is known as the *small-world phenomenon*, or the "six degrees of separation".
Social psychologist Stanley Milgram who asked randomly chosen "starter" individuals to each try forwarding a letter to a designated "target" person living in the town of Sharon, MA, a suburb of Boston.
He provided the target's name, address, occupation, and some personal information, but stipulated that the participants could not mail the letter directly to the target; rather, each participant could only advance the letter by forwarding it to a single acquaintance that he or she knew on a first-name basis, with the goal of reaching the target as rapidly as possible.
Roughly a third of the letters eventually arrived at the target, in a median of six steps, and this has since served as basic experimental evidence for the existence of short paths in the global friendship network, linking all (or almost all) of us together in society. This style of experiment, constructing paths through social networks to distant target people, has been repeated by a number of other groups in subsequent decades.

Small world property

According to six degrees of separation any two individuals, anywhere in the world, can be connected through a chain of six or fewer acquaintances. In the language of network science six degrees, also called the *small world property*, states that the distance between any two nodes in a network is unexpectedly small.

- What does short (or small) mean, i.e. short compared to what?
- How do we explain the existence of these short distances?

Consider a random network with average degree $\bar{k}$. A node in this network has on average:

- $\bar{k}$ nodes at distance one ($d = 1$)
- $\bar{k}^2$ nodes at distance two ($d = 2$).
- $\bar{k}^3$ nodes at distance three ($d = 3$).
- ...
- $\bar{k}^d$ nodes at distance d.

Proof.

The expected number of nodes up to distance d from our starting node is

$$n(d) \approx 1 + \bar{k} + \bar{k}^2 + \cdots + \bar{k}^d = \frac{\bar{k}^{d+1}}{\bar{k} - 1}.$$

$$\begin{aligned} n(d_{\max}) &= n \\ \bar{k}^{d_{\max}} &\approx n \\ d_{\max} &\approx \frac{\log n}{\log \bar{k}} \end{aligned}$$

Summary

1. Average degree: $\bar{k} = p(n-1)$
2. Average number of links: $\bar{L} = \frac{p(n-1)}{2}$
3. degree distribution $p_k = \binom{n-1}{k} p^k (1-p)^{n-1-k}$. If $k \ll n$, p_k is approximated as $e^{-\bar{k}} \frac{\bar{k}^k}{k!}$
4. Giant component n_G
   - $\bar{k} < 1$: no giant component $n_G \approx \ln n$
   - $1 < \bar{k} < \ln n$: one giant component and disconnected clusters $n_G \approx n^{2/3}$
   - $\bar{k} > \ln n$: all nodes join the giant component: $n_G \approx (p - p_j)n$
5. Average distance $\bar{d} \propto \frac{\log n}{\log \bar{k}}$
6. Clustering coefficient: $C = \frac{\bar{k}}{n}$