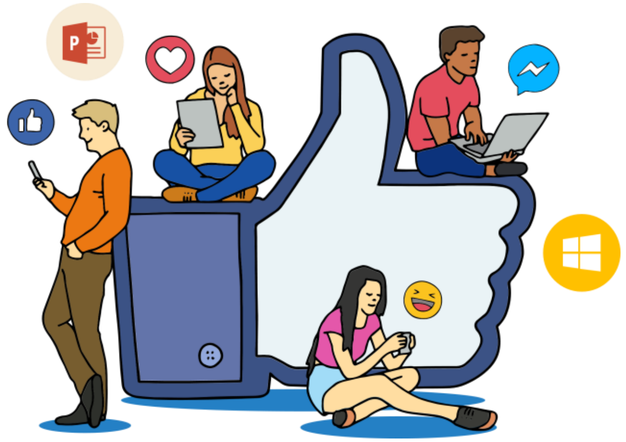
INFS7450 SOCIAL MEDIA ANALYTICS
Tutorial Week 6

School of ITEE

The University of Queensland



Outlines

- Quiz 3
- Knowledge Extension to Lecture 6
- Code Demo
- Q&A

Section 1: Quiz 3

Online quiz 3: Q1

1. Multiple Choice: Which of the following is correct?

Question Which of the following is correct?

Answer

A. When the frequency of an event changes as a power of an attribute, the frequency follows a power-law.

B. In a power-law distribution, small occurrences is common, and large instances is extremely rare

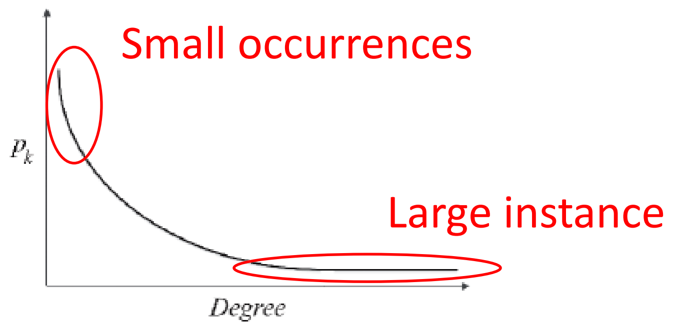
C. Most real-world social networks follow power-law distributions, which are also called scale-free networks.

D. A, B, C are all correct.

Correct Feedback

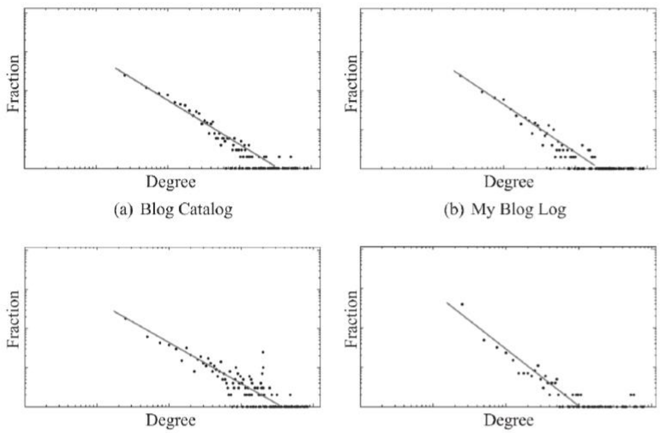
Incorrect Feedback

A typical shape of a power-law distribution



(a) Power-Law Degree Distribution

Real-world social networks follow power-law distributions (called **Scale-Free** networks)



(a) Blog Catalog (b) My Blog Log (c) Twitter (d) My Space

Power-Law Degree Distribution

- When the frequency of an event changes as a power of an attribute
  - The frequency follows a **power-law**

Power-law intercept

The power-law exponent and its value is typically in the range of **[2, 3]**

$$p_d = a d^{-b}$$

Fraction of users with degree d

Node degree

$$\boxed{\ln p_d} = -b \boxed{\ln d} + \ln a$$

online quiz 3: Q2

Question	To test whether a network exhibits a power-law distribution, we should: 1, Plot a log-log graph, where the x-axis represents lnk and the y-axis represents lnp_k 2, Pick a popularity measure and compute it for the whole network 3, If a power-law distribution exists, we should observe a straight line 4, Compute the fraction of individuals having popularity k.
Answer	A. 1,2,3,4 B. 2,4,1,3 C. 2,3,1,4 D. 3,1,2,4
Correct Feedback	you are correct
Incorrect Feedback	B is correct

To test whether a network exhibits a power-law distribution

1. Pick a popularity measure and compute it for the whole network
   – Example: number of friends for all nodes
2. Compute p_k, the fraction of individuals having popularity k.
3. Plot a log-log graph, where the x-axis represents $\ln k$ and the y-axis represents $\ln p_k$.
4. If a power-law distribution exists, we should observe a straight line

This is not a systematic approach!

1. Other distributions could also exhibit this pattern

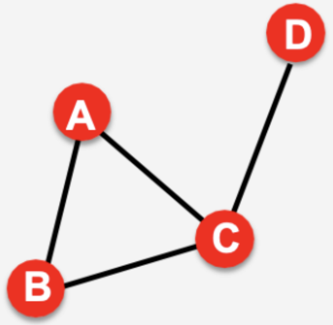
For a systematic approach see:

Clauset, Aaron, Cosma Rohilla Shalizi, and Mark EJ Newman. "Power-law distributions in empirical data." *SIAM review 51(4)* (2009): 661-703.

online quiz 3: Q3

Question

Given the following graph, the diameter of this graph is:



Answer

A. 0

B. 1

C. 2

D. 3

Correct Feedback

you are correct

Incorrect Feedback

C is correct

- **Diameter:** The maximum (shortest path) distance between any pair of nodes in a graph

Shortest paths:

(A,B):1

(A,C):1

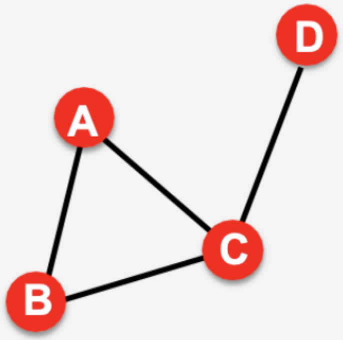
(A,D):2

(B,C):1

(B,D):2

(C,D):1

online quiz 3: Q4

Question	Given the following graph, the average shortest path length is:
	
Answer	A. 2/3
	✔ B. 4/3
	C. 1
	D. A,B,C are all wrong
Correct Feedback	you are correct
Incorrect Feedback	B is correct

- **Average shortest path length** for a connected graph (component) or a strongly connected (component of a) directed graph

$$\bar{h} = \frac{1}{2E_{\max}} \sum_{i,j \neq i} h_{ij}$$

where h_{ij} is the distance from node i to node j

E_{max} is max number of edges (total number of node pairs) $= n(n-1)/2$

Shortest paths:

(A,B):1 (B,A):1

(A,C):1 (C,A):1

(A,D):2 (D,A):2

(B,C):1 (C,B):1

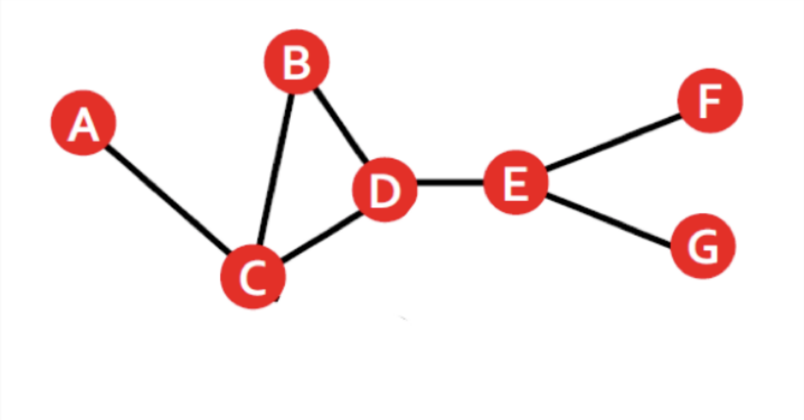
(B,D):2 (D,B):2

(C,D):1 (D,C):1

$$E_{max} = \frac{n(n-1)}{2} = \frac{4 \times (4-1)}{2} = 6$$

$$\bar{h} = \frac{1}{2 \times 6} \times 16 = \frac{4}{3}$$

Online quiz 3: Q5

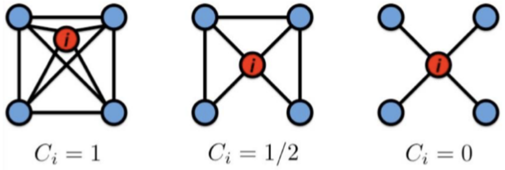
Question	What is the clustering coefficient of node C in the following graph?
	
Answer	✅ A. 1/3
	B. 1/2
	C. 1
	D. A, B, C are all wrong
Correct Feedback	you are correct
Incorrect Feedback	A is correct

- $C_i \in [0,1]$

$$C_i = \frac{2e_i}{k_i(k_i-1)}$$

where e_i is the number of edges between the neighbors of node i

$C_i = 0$ If the degree of node i is **1**.



- **Average clustering coefficient:** $C = \frac{1}{N}\sum_i^N C_i$

(B,D)

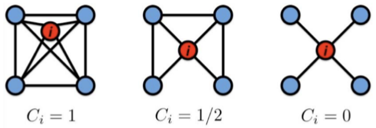
$$C_i = \frac{2e_i}{k_i(k_i-1)} = \frac{2 \times 1}{3(3-1)} = \frac{1}{3}$$

online quiz 3: Q6

- $C_i \in [0,1]$

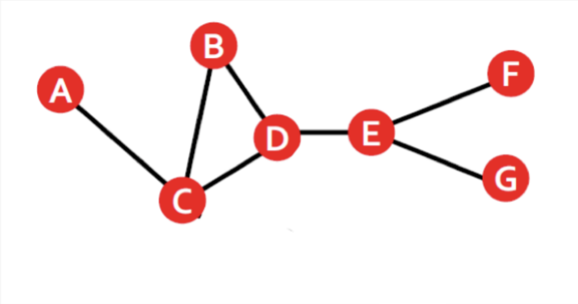
$$C_i = \frac{2e_i}{k_i(k_i-1)}$$ where e_i is the number of edges between the neighbors of node i

$C_i = 0$ If the degree of node i is **1**.



- **Average clustering coefficient:** $C = \frac{1}{N}\sum_i^N C_i$

Question	Which of the following values is closest to the average clustering coefficient of the following graph?
Answer	A. 0.33 B. 1 C. 0.24 (correct) D. 1/5
Correct Feedback	you are correct
Incorrect Feedback	C is correct



$$C_A = \frac{2e_i}{k_i(k_i-1)} = 0$$

$$C_B = \frac{2e_i}{k_i(k_i-1)} = \frac{2\times 1}{2(2-1)} = 1$$

$$C_C = \frac{2e_i}{k_i(k_i-1)} = \frac{2\times 1}{3(3-1)} = \frac{1}{3}$$

$$C_D = \frac{2e_i}{k_i(k_i-1)} = \frac{2\times 1}{3(3-1)} = \frac{1}{3}$$

$$C_E = \frac{2e_i}{k_i(k_i-1)} = \frac{2\times 0}{3(3-1)} = 0$$

$$C_F = \frac{2e_i}{k_i(k_i-1)} = 0$$

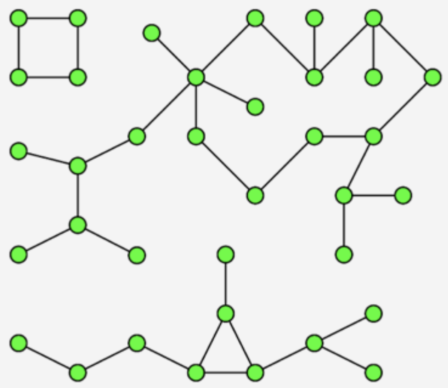
$$C_G = \frac{2e_i}{k_i(k_i-1)} = 0$$

$$C = \frac{1}{7}\left(1 + \frac{1}{3} + \frac{1}{3}\right) = 0.238$$

Online quiz 3: Q7

Question

How many components does the following graph have? and how many nodes does the giant component have?



Answer

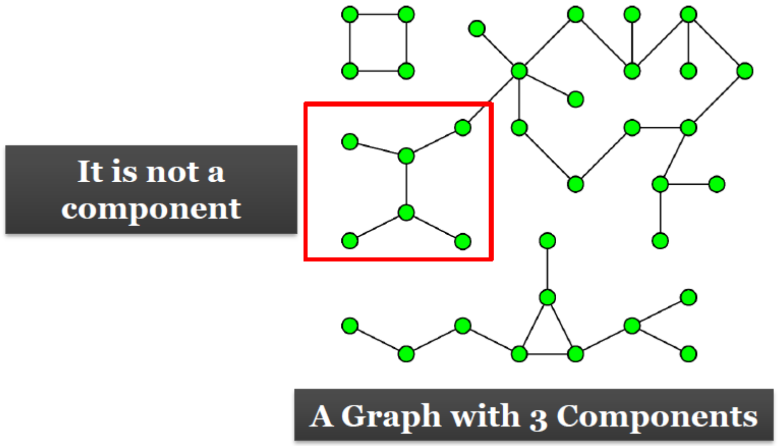
A. 3, 22

B. 3,10

C. 4,22

D. A,B,C are all wrong

- A component of an undirected graph is a subgraph
  - in which any two vertices are connected to each other by paths,
  - and which is connected to no additional vertices in the original graph.



Correct Feedback

you are correct

Incorrect Feedback

A is correct

Section 2

- In-class Quiz about random graph.
- Why is the (average) clustering coefficient of a regular lattice $\frac{3(\mathrm{c}-2)}{4(c-1)}$
- Properties of the configuration model.
- How to calculate Pearson correlation.

In-class Quiz about Random Graph

QUESTION 7

Which of the following is correct according to the random graph model?

- A. The graph is a result of a random process.
- B. The degree distribution is binomial.
- C. random graph can grow very large but nodes will be just a few hops apart.
- D. A,B,C are all correct.

QUESTION 8

Which of the following is WRONG according to the Random Graph Model?

- A. The graph is a result of a random process, and we can have many different realizations given the same n (node number) and p (edge probability).
- B. The average path length in a random graph is close to $\frac{ln|V|}{lnc}$, where V is the node set, and c is the expected degree.
- C. In random graphs, as we increase p, a large fraction of nodes start getting connected.
- D. In random graph, when $p=0$, the size of the giant component is n.

QUESTION 10

In random graphs, as we increase p (edge appears i.i.d. with probability p), a large fraction of nodes start getting connected. What's the size of the giant component when p=1?

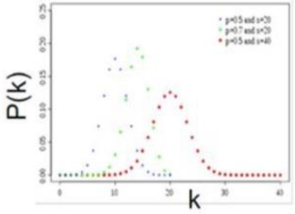
- A. the size of the giant component is 0
- B. the size of the giant component is n. n is the number of nodes.
- C. the size of the giant component is between 0-n. n is the number of nodes.
- D. the size is uncertain

Exercise Question

Question	Which of the following is correct according to the random graph model?
Answer	A. The graph is a result of a random process. B. The degree distribution is binomial. C. random graph can grow very large but nodes will be just a few hops apart. ✅ D. A,B,C are all correct.
Correct Feedback	you are correct
Incorrect Feedback	D is correct

- **Fact:** **Degree distribution of G_{np} is binomial.**
- Let $P(k)$ denote the fraction of nodes with degree k:

$$P(k) = \underbrace{\binom{n-1}{k}}_{\text{Select } k \text{ nodes out of } n-1} \underbrace{p^k}_{\text{Probability of having } k \text{ edges}} \underbrace{(1-p)^{n-1-k}}_{\text{Probability of missing the rest of the } n-1-k \text{ edges}}$$



Mean, variance of a binomial distribution

$$c = \bar{k} = p(n-1)$$

$$\sigma^2 = p(1-p)(n-1)$$

https://en.wikipedia.org/wiki/Binomial_distribution

Exercise Question

Question	Which of the following is WRONG according to the Random Graph Model?
Answer	A. The graph is a result of a random process, and we can have many different realizations given the same n (node number) and p (edge probability). B. The average path length in a random graph is close to $\frac{ln\|V\|}{lnc}$, where V is the node set, and c is the expected degree. C. In random graphs, as we increase p, a large fraction of nodes start getting connected. D. In random graph, when $p=0$, the size of the giant component is n.
Correct Feedback	you are correct
Incorrect Feedback	D is correct

- In random graphs:
  - $p = 0$
    - the size of the giant component is 0
  - $p = 1$
    - the size of the giant component is n

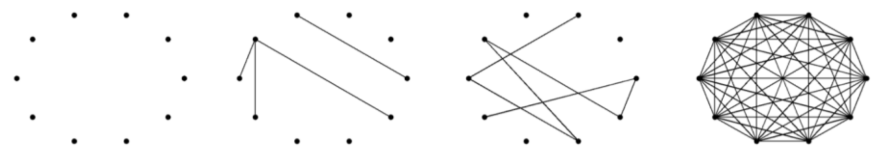
The average path length in a random graph is

$$h \approx \frac{\ln |V|}{\ln c}$$

Proof

Exercise Question

Question	In random graphs, as we increase p (edge appears i.i.d. with probability p), a large fraction of nodes start getting connected. What's the size of the giant component when p=1?
Answer	A. the size of the giant component is 0 B. the size of the giant component is n. n is the number of nodes. C. the size of the giant component is between 0-n. n is the number of nodes. D. the size is uncertain
Correct Feedback	you are correct
Incorrect Feedback	B is correct



Probability (p)	0.0	0.088	0.11	1.0
Average Node Degree (c)	0.0	0.8	≈1	n-1=9
Diameter	0	2	6	1
Giant Component Size	0	4	7	10
Average Path Length	0.0	1.5	2.66	1.0

- The phase transition we focus on happens when
  - average node degree $c = 1$ (or when $p = 1/(n-1)$)

Regular Lattice

- In real-world interactions, many individuals have a limited and often at least, a fixed number of connections

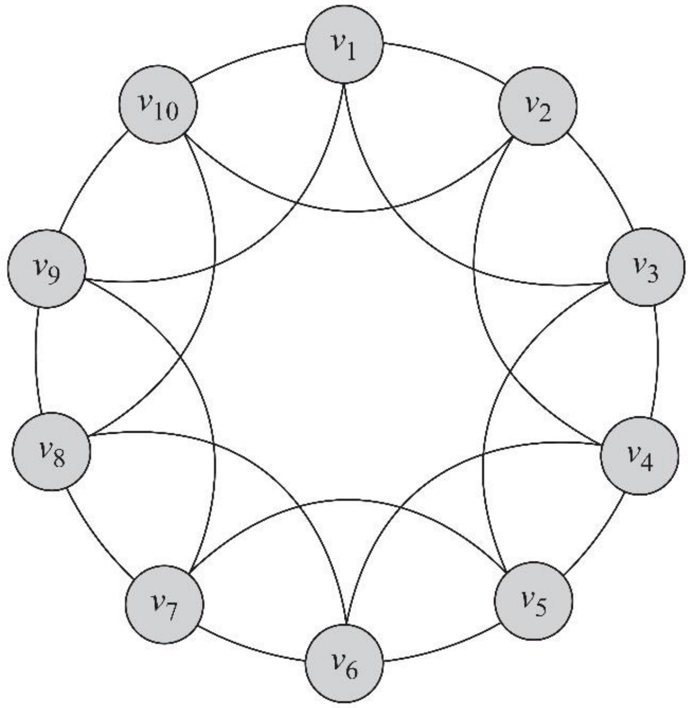
- In graph theory terms, this assumption is equivalent to embedding users in a regular network

- A regular (ring) lattice is a special case of regular networks where there exists a certain pattern on how ordered nodes are connected to one another

- In a regular lattice of degree c, nodes are connected to their previous $c/2$ and following $c/2$ neighbors

- Formally, for node set V={v_1, v_2, v_3,…, v_n}, an edge exists between node i and j if and only if

$$0 \leq \min(n - |i - j|, |i - j|) \leq c/2$$



Homework

- Why is the (average) clustering coefficient of a regular lattice? Try to prove it.

$$\frac{3(c-2)}{4(c-1)}$$

If we want to calculate the clustering coefficient of node v_i, we first list its neighbor nodes.

$$\underbrace{v_{i-\frac{c}{2}}, v_{i-\frac{c}{2}+1} \dots, v_{i-2}, v_{i-1}}_{\text{Previous c/2 neighbors}}, v_i, \underbrace{v_{i+1}, v_{i+2}, \dots, v_{i+\frac{c}{2}-1}, v_{i+\frac{c}{2}}}_{\text{Following c/2 neighbors}}$$

$$C_i = \frac{2e_i}{k_i(k_i-1)}$$

where e_i is the number of edges between the neighbors of node i

We start calculate e_i from the previous c/2 neighbors

We calculate the number of common neighbor nodes between node v_i and node $v_{i-\frac{c}{2}}$

$$v_{i-\frac{c}{2}}, \underbrace{\boxed{v_{i-\frac{c}{2}+1} \dots, v_{i-2}, v_{i-1}, v_i}}_{\frac{c}{2}-1}, v_{i+1}, v_{i+2}, \dots, v_{i+\frac{c}{2}-1}, v_{i+\frac{c}{2}}$$

$$v_{i-\frac{c}{2}}, v_{i-\frac{c}{2}+1} \underbrace{\boxed{\dots, v_{i-2}, v_{i-1}, v_i, v_{i+1}}}_{\frac{c}{2}-1}, v_{i+2}, \dots, v_{i+\frac{c}{2}-1}, v_{i+\frac{c}{2}}$$

$$v_{i-\frac{c}{2}}, v_{i-\frac{c}{2}+1} \dots, v_{i-2}, v_{i-1}, \boxed{v_i, \underbrace{v_{i+1}, v_{i+2}, \dots, v_{i+\frac{c}{2}-1}, v_{i+\frac{c}{2}}}_{\frac{c}{2}-1}}$$

$$\underbrace{v_{i-\frac{c}{2}}, v_{i-\frac{c}{2}+1} \dots, v_{i-2}, v_{i-1}}_{\text{Previous c/2 neighbors}}, v_i, \underbrace{v_{i+1}, v_{i+2}, \dots, v_{i+\frac{c}{2}-1}, v_{i+\frac{c}{2}}}_{\text{Following c/2 neighbors}}$$

Previous c/2 neighbors part

following c/2 neighbors part

$$e_i = \boxed{\frac{c}{2}\left(\frac{c}{2}-1\right)} + \boxed{\left(\frac{c}{2}-1\right)\frac{c}{4}}$$
$$= \frac{3(c^2-2c)}{8}$$

$$v_{i-\frac{c}{2}}, v_{i-\frac{c}{2}+1} \dots, v_{i-2}, v_{i-1}, v_i, v_{i+1}, \boxed{\underbrace{v_{i+2}, \dots, v_{i+\frac{c}{2}-1}, v_{i+\frac{c}{2}}}_{\frac{c}{2}-1}, v_{i+\frac{c}{2}+1}}$$

$$C = \frac{2e_i}{c(c-1)}$$
$$= 2 \times \frac{3(c^2-2c)}{8} \times \frac{1}{c(c-1)}$$
$$= \frac{3(c-2)}{4(c-1)}$$

$$v_{i-\frac{c}{2}}, v_{i-\frac{c}{2}+1} \dots, v_{i-2}, v_{i-1}, v_i, v_{i+1}, \underbrace{v_{i+2}, \boxed{\dots, v_{i+\frac{c}{2}-1}, v_{i+\frac{c}{2}}}}_{\frac{c}{2}-2}\boxed{, v_{i+\frac{c}{2}+1}, v_{i+\frac{c}{2}+2}}$$

$$v_{i-\frac{c}{2}}, v_{i-\frac{c}{2}+1} \dots, v_{i-2}, v_{i-1}, v_i, v_{i+1}, v_{i+2}, \dots, v_{i+\frac{c}{2}-1}, v_{i+\frac{c}{2}}, \boxed{\underbrace{v_{i+\frac{c}{2}+1}, v_{i+\frac{c}{2}+2}, \dots}_{0}}$$

Properties of the configuration model.

Properties of the Configuration Model

The probability that node v_i gets connected to node v_j is approximately

$$\frac{d_i d_j}{2m}$$

Proof:

In the shuffled list, for each v_i instance:

- There are d_j instances of v_j that it could be next to
- The probability of being next to v_j is $\frac{d_j}{2m-1}$

- There are d_i instances of v_i ; therefore, the total probability is $(d_i d_j)/(2m-1) \approx (d_i d_j)/2m$

Properties of the configuration model.

How to generate a Configuration model

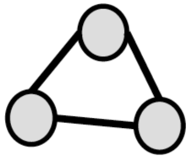
1. Create a list where each node v_i with degree d_i is repeated d_i times
2. Shuffle the list
3. Starting from the first index, join adjacent nodes

Example: Degree sequence (2,2,2)

v_1	v_1	v_2	v_2	v_3	v_3

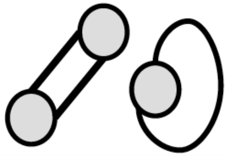
Random Shuffle 1:

v_1	v_2	v_2	v_3	v_3	v_1



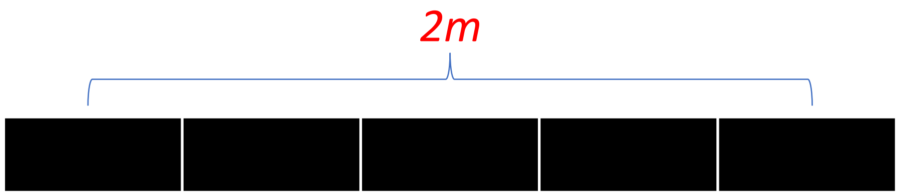
Random Shuffle 2:

v_1	v_1	v_2	v_3	v_3	v_2



Properties of the configuration model.

1. Assume that there are *m* edges in the graph, so the total number of degree is *2m*



– Handshake Theorem: Let G=(V,E) be an undirected graph with m edges. Then

$$2m = \sum_{v \in V} \deg(v)$$

Proof: Each edge contributes twice to the total degree count of all vertices. Thus, both sides of the equation equal to twice the number of edges.

Properties of the Configuration Model

The probability that node v_i gets connected to node v_i is approximately

$$\frac{d_i d_j}{2m}$$

Proof:

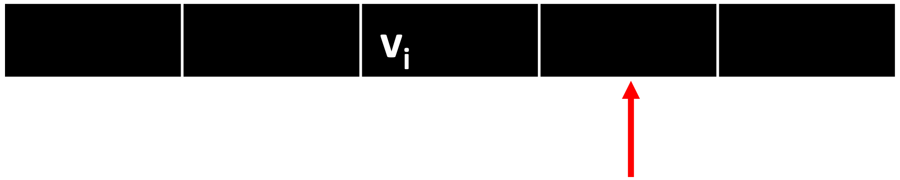
In the shuffled list, for each v_i instance:

- There are d_j instances of v_j that it could be next to
- The probability of being next to v_j is $\frac{d_j}{2m-1}$

- There are d_i instances of v_i ; therefore, the total probability is $(d_i d_j)/(2m-1) \approx (d_i d_j)/2m$

Properties of the configuration model.

1. Assume that there are *m* edges in the graph, so the total number of degree is *2m*



2. The number of all possible values that this position can take is $\frac{1}{2m-1}$.

Properties of the Configuration Model

The probability that node v_i gets connected to node v_i is approximately

$$\frac{d_i d_j}{2m}$$

Proof:

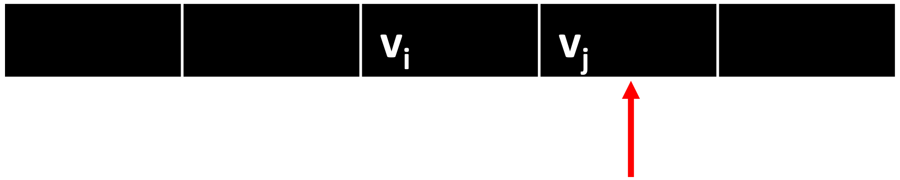
In the shuffled list, for each v_i instance:

- There are d_j instances of v_j that it could be next to
- The probability of being next to v_j is $\frac{d_j}{2m-1}$

- There are d_i instances of v_i ; therefore, the total probability is $(d_i d_j)/(2m-1) \approx (d_i d_j)/2m$

Properties of the configuration model.

1. Assume that there are *m* edges in the graph, so the total number of degree is *2m*



2. **The number of all possible values that this position can take is $\frac{1}{2m-1}$.**

3. **The probability of being next to v_j is $\frac{d_j}{2m-1}$.**

Properties of the Configuration Model

The probability that node v_i gets connected to node v_i is approximately

$$\frac{d_i d_j}{2m}$$

Proof:

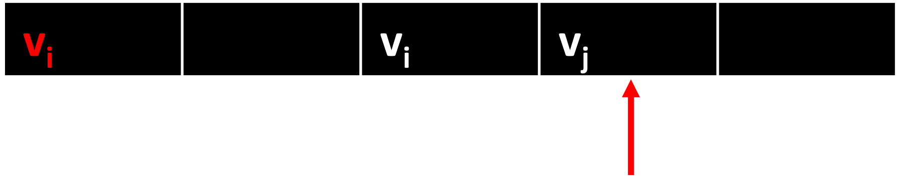
In the shuffled list, for each v_i instance:

- There are d_j instances of v_j that it could be next to
- The probability of being next to v_j is $\frac{d_j}{2m-1}$

- There are d_i instances of v_i ; therefore, the total probability is $(d_i d_j)/(2m-1) \approx (d_i d_j)/2m$

Properties of the configuration model.

1. Assume that there are *m* edges in the graph, so the total number of degree is *2m*



2. The number of all possible values that this position can take is $\frac{1}{2m-1}$.

3. The probability of being next to v_j is $\frac{d_j}{2m-1}$.

4. There are d_i instances of v_i. Therefore, the total probability is $\frac{d_i d_j}{2m-1}$.

Properties of the Configuration Model

The probability that node v_i gets connected to node v_i is approximately

$$\frac{d_i d_j}{2m}$$

Proof:

In the shuffled list, for each v_i instance:

– There are d_j instances of v_j that it could be next to

– The probability of being next to v_j is $\frac{d_j}{2m-1}$

- There are d_i instances of v_i ; therefore, the total probability is $(d_i d_j)/(2m-1) \approx (d_i d_j)/2m$

Pearson correlation $\rho(X,Y)$ is the normalized version of covariance

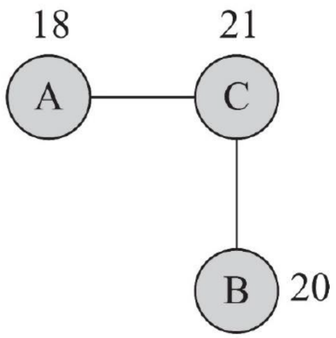
$$\rho(X_L, X_R) = \frac{\sigma(X_L, X_R)}{\sigma(X_L)\sigma(X_R)}.$$

In our case: $\sigma(X_L) = \sigma(X_R)$

Standard deviation

$$\sigma_X^2 = \mathrm{E}[(X - \mathrm{E}[X])^2] = \mathrm{E}[X^2] - [\mathrm{E}[X]]^2$$

Variance



$$X_L = \begin{bmatrix} 18 \\ 21 \\ 21 \\ 20 \end{bmatrix} \qquad X_R = \begin{bmatrix} 21 \\ 18 \\ 20 \\ 21 \end{bmatrix}$$

How to calculate it?

$$\boxed{\rho(X_L, X_R) = -0.67}$$

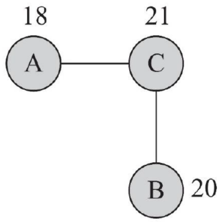
Pearson correlation $\rho(X,Y)$ is the normalized version of covariance

$$\rho(X_L, X_R) = \frac{\sigma(X_L, X_R)}{\sigma(X_L)\sigma(X_R)}.$$

$$\sigma(X_L, X_R) = \mathbf{E}[X_L X_R] - \mathbf{E}[X_L]\mathbf{E}[X_R]$$
$$= \frac{\sum_{ij} A_{ij} x_i x_j}{2m} - \frac{\sum_{ij} d_i d_j x_i x_j}{(2m)^2}$$
$$= \frac{1}{2m} \boxed{\sum_{ij} \left(A_{ij} - \frac{d_i d_j}{2m} \right)} x_i x_j$$

In our case: $\sigma(X_L) = \sigma(X_R)$ $\quad \sigma_X^2 = \mathrm{E}[(X - \mathrm{E}[X])^2] = \mathrm{E}[X^2] - [\mathrm{E}[X]]^2$

Standard deviation **Variance**



$$X_L = \begin{bmatrix} 18 \\ 21 \\ 21 \\ 20 \end{bmatrix} \qquad X_R = \begin{bmatrix} 21 \\ 18 \\ 20 \\ 21 \end{bmatrix}$$

$$\rho(X_L, X_R) = -0.67$$

In [6]:
```
A = [[0,0,1],
     [0,0,1],
     [1,1,0]]
d = [1,1,2]
x = [18,20,21]
m = 2

def cal_cov(A,d,x,m):
    value = 0
    for i in range(len(A)):
        for j in range(len(A[i])):
            value += (A[i][j]-(d[i]*d[j])/(2*m))*x[i]*x[j]
    value = value/(2*m)

    return value

cov = cal_cov(A,d,x,m)
print(cov)
```

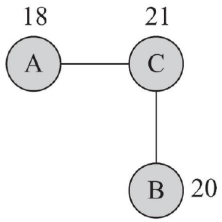
```
-1.0
```

Pearson correlation $\rho(X, Y)$ is the normalized version of covariance

$$\rho(X_L, X_R) = \frac{\sigma(X_L, X_R)}{\sigma(X_L)\sigma(X_R)}.$$

In our case: $\sigma(X_L) = \sigma(X_R)$ $\quad \sigma_X^2 = \mathrm{E}[(X - \mathrm{E}[X])^2] = \mathrm{E}[X^2] - [\mathrm{E}[X]]^2$

Standard deviation Variance



$$X_L = \begin{bmatrix} 18 \\ 21 \\ 21 \\ 20 \end{bmatrix} \qquad X_R = \begin{bmatrix} 21 \\ 18 \\ 20 \\ 21 \end{bmatrix}$$

$$\rho(X_L, X_R) = -0.67$$

$$E(X_L) = E(X_R) = \frac{\sum_i (X_L)_i}{2m} = \frac{\sum_i d_i x_i}{2m}$$

$$\sigma_X^2 = \mathrm{E}[(X - \mathrm{E}[X])^2] = \mathrm{E}[X^2] - [\mathrm{E}[X]]^2$$

Variance

```
x_l = [18,21,20]
d_l = [1,2,1]
m=2
def cal_e_xl_2(x_l,d_l,m):
    value = 0
    for i in range(len(x_l)):
        value += d_l[i]*x_l[i]*x_l[i]
    value = value/(2*m)
    return value
e_xl_2 = cal_e_xl_2(x_l,d_l,m)
print(e_xl_2)
```

```
401.5
```

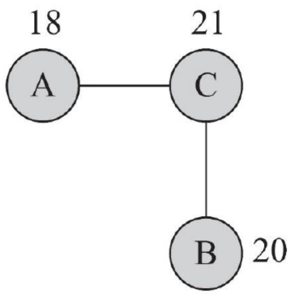
```
x_l = [18,21,20]
d_l = [1,2,1]
m=2

def cal_e_xl(x_l,d_l,m):
    value = 0
    for i in range(len(x_l)):
        value += d_l[i]*x_l[i]
    value = value/(2*m)
    return value

e_xl2 = cal_e_xl(x_l,d_l,m)*cal_e_xl(x_l,d_l,m)
print(e_xl2)
```

```
400.0
```

Pearson correlation $\rho(X, Y)$ is the normalized version of covariance

$$\rho(X_L, X_R) = \frac{\sigma(X_L, X_R)}{\sigma(X_L)\sigma(X_R)}.$$

In our case: $\underbrace{\sigma(X_L) = \sigma(X_R)}_{\text{Standard deviation}}$ $\qquad \underbrace{\sigma_X^2 = \mathrm{E}[(X - \mathrm{E}[X])^2] = \mathrm{E}[X^2] - [\mathrm{E}[X]]^2}_{\text{Variance}}$



$$X_L = \begin{bmatrix} 18 \\ 21 \\ 21 \\ 20 \end{bmatrix} \qquad X_R = \begin{bmatrix} 21 \\ 18 \\ 20 \\ 21 \end{bmatrix}$$

$$\rho(X_L, X_R) = -0.67$$

$$\rho(X_L, X_R) = \frac{-1}{\sqrt{1.5}\sqrt{1.5}} = -\frac{2}{3} = -0.67$$

See you next week☺