

The Degree Sequence

and

Polynomial-Time Graph Isomorphism

for

Random Graphs

- Let's say we want to construct a random graph that will be in the probability space $G(n, p)$. As explained in Lecture 43, starting with n isolated vertices, we construct such a graph one edge at a time. One way to populate the graph with edges would be to stand at each vertex and to carry out one Bernoulli trial for deciding whether to link up with each potential destination vertex. When we start this construction at the very first vertex, we will conduct $n-1$ Bernoulli trials. When we move to the next vertex, if that vertex is found isolated, we simply repeat what we did at the first vertex. On the other hand, if this vertex is already connected to the first vertex, we carry out $n-2$ Bernoulli trials for determining the connectivity vis-a-vis the remaining $n-2$ vertices; and so on. (Each Bernoulli trial could consist of flipping a loaded coin that shows heads with probability p . If we see a head vis-a-vis a destination vertex, we throw a link to that vertex.)
- An important observation to be made in this construction is that even when we run a different number of Bernoulli trials at the different vertices of a graph, for the purpose of deriving statistical properties of the edge counts at the vertices, we can pretend that we have carried out $n-1$ Bernoulli trials at each vertex. We can do so because each Bernoulli trial is run independently of the trials in the past and the trials to come. So the temporal order in which the trials are carried out is unimportant. Therefore, when we move to a new vertex and we find that vertex already connected to a certain number of previously considered vertices, we can assume those connections to be the result of locally conducted Bernoulli trials.

- Therefore, we can claim that for a graph in $G(n,p)$ the degree of a vertex (which is the number of edges incident at the vertex) must be a binomial random variable with parameters $n-1$ (for the number of trials) and p (for the probability of success). We will use $b(k; n-1, p)$ to represent this binomial distribution.

- Therefore, the probability of ^{our} finding the degree of a vertex to be equal to k is

$$b(k; n-1, p) = \binom{n-1}{k} p^k q^{n-1-k}$$

$$q = 1-p$$

- We now assume that the n vertices of a graph are labeled with integers in the descending order of their degrees. That is, the n vertices will be assumed to be labeled $1, 2, 3, \dots, n$ such that

$$d_1 \geq d_2 \geq d_3 \geq \dots \geq d_n$$

where d_i is the degree of the vertex labeled i . We will also use the notation $d_i(G)$ for this purpose, where G names a specific graph.

- We will use the notation $\Delta(G)$ for the maximum degree and $\delta(G)$ of a graph G . Obviously, $\Delta(G) = d_1(G)$ and $\delta(G) = d_n(G)$.

- The sequence (d_1, d_2, \dots, d_n) will be called the degree sequence of a graph. A degree sequence is a rich source of information for the characterization of random graphs.

- We now introduce the following three random variables that will help us better understand the information contained in a degree sequence:

X_k : the random variable whose value is the number of vertices of degree k

y_k : the random variable whose value is the number of vertices of degree at least k

Z_k : the random variable whose value is the number of vertices of degree at most k

Obviously, we have $y_k = \sum_{l \geq k} X_l$ and $Z_k = \sum_{l \leq k} X_l$

- Here are two direct consequences of the definition of a degree sequence:

$$d_r \geq k \text{ iff } y_k \geq r$$

and

$$d_{n-r} \leq k \text{ iff } Z_k \geq r+1$$

- Let's now focus on the random variable y_2 whose value is the total number of vertices of degree at least 2 in a graph of degree n . For a densely connected graph of order n , we can expect $E\{y_2\}$ to be close to n . On the other hand, for a sparsely connected graph, we can expect $E\{y_2\}$ to be close to 0. y_2 is 0 for the graph shown at right →

- We will now show that if $p = o(n^{-3/2})$ [that is if p decays faster



44-3

than $n^{-3/2}$ as $n \rightarrow \infty$ since we want $\frac{p}{n^{3/2}} \rightarrow 0$ as $n \rightarrow \infty$] then $E\{Y_2\} \rightarrow 0$.

- Before we establish the above claim, we write the following for $E\{X_k\}$ for a graph G in $G(n, p)$. As $n \rightarrow \infty$, we can say:

$$E\{X_k\} = n \cdot b(k; n-1, p) = n \binom{n-1}{k} p^k q^{n-1-k}$$

This is justified by the fact that $b(k; n-1, p)$ is the probability that the degree of a vertex is exactly k . Using the frequentist interpretation of probability, if we multiply $b(k; n-1, p)$ by n , that has to give us an estimate for how many of the n vertices will be of degree k . This estimate will only become more precise as $n \rightarrow \infty$.

- Let's now get back to the estimation of an upper bound for $E\{Y_2\}$:

$$\begin{aligned} E\{Y_2\} &= \sum_{j=2}^{n-1} E\{X_k\} = n \sum_{j=2}^{n-1} \binom{n-1}{j} p^j q^{n-1-j} \\ &= n q^{n-1} \sum_{j=2}^{n-1} \frac{(n-1)(n-2)\dots(n-j)}{j!} p^j q^{n-j} \leq q^{n-1} \sum_{j=0}^{n-1} \frac{\left(\frac{np}{q}\right)^j}{j!} = q^{n-1} e^{\frac{pn}{q}} \end{aligned}$$

Substituting $q = 1 - p$ in the bound and simplifying the bound for the case when $n \rightarrow \infty$ and $p \rightarrow 0$ such that $pn \rightarrow 0$, we can show that $E\{Y_2\} \rightarrow 0$ when $p = o(n^{-3/2})$.

- A complement of the above situation takes place when $q = o(n^{-3/2})$. As shown above, with $p = o(n^{-3/2})$, the graph consists mostly of isolated edges. With $q = o(n^{-3/2})$, the graph is densely connected, with most vertices possessing the maximum possible degree of $n-1$.

The variance of Poisson r.v. is also λ

THEOREM (X_k is Poisson Distributed):

This theorem applies when the edge placement probability p is neither too low ($p = o(n^{-3/2})$) so as to make a graph too sparsely connected, nor too high ($q = o(n^{-3/2})$) so as to make a graph too densely connected. For all other values of p , $n^{-3/2} < p < 1 - n^{-3/2}$, we have

$$P(X_k = r) \approx \frac{e^{-\lambda_k} \lambda_k^r}{r!}$$

with $\lambda_k = n \cdot b(k; n-1, p)$ and if $0 < \lim_{n \rightarrow \infty} \lambda_k < \infty$

Note that $E\{X_k\} = \lim_{n \rightarrow \infty} \lambda_k$

PROOF: The proof of the theorem is based on demonstrating the following for the r th factorial moment of X_k :

$$E_r\{X_k\} = \lambda_k^r \quad \text{as } n \rightarrow \infty$$

and the fact that Poisson is the only distribution for which the r th factorial moment equals the r th power of the mean. The r th factorial moment of a random variable X_k is defined to be

$$E_r\{X_k\} = E\{X_k (X_k - 1) \dots (X_k - r + 1)\}$$

Poisson Distribution:

A random variable X whose mean value is λ is said to have a Poisson distribution if

$$P(X = r) = \frac{e^{-\lambda} \lambda^r}{r!}, \quad k = 0, 1, 2, \dots$$

Obviously, this is a one-parameter distribution, the parameter being λ .

The Poisson distribution is also denoted $p(r; \lambda)$. So,

$$p(r; \lambda) = P(X = r) = \frac{e^{-\lambda} \lambda^r}{r!}$$

A Poisson distribution expresses the probability of k events occurring in, say, one unit of time when the average rate of event-occurrence is known and when each event occurs independently of the previous event.

► A binomial distribution can be upper-bounded and lower-bounded by Poisson distributions. For example,

$$b(k; n, p) \leq p(k; \lambda) e^{-\lambda}$$

To gain a better understanding of what is meant by a factorial moment, consider the case when $r = 2$. That is, we will consider the second factorial moment $E_2\{X_k\} = E\{X_k(X_k-1)\}$. We now claim that the second factorial moment is the ^{expected} number of ordered pairs of k -degree vertices in a graph. For a graph with 4 k -degree vertices labeled A, B, C, and D, you can have $12 = 4 \times 3$ ordered pairs of such vertices: AB, AC, AD, BA, BC, Similarly, for a graph with X_k number of k -degree vertices, you can have $X_k(X_k-1)$ number of ordered pairs of such vertices.

As established in the box at right,

$$E_2\{X_k\} = E\{X_k(X_k-1)\} = \lim_{n \rightarrow \infty} \lambda_k^2$$

Generalizing this argument to the case of r^{th} factorial moment, we can show

$$E_r\{X_k\} = \lim_{n \rightarrow \infty} \lambda_k^r$$

As mentioned earlier, Poisson random variables are the only known random variables that exhibit this property.

Does that mean we have proved that X_k is Poisson? Not really.

What we have proved is the necessary condition for a sequence of random variables (in our case, the different versions of X_k as $n \rightarrow \infty$) to converge in probability distribution to a target random variable (the Poisson random variable in our case). There is also a sufficient condition one must take care of. See Bollobas for that.

How many ordered pairs of k -degree vertices exist in a graph G in $G(n,p)$?

To answer this question, let's go back to how G "evolves" one edge at a time through Bernoulli experiments at each edge placement site. Let's choose two candidate vertices randomly for our ordered pair. Let P and Q be their labels. There are two ways for both P and Q to acquire k edges:

- ① P acquires $k-1$ edges vis-a-vis all other $n-2$ vertices (but not including Q) with probability $b(k-1; n-2, p)$. Q does the same vis-a-vis all other $n-2$ vertices (not including P) with probability $b(k-1; n-2, p)$. We next consider the Bernoulli experiment for possibly linking P and Q. This will succeed with probability p and the success will cause P and Q to have degree k . So the probability of P and Q becoming k -degree vertices through this mechanism is $p \cdot b^2(k-1; n-2, p)$.

- ② P acquires k edges vis-a-vis the other $n-2$ vertices (not including Q) with probability $b(k; n-2, p)$. Q also acquires k edges vis-a-vis the other $n-2$ vertices (not including P) with probability $b(k; n-2, p)$.

Now we need to make certain that when the Bernoulli trial takes place for possibly linking P and Q, it fails since the degrees at P and Q are already k . Thus the probability of P and Q turning out to be k -degree edges through the second mechanism is $q \cdot b^2(k; n-2, p)$.

Combining the two mechanisms, the probability that P and Q are both k -degree vertices is

$$p \cdot b^2(k-1; n-2, p) + q \cdot b^2(k; n-2, p)$$

THEOREM (Y_k and Z_k are also Poisson):

With the same conditions on p as in the previous theorem and using the same overall logic as before, we can show that Y_k and Z_k are also Poisson distributed asymptotically.

The role that λ_k plays in the Poisson distribution for X_k is played by μ_k for Y_k and ν_k for Z_k :

$$\mu_k = nB(k; n-1, p)$$

$$y_k = n[1 - B(k+1; n-1, p)]$$

where

$$B(l; m, p) = \sum_{j \geq l}^m b(j; m, p)$$

$B(l; m, p)$ gives us the probability of seeing **at least** l successes if we carry out m trials of a Bernoulli experiment in which the probability of an individual success is p .

Since the total number of ordered pairs of vertices in an n -vertex graph is $n(n-1)$ and since the probability that any particular ordered pair consists solely of k -degree vertices is as shown above, the expected number of ordered pairs of k -degree vertices is given by

$$n(n-1)[pb^2(k-1; n-2, p) + qb^2(k; n-2, p)]$$

as $n \rightarrow \infty$. This can be simplified to

$$n^2 b^2(k; n-1, p) \quad \text{as } n \rightarrow \infty$$

which is the same as π_k^2

- We will now present **upper and lower bounds** on the degree values in the beginning portion of a degree sequence.

THEOREM (Vertex Degrees Are Bounded) :

The degree d_m at the m^{th} element in a degree sequence obeys the following bounds as $n \rightarrow \infty$:

$$|d_m - pn - x\sqrt{pqn}| \leq w(n) \sqrt{\frac{pqn}{m \log \frac{n}{m}}}$$

where x is related to m by $x \approx \sqrt{2 \log \frac{n}{m}}$ and where $w(n)$ is a function that rises to infinity arbitrarily slowly as $n \rightarrow \infty$. m is also constrained by the requirement $x \rightarrow \infty$

PROOF : We will derive bounds on d_m from the probability distribution for the random variable y_k . The degree d_m and the value of y_k are intimately related since, as mentioned previously,

$$d_m \geq k \text{ iff } y_k \geq m$$

From which also follows :

$$d_m < k \text{ iff } y_k < m$$

From the previous theorem we know that

$$E\{y_k\} = \mu_k = nB(k; n-1, p) = n \sum_{j \geq k}^{n-1} b(j; n-1, p) = n \cdot P(S_{n-1, p} \geq k)$$

where, in the last equality, $S_{n-1, p}$ is the number of successes in $n-1$ trials of a Bernoulli experiment in which p is the probability of an individual success. As discussed in Lecture 43, we can use the DeMoivre-Laplace

Theorem to say

$$P(S_{n-1, p} \geq pn + h) \approx \frac{1}{\sqrt{2\pi t}} e^{-\frac{t^2}{2}}, \quad t \rightarrow \infty$$

where

$$t = \frac{h}{\sqrt{pq(n-1)}}$$

In the equation (B) above, we will adjust the value of h so that

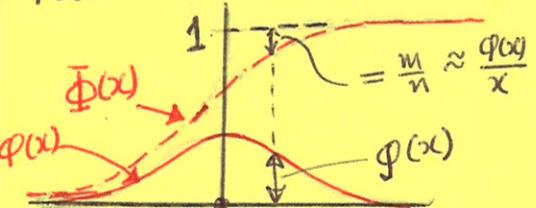
$$pn + h = pn + x\sqrt{pqn} + w(n) \sqrt{\frac{pqn}{m \log \frac{n}{m}}} = k$$

This requires that h be set to

A better way to visualize the relationship between m , n , and x is

$$1 - \Phi(x) = \frac{m}{n}$$

where $\Phi(x)$ is the distribution function of a zero-mean unit-variance Gaussian random variable.



Since $1 - \Phi(x) \approx \frac{\phi(x)}{x}$ we can write

$$\frac{m}{n} \approx \frac{1}{\sqrt{2\pi}} \frac{1}{x} e^{-\frac{x^2}{2}}$$

Taking log of both sides:

$$\log\left(\frac{m}{n}\right) \approx \log\frac{1}{\sqrt{2\pi}} - \frac{x^2}{2} - \log x$$

Since $\lim_{x \rightarrow \infty} \frac{\log x}{x^2} \rightarrow 0$, we can ignore the last term on the right. We can also ignore $\log\frac{1}{\sqrt{2\pi}}$ as $x \rightarrow \infty$. So

$$x \approx \sqrt{2 \log \frac{n}{m}}$$

$$h = p + x\sqrt{pqn} + w(n)\sqrt{\frac{pqn}{m \log \frac{n}{m}}}$$

If we substitute this value for h in the expression for t at ④ above, we get

$$t \approx \frac{1}{Nn} \sqrt{\frac{p}{q}} + x + \frac{w(n)}{\sqrt{m \log \frac{n}{m}}}$$

Setting $p(m-1)th$ to k , as indicated by the bubble in red on the right hand side of ⑤, and substituting the above value for t in the equation at ③, we get

$$P(S_{n-1,p} \geq k) \approx \frac{1}{\sqrt{2\pi} t} e^{-\frac{1}{2}(x^2 + \frac{w^2(n)}{m \log \frac{n}{m}} + \frac{2xw(n)}{\sqrt{m \log \frac{n}{m}}})}$$

where we have ignored terms in the exponent that go to zero as $n \rightarrow \infty$. Also, as $n \rightarrow \infty$, we can replace t by x and $\frac{1}{\sqrt{2\pi} x} e^{-\frac{x^2}{2}}$ by $\frac{1}{n}$ on the basis of our previous comments. Additionally, by the last equality at ①, the left hand side at ⑥ above can be replaced by $E\{y_k\}$. So we can write

$$E\{y_k\} \approx m e^{-\frac{xw(n)}{\sqrt{m \log \frac{n}{m}}} - \frac{w^2(n)}{2m \log \frac{n}{m}}}$$

Since $w(n)$ can be chosen to rise to ∞ slower than any other function as $n \rightarrow \infty$, the above equation can be massaged to yield:

$$m \geq E\{y_k\} + \sqrt{w(n)} \sqrt{E\{y_k\}}$$

We will use the above inequality in the following result we obtain from the Chebyshov's inequality as expressed at ⑦ in the box after we substitute $d = m - E\{y_k\}$ in it:

$$P((y_k - E\{y_k\}) \geq (m - E\{y_k\})) \leq \frac{E\{y_k\}}{(m - E\{y_k\})^2}$$

The bound as expressed by the right hand side of this inequality is not violated if we replace the denominator on the right by its lower bound that can be obtained from ① above. So we get

$$P\{y_k \geq m\} \leq \frac{E\{y_k\}}{2w^2(n) E\{y_k\}} = \frac{O(1)}{w^2(n)}$$

Since by specification, $w(n) \rightarrow \infty$ as $n \rightarrow \infty$, we are led to the conclusion:

$$\lim_{n \rightarrow \infty} P(y_k \geq m) = 0$$

In other words, we must have $y_k < m$ as $n \rightarrow \infty$.

But since $y_k < m$ iff $d_m < k$, we claim the following bound on d_m :

$$d_m < pn + x\sqrt{pqn} + w(n)\sqrt{\frac{pqn}{m \log \frac{n}{m}}}$$

Similar reasoning leads to the following lower bound:

$$d_m > pn + x\sqrt{pqn} - w(n)\sqrt{\frac{pqn}{m \log \frac{n}{m}}}$$

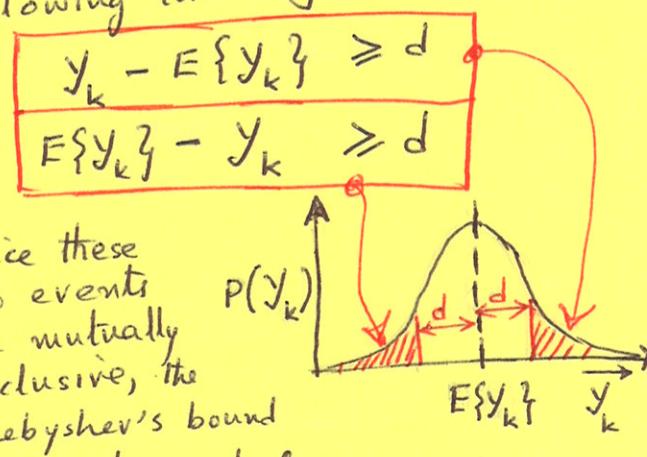
Chebyshov's inequality:

For any real valued random variable X with mean μ and variance σ^2 , the following is always true $P(|X - \mu| \geq d) \leq \frac{\sigma^2}{d^2}$

In our case:

$$P(|y_k - E\{y_k\}| \geq d) \leq \frac{\sigma_{y_k}^2}{d^2}$$

The probability space event $|y_k - E\{y_k\}| \geq d$ is a union of the following two disjoint events:



Since these two events are mutually exclusive, the Chebyshov's bound applies to each:

$$P((y_k - E\{y_k\}) \geq d) \leq \frac{\sigma_{y_k}^2}{d^2}$$

$$\& P((E\{y_k\} - y_k) \geq d) \leq \frac{\sigma_{y_k}^2}{d^2}$$

Since the mean and the variance are the same for Poisson random variables, we can write the above as

$$P((y_k - E\{y_k\}) \geq d) \leq \frac{E\{y_k\}}{d^2} \quad \text{⑧}$$

$$\& P((E\{y_k\} - y_k) \geq d) \leq \frac{E\{y_k\}}{d^2} \quad \text{⑨}$$

- The next theorem demonstrates that unless p is close to 0 or 1, the first several values of a degree sequence are **strictly decreasing** for the graphs in $G(n, p)$.

THEOREM (Jumps in Degree Sequence) :

The first several elements of a degree sequence exhibit the following property :

$$d_i - d_{i+1} \geq \frac{\alpha(n)}{m^2} \sqrt{\frac{pqn}{\log n}} \quad \text{for every } i < m$$

where the index m obeys the following constraints

$$m = o\left(\frac{pqn}{\log n}\right)^{1/4}$$

and with $\alpha(n) \rightarrow 0$ as $n \rightarrow \infty$.

As you will see in the BES algorithm described later, we can make $\alpha(n)$ be dependent on m to control the size of the jumps.

PROOF : In order to control the relationship between m and n as $n \rightarrow \infty$, we will again define an "anchor" variable x , as in the previous theorem, through $1 - \Phi(x) = \frac{m}{n}$. As seen previously, this gives us

$$x \approx \sqrt{2 \log \frac{n}{m}}$$

As $n \rightarrow \infty$, we also require that $x \rightarrow \infty$. With x defined in this manner, we know from the previous theorem that the degree d_m of a vertex m is lower-bounded as shown below :

$$d_m > pn + \sqrt{pqn}(x - \varepsilon)$$

$$\text{with } \varepsilon = \frac{w(n)}{\sqrt{m \log \frac{n}{m}}}$$

We will reexpress the lower-bound as

$$d_m \geq K$$

$$\text{with } K = [pn + \sqrt{pqn}(x - \varepsilon)]$$

- We now make the claim that for any given vertex labeled u in the degree sequence, we will **NOT** be able to find an ordered vertex pair (u, v) such that

$$\begin{array}{l} d_u \geq K \\ d_v \geq d_u \\ d_v - d_u \leq J - 1 \end{array}$$

$$\text{with } J = \frac{\alpha(n)}{m^2} \sqrt{\frac{pqn}{\log n}}$$

If we can prove this claim, we will have proved the theorem (subject to our articulating the constraint on m as in the statement of the theorem).

- In order to establish the above claim, we will determine the probability of finding an ordered vertex pair (u, v) such that $d_u \geq K$, $d_v \geq d_u$, and

and $d_v - d_u \leq J-1$. As explained in the sidebar on the right, this probability is given by

$$p \sum_{j=K-1}^{n-2} b(j; n-2, p) \sum_{l=j}^{J-1} b(l; n-2, p) + q \sum_{j=K}^{n-2} b(j; n-2, p) \sum_{l=j}^{J-1} b(l; n-2, p)$$

We will now use L to denote the expected number of such ordered pairs of vertices. L must equal the total number of ordered pairs of vertices, which is $n(n-1)$, and the probability of a pair being of the desired kind. Therefore,

$$L = n(n-1) \left[p \sum_{j=K-1}^{n-2} b(j; n-2, p) \sum_{l=j}^{J-1} b(l; n-2, p) + q \sum_{j=K}^{n-2} b(j; n-2, p) \sum_{l=j}^{J-1} b(l; n-2, p) \right]$$

The rest of the proof consists of showing that $L \rightarrow 0$ as $n \rightarrow \infty$ assuming that m obeys the constraints specified by the theorem. However, before we can show this, the expression for L must be simplified considerably:

- ① In the summation $\sum_{l=j}^{J-1} b(l; n-2, p)$ that shows up in both components of L , the index l is lower-bounded by j , but j is lower-bounded by $K-1$. Now, $K = pn + \sqrt{pnq}(x-\epsilon)$. Since we have much latitude in choosing $w(n)$, we will assume $x \geq \epsilon$ always. [This requires that $\sqrt{2 \log \frac{n}{m}} \geq \frac{w(n)}{\sqrt{m \log n}}$, which translates into $\sqrt{2m \log \frac{n}{m}} \geq w(n)$ — something that we can easily guarantee since $w(n)$ is supposed to rise arbitrarily slowly as $n \rightarrow \infty$.] So, we have $K \geq pn$. Since $K \geq pn$ implies $l \geq pn$ in $\sum_{l=j}^{J-1} b(l; n-2, p)$, we can safely say that every binomial in this summation will be less than $b(j; n-2, p)$. [Recall that, as $n \rightarrow \infty$, $b(l; n-2, p)$ is unimodal with its mode (the point where the probability distribution peaks) at $\lfloor (n-2+1)p \rfloor$.]
- ② Therefore, the expression for L can be bounded by:

Estimating the probability that a pair of randomly selected vertices constitutes an ordered pair (u, v) such that $d_u \geq K$, $d_v \geq d_u$, and $d_v - d_u \leq J-1$:

As we populate an n -vertex graph with edges, an ordered pair (u, v) with the desired properties can come into existence through the following two mechanisms:

- ① For u we choose a vertex whose current degree is at least $K-1$. (As to why $K-1$, that will become clear shortly.) Let's say we have chosen a vertex u of degree $j \geq K-1$. For every such u , we will choose a vertex v that is not yet connected to u and whose degree is at least j and at most $J-1$. For a given u of degree j , the probability of finding such a v is $\sum_{l=j}^{J-1} b(l; n-2, p)$. Note $n-2$ for the number of Bernoulli trials. That is because we do not want these trials to include u . This probability is for a given u . Since u is allowed to be any vertex whose degree is at least $K-1$, the probability of finding a vertex pair in the state described here is $\sum_{j=K-1}^{n-2} b(j; n-2, p) \sum_{l=j}^{J-1} b(l; n-2, p)$.

When we take such vertex pairs and join the two vertices (an operation accomplished with probability p), we end up with vertices u of degree at least K and with vertices v of degree that is no more than $J-1$ greater than that of u . The probability of finding a randomly chosen vertex pair (u, v) in the resulting state is:

$$p \sum_{j=K-1}^{n-2} b(j; n-2, p) \sum_{l=j}^{J-1} b(l; n-2, p) \quad (M)$$

- ② For the second mechanism, for u we choose a vertex whose degree is at least K . (We now allow for the degree of u to be at least K right from the beginning because we will not be connecting this u to its corresponding v later on.) Let's say we have chosen a vertex u of degree $j \geq K$. For every such u , we will choose a vertex v whose degree is at least j and at most $J-1$. The probability of finding a randomly selected v in such a state is $\sum_{l=j}^{J-1} b(l; n-2, p)$ where we use only $n-2$ Bernoulli trials because such a v is not allowed to connect with the chosen u . This probability is for a given u . Since u is allowed to be any vertex whose degree is

$$L \leq \left\{ p \sum_{j=K-1}^{n-2} b(j; n-2, p) J b(j; n-2, p) + q \sum_{j=K}^{n-2} b(j; n-2, p) J b(j; n-2, p) \right\}$$

The contribution made by the second summation to the bound will not be violated if we change the lower limit from $j=K$ to $j=K-1$. Since $p+q=1$, with this change the two summations collapse into one, to yield

$$L \leq n^2 J \sum_{j=K-1}^{n-2} b(j; n-2, p) b(j; n-2, p)$$

Although the two terms in the product on the right are identical, we will treat them differently in order to show that

$L \rightarrow 0$ as $n \rightarrow \infty$. Since all

$b(j; n-2, p)$ are less than $b(K-1; n-2, p)$ for reasons explained earlier, the bound for L can be expressed as

$$L \leq \left[n^2 J \sum_{j=K-1}^{n-2} b(j; n-2, p) \cdot b(K-1; n-2, p) \right] = n^2 J B(K-1; n-2, p) \cdot b(K-1; n-2, p)$$

where we used B in the second form of the bound as defined earlier when we first talked about y_k and Z_k being Poisson. The bound for L can be simplified further by using the Gaussian approximation for $b(K-1; n-2, p)$

[this approximation says that as $n \rightarrow \infty$, the binomial distribution looks more and more like a Gaussian distribution with mean pn and variance npq] and the DeMoivre-Laplace bound for $B(K-1; n-2, p)$. The result is

$$L \leq n^2 J \frac{x}{\sqrt{n} \sqrt{npq}} \left(\frac{m}{n} \right)^{3-2\sqrt{2}}$$

If we substitute in this the value of J shown near the beginning of this proof, we get

$$L \leq \alpha(n) \frac{x}{\sqrt{\log n}} \left(\frac{n}{m} \right)^{2\sqrt{2}-1}$$

The final step of the proof consists of demonstrating that this upper bound goes to 0 for $m = o\left(\frac{pn}{\log n}\right)^{1/4}$.

A Fast Algorithm for Graph Isomorphism :

This algorithm, presented by Babai, Erdős, and Selkow (BES) in 1980, is based on the following properties of random graphs :

- 1) The first several elements of a degree sequence are monotonically decreasing
- 2) The smallest jump one is likely to see between the successive elements of a degree sequence depends on how many of the beginning elements we wish to examine.
- 3) The BES algorithm selects for m , the number of the starting elements of a degree sequence to examine as :

at least K , the probability of finding a vertex pair (u, v) in the state described here is 44-9

$\sum_{j=K}^{n-2} b(j; n-2, p) \sum_{l=j}^{j+J-1} b(l; n-2, p)$. Now we must guarantee that subsequently there will not come into existence an edge that connects u with v . We can supply this guarantee with probability q . Therefore, the probability of finding a randomly selected vertex pair in the state that can be attributed to the second mechanism is

$$q \sum_{j=K}^{n-2} b(j; n-2, p) \sum_{l=j}^{j+J-1} b(l; n-2, p)$$
N

Therefore, the overall probability of finding a randomly chosen vertex pair with the desired properties is the sum of the probabilities at M and N.

$$m = \lceil 3 \log_{1/q} n \rceil$$

$$= \lceil -\frac{3 \log n}{\log q} \rceil$$

$$\log_{1/q} n = \frac{\log n}{\log \frac{1}{q}}$$

- 4) With m set as above, the BES algorithm sets the value of $\alpha(n)$ that appears in the statement of the "Jump" theorem on page 44-7 to

$$\alpha(n) = 3m^2 \sqrt{\frac{\log n}{\log q}}$$

this satisfies the required condition that as $n \rightarrow \infty$, we want $\alpha(n) \rightarrow 0$

- 5) With m and $\alpha(n)$ set as above, we can show that the jumps between the successive elements of the degree sequence for its first m elements will obey the inequalities:

$$\begin{aligned} d_1 &\geq d_2 + 3 \\ d_2 &\geq d_3 + 3 \\ \vdots &\quad \vdots \\ d_m &\geq d_{m+1} + 3 \end{aligned}$$

- 6) The first m elements of the degree sequence play a key role in the BES algorithm. It obviously will NOT do to compare two graphs on the basis of just these m vertex degrees. Two graphs may exhibit very similar degree sequences but very different connectivity patterns. Crucial to establishing isomorphism is to make certain that the connectivity patterns in the two graphs are the same. How the similarities between two connectivity patterns is established by the BES algorithm is explained in what follows.

- With the "understanding" conveyed by the above six "points", we will now present the following key step of the BES algorithm. This step will calculate a "connectivity weight" at each of the n vertices of an n -vertex graph on the basis of how that vertex is connected to the m vertices that correspond to the first m elements of the degree sequence. We will use $f(x_i)$ to denote the connectivity weight at the vertex labeled x_i . We assume that after the vertex degrees are calculated at all n vertices, the vertices are labeled as

$$x_1, x_2, \dots, x_n$$

and their vertex degrees designated

$$d_1, d_2, \dots, d_n$$

with the understanding that $d_1 \geq d_2 \geq \dots \geq d_n$.

- The connectivity weight $f(x_i)$ at vertex x_i is calculated as

$$f(x_i) = \sum_{j=1}^m a(i,j) 2^j \quad i = 1, 2, \dots, n$$

where $a(i,j) = 1$ if $\{x_i, x_j\} \in E(G)$ and 0 otherwise.

- You can think of $f(x_i)$ as m-bit binary words we associate with each vertex of an n-vertex graph, the binary word being an encoding of how that vertex is connected with the m vertices that correspond to the first m elements of the degree sequence.
- One can show theoretically that no two vertices of a graph G in $G(n,p)$ will have the same value for $f(x_i)$ as $n \rightarrow \infty$. This point is explained below in the sidebar.
- The BES algorithm re-labels the n vertices in the descending order of the connectivity weights.
(The descending order can be expected to be strict since $f(x)$ is unique for every vertex.)
- The relabeling of the vertices can be considered to be a permutation π_G of the vertices under which we list the vertices as

$$x_{\pi_G(1)}, x_{\pi_G(2)}, \dots, x_{\pi_G(n)}$$
 with the vertices obeying:

$$f(x_{\pi_G(1)}) > f(x_{\pi_G(2)}) \dots > f(x_{\pi_G(n)})$$
- Such relabeling of the vertices of a graph is referred to as **canonical labeling**
- Canonical labeling of a graph can be carried out in $O(pn^2)$ time. As you already know, the average degree of a vertex is pn . Calculating the degrees at all n vertices is $O(pn^2)$ activity. Finding $f(x)$ at each vertex is $O(pn)$ work. Sorting all n vertices on the basis of their $f(x)$ values takes time that is less than $O(n)$.
- If two graphs are isomorphic, then after canonical labeling they will look like the same graph.

Proof of the assertion that $f(x_i) = f(x_j)$ will occur with 0 probability as $n \rightarrow \infty$:

$f(x_i) = f(x_j)$ implies that x_i and x_j are joined to exactly the same vertices in the set $V_{ij} = \{x_1, x_2, \dots, x_m\} - \{x_i, x_j\}$

Let W_{ij} be the set of $m-2$ vertices of the highest degree in $G - \{x_i, x_j\}$. (Note that the graph $G - \{x_i, x_j\}$ will not only have the vertices x_i and x_j excised from G but also all the edges incident on these two vertices.) We now note the following three aspects of the relationship between V_{ij} and W_{ij} :

① $W_{ij} \subseteq V_{ij}$ The cardinality of V_{ij} may be $m-2$, $m-1$, or m depending on whether none of, only one of, or both of x_i and x_j are in the set $\{x_1, x_2, \dots, x_m\}$. On the other hand the cardinality of W_{ij} is always $m-2$. The subset relationship is further justified by the point we make below.

② For the above subset property to hold, we must also establish that all $m-2$ vertices of W_{ij} are in V_{ij} . This is a consequence of the fact that the $(m-2)$ highest degree vertices in the subgraph $G - \{x_i, x_j\}$ will maintain their order of appearance and will be contained in the sequence of m highest degree vertices in G . That is because these m vertices obey the constraint $d_i - d_{i+1} \geq 3$. So, at the most, any or all of the vertices in $\{x_1, x_2, \dots, x_m\}$ may

- So if we want to test whether two graphs G and H are isomorphic, all we need do is to re-label the vertices with the canonical labeling algorithm. The relabeled graphs must be identical if G and H are isomorphic.

44-12

lose a maximum of two edges when we remove x_i and x_j from the graph. But since the difference between the consecutive degrees is 3 or greater, a degree reduction by 2 (in the worst case) does not cause a vertex to lose its place in a degree-ordered listing of the $m-2$ highest degree vertices in $G - \{x_i, x_j\}$.

[The reason we inject W_{ij} into the proof is because W_{ij} is easier to work with than V_{ij} since the former has fixed cardinality of $m-2$.]

- (3) We can show that

$$P(f(x_i) = f(x_j)) = P(x_i \text{ and } x_j \text{ are joined to the same vertices in } V_{ij}) \leq P(x_i \text{ and } x_j \text{ are joined to the same vertices in } W_{ij})$$

The reason for the inequality shown above

is subtle but important. Recall that W_{ij} is the set of $m-2$ highest-degree vertices in the subgraph $G - \{x_i, x_j\}$. The selection of a vertex for inclusion in W_{ij} is made regardless of how it is connected with x_i and/or x_j . This can only increase the probability that x_i and x_j are connected to the same vertices in W_{ij} as compared to the $m-2$ vertices in V_{ij} . [As to why, note that the degree differences in W_{ij} can only be smaller compared to the degree differences in V_{ij} — as explained earlier. With the degrees in W_{ij} likely to be more uniform, the probability that x_i and x_j will be connected to the same vertices in W_{ij} can only be larger.]

- Bounding $P(f(x_i) = f(x_j))$ therefore boils down to estimating the probability that x_i and x_j are connected to same subset of vertices in W_{ij} . This probability can be estimated by considering : (i) x_i and x_j are both connected to a single vertex of W_{ij} ; (ii) x_i and x_j are connected to exactly same two vertices in W_{ij} ; and so on. This gives us

$$P(x_i \text{ and } x_j \text{ are connected to the same vertices in } W_{ij})$$

$$= (m-2)p^2q^{2m-6} + \binom{m-2}{2}p^4q^{2m-8} + \dots + \binom{m-2}{m-2}p^{2m-4}$$

$$= (p^2 + q^2)^{m-2} - q^{2m-4} \leq (p^2 + q^2)^{m-2} = q^{m-2} \left(\frac{1-2pq}{q}\right)^{m-2} \leq q^{m-2}$$

The last inequality follows from the fact that $1-2pq \leq q$ when $\frac{1}{2} \leq q \leq 1$. (The BES algorithm as presented is only for the case $0 < p < \frac{1}{2}$. When that is not the case, we can always work with the complement of the graph.)

- We can show that $q^{m-2} = O(n^{-3})$. This follows from $m \approx -3 \frac{\log n}{\log q}$
- That completes the proof of the assertion that as $n \rightarrow \infty$, $f(x)$ will take on unique values for every vertex in a graph.