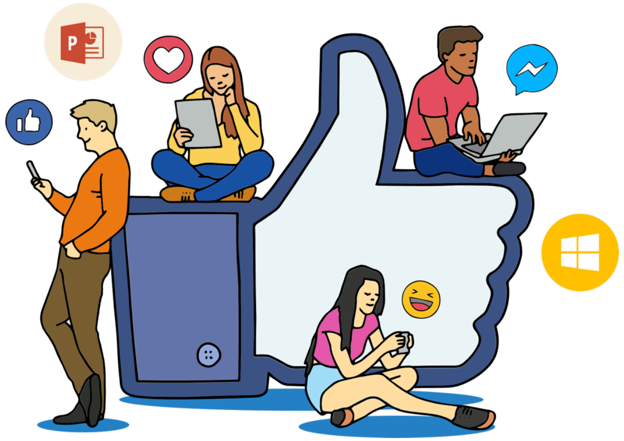
INFS7450 SOCIAL MEDIA ANALYTICS
Tutorial Week 5

School of EECS

The University of Queensland



Outlines

- Feedback of Quiz 2
- Knowledge Extension to Lecture 5
- Code Demo
- Q&A

Section 1: Feedback of Quiz 2

online quiz 2: Q1

Which of the following is correct?

A. Centrality measures were proposed to account for the importance of the nodes in a network
B. Centrality Measures mainly include: Geometric Measures, Spectral Measures, and Path-based Measures.
C. (In)Degree Centrality, Closeness Centrality, Harmonic Centrality belong to Geometric Centrality Measures.
D. A, B, C are all correct. ✅

(In)Degree Centrality
Closeness Centrality
Harmonic Centrality

Eigenvector Centrality
Kat’s Index
PageRank
Hits

Edge Betweenness
Node Betweenness

online quiz 2: Q2

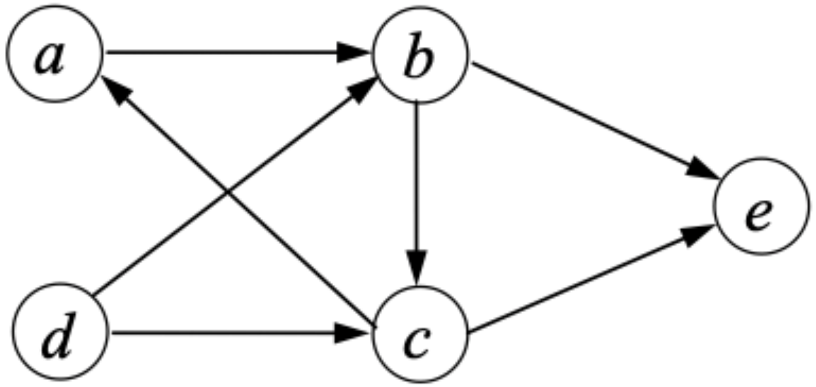
Which of the followings belong to Spectral Centrality Measures?

A. Eigenvector Centrality, (In)Degree Centrality, PageRank, Hits
B. Eigenvector Centrality, Kat's Index, PageRank, Hits. ✅
C. (In)Degree Centrality, Closeness Centrality, Harmonic Centrality
D. Closeness Centrality, Harmonic Centrality, PageRank.

online quiz 2: Q3

The (In)Degree Centrality of node c is:

A. 1
B. 2 ✅
C. 3
D. 4



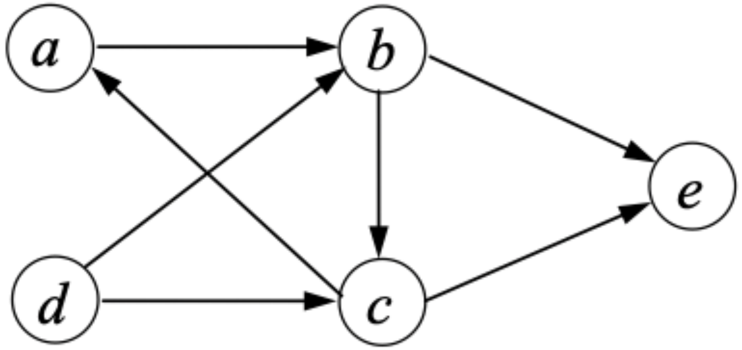
In some other third-party libraries, Degree Centrality might be normalised by the max degree, for example:

```
s = 1.0 / (len(G) - 1.0)
centrality = {n: d * s for n, d in G.degree()}
return centrality
```

online quiz 2: Q4

Having compared to the definition of in-degree centrality, can you try to calculate the OUT-degree centrality of node c in the following graph?

A. 1
B. 2 ✅
C. 3
D. 4



online quiz 2: Q5

Having compared the definition of (In)degree centrality for directed graphs, can you calculate the Degree Centrality of node v3 in the following undirected graph? Let's say, node index starts from 1. We show the adjacent matrix of this graph, and you need to reconstruct the graph according to this matrix, or you can also calculate the degree centrality directly from the matrix.

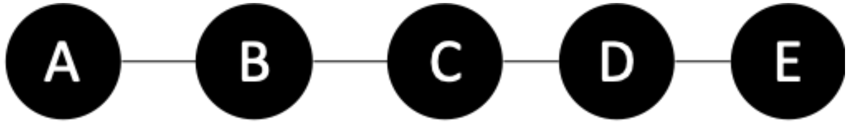
A. 1
B. 2
C. 3 ✅
D. 4

$$A = \begin{bmatrix} 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \end{bmatrix}.$$

online quiz 2: Q6

What is the closeness centrality of node A in the following graph (results accurate to 0.1)

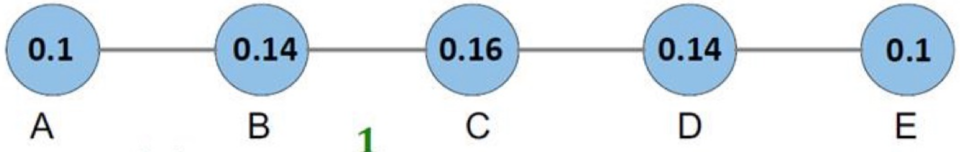
A. 1
B. 2
C. 0.1 ✅
D. 0.5



Closeness Centrality:

$$c_{\text{clos}}(x) = \frac{1}{\sum_y \boxed{d(y,x)}}$$

length of the shortest path from x to y

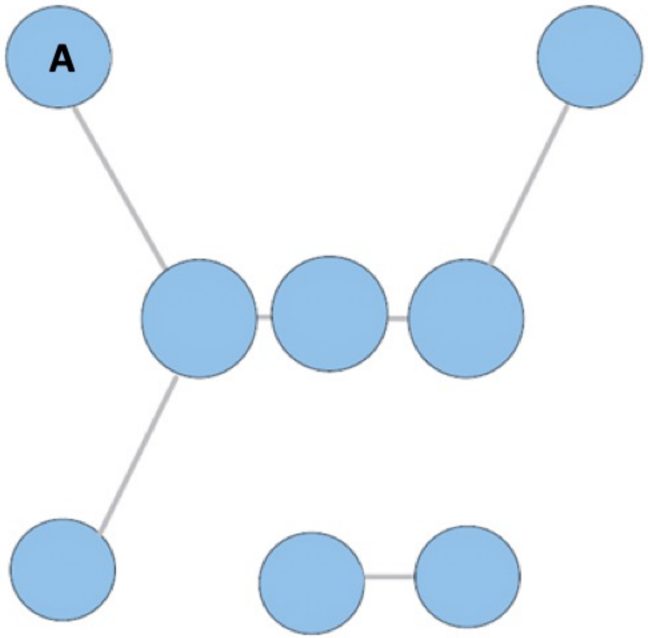


$$c_{clos}(A) = \frac{1}{1+2+3+4} = 0.1$$

online quiz 2: Q7

The Harmonic Centrality of node A in the following graph is: (The result remains two decimal places

A. 3.55
B. 3.88
C. 2.58 ✅
D. 1.0



Harmonic Centrality

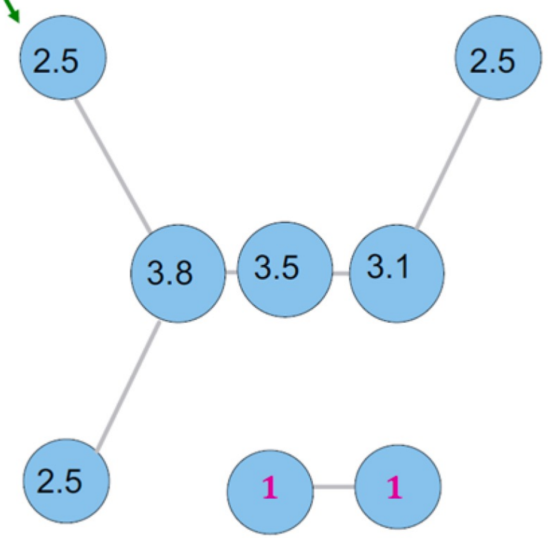
Rather than summing the distances of a node to all other nodes, the harmonic centrality algorithm sums the inverse of those distances. This enables it to deal with infinite values.

$$c_{\text{har}}(x) = \boxed{\sum_{y \neq x} \frac{1}{d(y,x)}} = \sum_{d(y,x)<\infty, y \neq x} \frac{1}{d(y,x)}$$

$$c_{\text{har}}(x) = \frac{1}{\text{n-1}} \boxed{\sum_{y \neq x} \frac{1}{d(y,x)}} = \frac{1}{\text{n-1}} \sum_{d(y,x)<\infty, y \neq x} \frac{1}{d(y,x)}$$

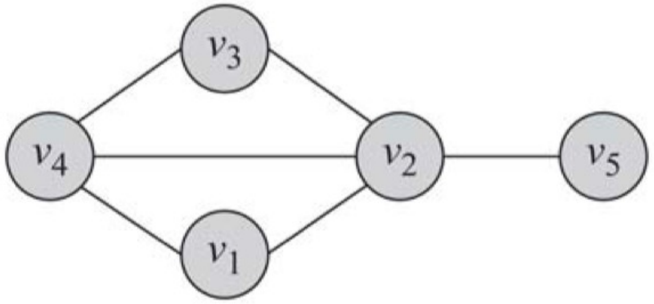
- Strongly correlated to closeness centrality
- Naturally also accounts for nodes y that cannot reach x
- Can be applied to graphs that are **not strongly connected**

$$c_{harm} = \frac{1}{1} + \frac{1}{2} + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} = \mathbf{2.58}$$



online quiz 2: Q8

For the following graph, the adjacency matrix is as follows (matrix A).



$$A = \begin{bmatrix} 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix}.$$

The eigenvalues of A are (−1.74, −1.27, 0.00, +0.33, +2.68). For eigenvector centrality, the largest eigenvalue is selected: 2.68. The corresponding eigenvector is the eigenvector centrality vector and is

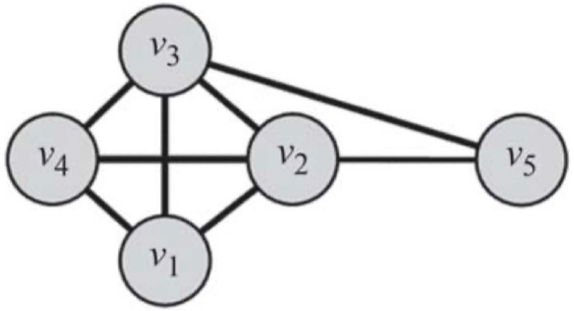
$$\mathbf{C}_e = \begin{bmatrix} 0.4119 \\ 0.5825 \\ 0.4119 \\ 0.5237 \\ 0.2169 \end{bmatrix}.$$

Based on eigenvector centrality, which node is the most central node?

A. v5
B. v4
C. v3
D. v2 ✅

online quiz 2: Q9

For the following graph, the adjacency matrix is as matrix A).



$$A = \begin{bmatrix} 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \end{bmatrix} = A^T.$$

The eigenvalues of A are (-1.68, -1.0, -1.0, $+0.35$, $+3.32$). The largest eigenvalue of A is $\lambda = 3.32$. We assume $\alpha = 0.25 < 1/\lambda$ and $\beta = 0.2$. Then, Katz centralities are

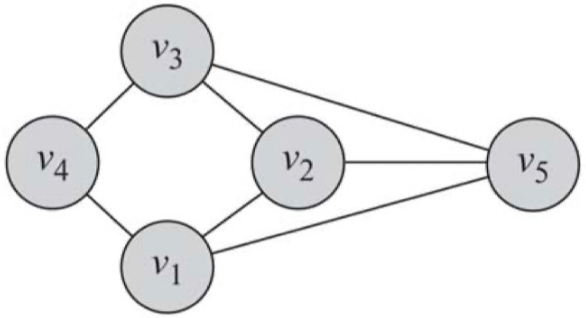
$$\mathbf{C}_{Katz} = \beta(\mathbf{I} - \alpha A^T)^{-1} \cdot \mathbf{1} = \begin{bmatrix} 1.14 \\ 1.31 \\ 1.31 \\ 1.14 \\ 0.85 \end{bmatrix}.$$

According to the Katz centrality. Which is/are the most important node/nodes?

A. only v2
B. only v3
C. v2, and v3 ✅
D. v5

online quiz 2: Q10

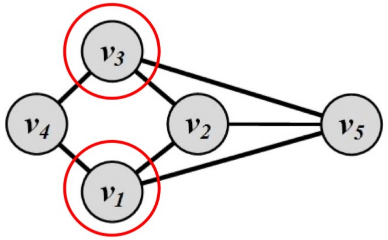
For the following graph, the adjacency matrix is as matrix A:



$$A = \begin{bmatrix} 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \end{bmatrix}.$$

We assume $\alpha = 0.95 < 1$ and $\beta = 0.1$. Then, please calculate the PageRank values (values accurate to 0.01) and find the most importance node/nodes.

A. v1
B. v3
C. v1, and v3 ✅
D. v5



$$A = \begin{bmatrix} 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \end{bmatrix} \quad D = \begin{bmatrix} 3 & 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 3 \end{bmatrix}$$

$$\mathbf{C}_p = \beta(\mathbf{I} - \alpha A^T D^{-1})^{-1} \cdot \mathbf{1} = \begin{bmatrix} 2.14 \\ 2.13 \\ 2.14 \\ 1.45 \\ 2.13 \end{bmatrix}$$

Very similar to Katz

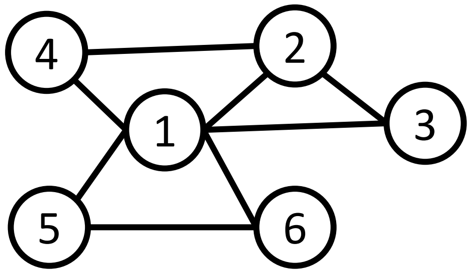
We can also calculate pagerank via power iteration method

Section 2: Knowledge Extension to Lecture 5

- In-class quiz about computing:
  - Degree distribution
  - Diameter
  - Average shortest path length
  - Average clustering coefficient
- Why is the clustering coefficient of random graph p?
  - What is the expected value?
  - What is binomial distribution?
  - What is Bernoulli distribution?

In-class Quiz

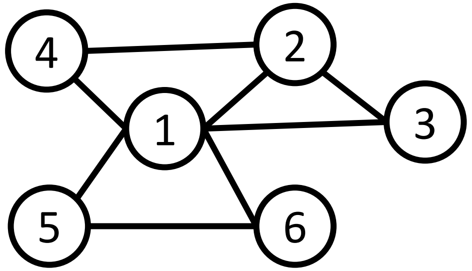
Consider the undirected graph in the figure below:



(a) Write down the degree distribution of the graph
(b) Compute the diameter of the graph, and indicate at least one pair of nodes which are at distance
(c) Compute the average shortest path length
(d) Compute the average clustering coefficient

In-class Quiz

Consider the undirected graph in the figure below:



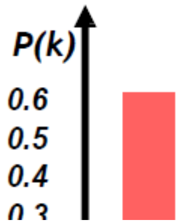
(a) Write down the degree distribution of the graph

- **Degree distribution $P(k)$:** Probability that a randomly chosen node has degree k

  N_k = # nodes with degree k

- Normalized histogram:

  $P(k) = N_k / N$ → plot



Typical wrong answer in the past exams:

$p(1) = \frac{5}{6}, p(2) = \frac{3}{6}, p(3) = \frac{2}{6}, p(4) = \frac{2}{6}, p(5) = \frac{2}{6}, p(6) = \frac{2}{6},$

Solution:

Step 1: list out the degrees of all the nodes in the graph

Node 1: 5

Node 2: 3

Node 3: 2

Node 4: 2

Node 5: 2

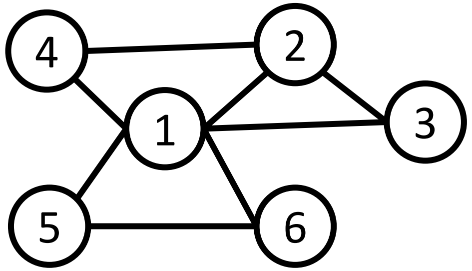
Node 6: 2

Step 2: Count the frequency of each degree value

$$p(degree = 2) = \frac{4}{6}, p(degree = 3) = \frac{1}{6}, p(degree = 5) = \frac{1}{6}$$

n-class Quiz

Consider the undirected graph in the figure below:



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

(a) Compute the diameter of the graph, and indicate at least one pair of nodes which have maximum shortest path

Solution:

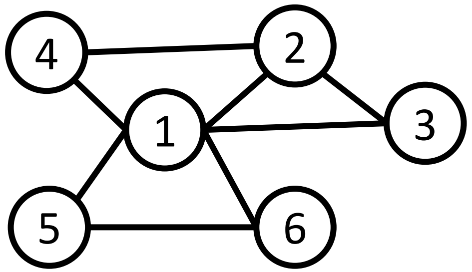
Step 1: List out the shortest path for every pair of nodes

	Node 1	Node 2	Node 3	Node 4	Node 5	Node 6
Node 1						
Node 2	1					
Node 3	1	1				
Node 4	1	1	2			
Node 5	1	2	2	2		
Node 6	1	2	2	2	1	

Step 2: Identify the longest of these shortest paths to determine the diameter

In-class Quiz

Consider the undirected graph in the figure below:



(a) Compute the average shortest path length

	Node 1	Node 2	Node 3	Node 4	Node 5	Node 6
Node 1		1	1	1	1	1
Node 2	1		1	1	2	2
Node 3	1	1		2	2	2
Node 4	1	1	2		2	2
Node 5	1	2	2	2		1
Node 6	1	2	2	2	1	

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) $= n(n-1)/2$

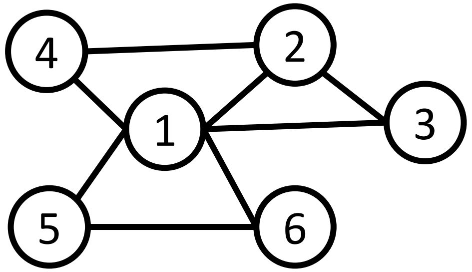
Solution:

$$E_{\max} = \frac{n(n-1)}{2} = \frac{6\times5}{2} = 15$$

$$\bar{h} = \frac{1}{2\times15}\times44 = \frac{44}{30} = 1.47$$

In-class Quiz

Consider the undirected graph in the figure below:



(a) Compute the average clustering coefficient

Solution:

Step 1: Compute clustering coefficient for every node

$$C_1 = \frac{2e_1}{k_1(k_1-1)} = \frac{2\times 3}{5\times 4} = \frac{3}{10} = 0.3$$

$$C_2 = \frac{2e_2}{k_2(k_2-1)} = \frac{2\times 2}{3\times 2} = \frac{2}{3} = 0.67$$

$$C_3 = \frac{2e_3}{k_3(k_3-1)} = \frac{2\times 1}{2\times 1} = 1$$

$$C_4 = \frac{2e_4}{k_4(k_4-1)} = \frac{2\times 1}{2\times 1} = 1$$

$$C_5 = \frac{2e_5}{k_5(k_5-1)} = \frac{2\times 1}{2\times 1} = 1$$

$$C_6 = \frac{2e_6}{k_6(k_6-1)} = \frac{2\times 1}{2\times 1} = 1$$

Clustering coefficient:

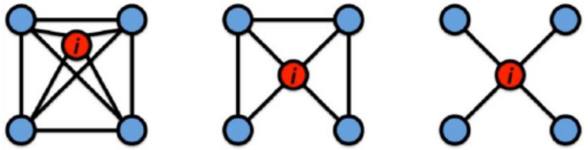
- What portion of ***i***'s neighbors are linked?
- Node ***i*** with degree $\boldsymbol{k_i}$
- $C_i \in [0,1]$

$k_i(k_i-1)/2$ The maximum number of edges between the neighbors of node i

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

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**.



$C_i = 1$ $C_i = 1/2$ $C_i = 0$

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

Step 2: Compute average clustering coefficient

$$C = \frac{1}{6}(0.3 + 0.67 + 1 + 1 + 1 + 1) = 0.83$$

Why is the clustering coefficient of random graph p?

What is the expected value?
What is binomial distribution?
What is Bernoulli distribution?

Clustering Coefficient of G_{np}

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

  Where e_i is the number of edges between i's neighbors

- Edges in G_{np} appear i.i.d. with prob. p
- **So, expected E[e_i] is:** $= p\frac{k_i(k_i-1)}{2}$

  Each pair is connected with prob. p

  Number of distinct pairs of neighbors of node i of degree k_i

- **Then *E[C]*:** $= \frac{p \cdot k_i(k_i - 1)}{k_i(k_i - 1)} = p = \frac{\bar{k}}{n-1} \approx \frac{\bar{k}}{n}$

Clustering coefficient of a random graph is small.
If we generate bigger and bigger graphs with fixed avg. degree k (that is we set $p = k \cdot 1/n$), then C decreases with the graph size n.

The ***expectation*** of a **discrete random variable** X taking the values a1, a2, . . . and with probability mass function p is the number:

$$E[X] = \sum_i a_i P(X = a_i) = \sum_i a_i p(a_i)$$

We also call E[*X*] the ***expected value*** or ***mean*** of *X*. Since the expectation is determined by the probability distribution of *X* only, we also speak of the expectation or mean of the distribution.

Expected values

of discrete random variable

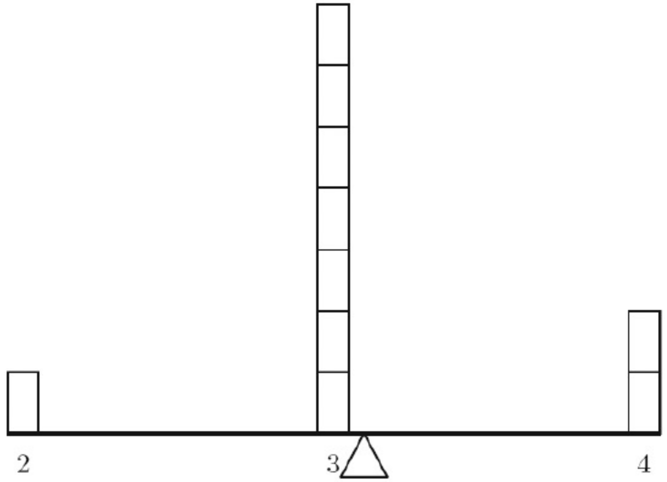


Fig. 7.1. Expected value as center of gravity.

Example

- Let *X* be the discrete random variable that takes the values 1, 2, 4, 8, and 16, each with probability 1/5. Compute the expectation of *X*.

$$\mathrm{E}[X] = \sum_i a_i \mathrm{P}(X = a_i) = \sum_i a_i p(a_i)$$

$$\mathrm{E}[X] = \sum_i a_i \mathrm{P}(X = a_i) = 1 \cdot \frac{1}{5} + 2 \cdot \frac{1}{5} + 4 \cdot \frac{1}{5} + 8 \cdot \frac{1}{5} + 16 \cdot \frac{1}{5} = \frac{31}{5} = 6.2.$$

Bernoulli Distribution

- Let X have Bernoulli distribution with the probability of success p.

$$\mathrm{E}[X] = \sum_i a_i \mathrm{P}(X = a_i) = \sum_i a_i p(a_i)$$

$$E(X) = \sum_x xP(x) = (0)(1-p) + (1)(p) = p$$

$$Var(X) = \sum_x (x-p)^2 P(x) = (0-p)^2(1-p) + (1-p)^2(p)$$

$$= p(1-p)(p+1-p) = p(1-p)$$

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

Binomial Distribution

- Let X have Binomial distribution with the probability of success p and the number of trails n.
- Computing the expectation of X directly leads to a complicated formula, but we can use the fact that X can be represented as the sum of n independent Bernoulli variables:

$$X = X_1 + ... + X_n$$

$$E(X) = E(X_1 + ... + X_n) = E(X_1) + ... + E(X_n) = p + ... + p = np$$

$$Var(X) = Var(X_1 + ... + X_n) = Var(X_1) + ... + Var(X_n) = np(1 - p)$$

Note: We do not need the independence assumption for the expected value, since it is a linear function of RVs, but we need it for variance.

Why is the clustering coefficient of random graph p?

Clustering Coefficient of G_{np}

- **Remember:** $C_i = \dfrac{2e_i}{k_i(k_i-1)}$

  Where e_i is the number of edges between i's neighbors

- Edges in G_{np} appear i.i.d. with prob. p

- **So, expected $\boxed{E[e_i]}$ is:** $= p\dfrac{k_i(k_i-1)}{2}$

  Each pair is connected with prob. p

  Number of distinct pairs of neighbors of node i of degree k_i

- **Then $E[C]$:** $= \dfrac{p \cdot k_i(k_i-1)}{k_i(k_i-1)} = p = \dfrac{\bar{k}}{n-1} \approx \dfrac{\bar{k}}{n}$

Clustering coefficient of a random graph is small.
If we generate bigger and bigger graphs with fixed avg. degree k (that is we set $p = k \cdot 1/n$), then C decreases with the graph size n.

Binomial Distribution

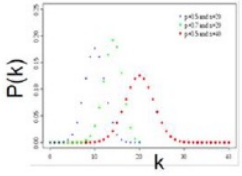
- **Fact:** Degree distribution of G_{np} is binomial.
- Let $P(k)$ denote the fraction of nodes with degree k:

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

Select k nodes out of n-1

Probability of having k edges

Probability of missing the rest of the n-1-k 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

Section 3: Code Demo

See you next week☺