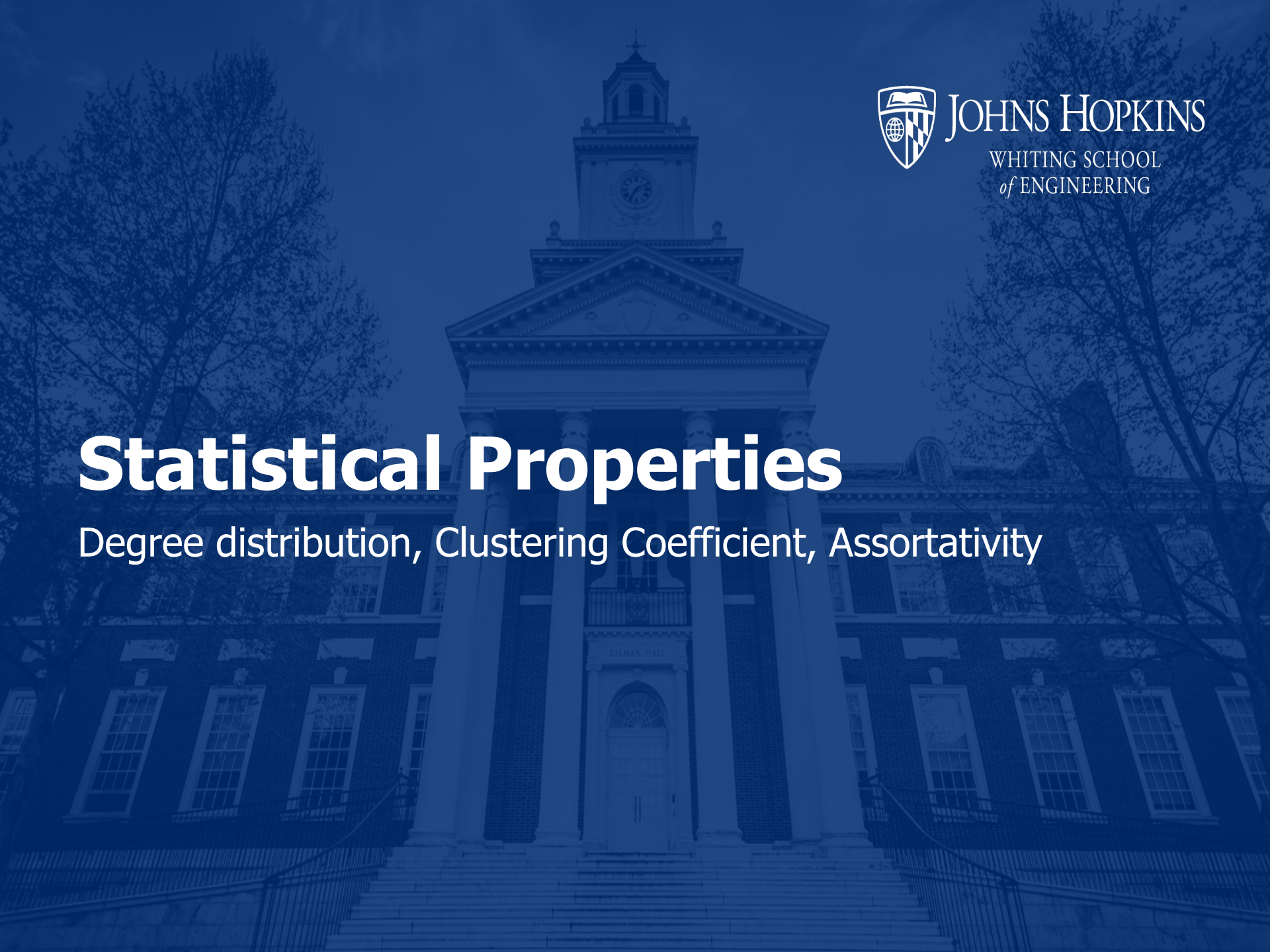


Statistical Properties

Degree distribution, Clustering Coefficient, Assortativity

Topics

Topics:

- **Introduction to Probability & Statistical Analysis**
- **Degree distribution**
- **Clustering coefficient**
- **Assortativity**

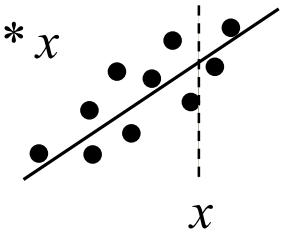
Some Preliminaries on Probability & Statistics

Statistical Learning

Summarizing information about data and serve for decision making

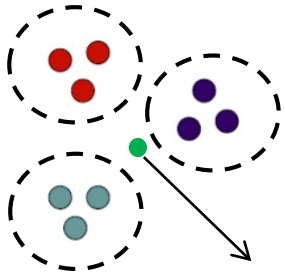
A relationship:

$$y = a + b * x$$



What is the expected value of y for this x ?

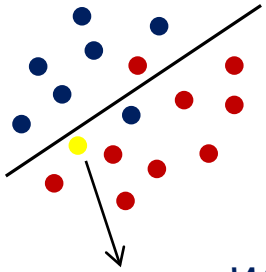
A clustering:



Which cluster does this measurement belong to ?

Is this measurement normal ?

A decision boundary:

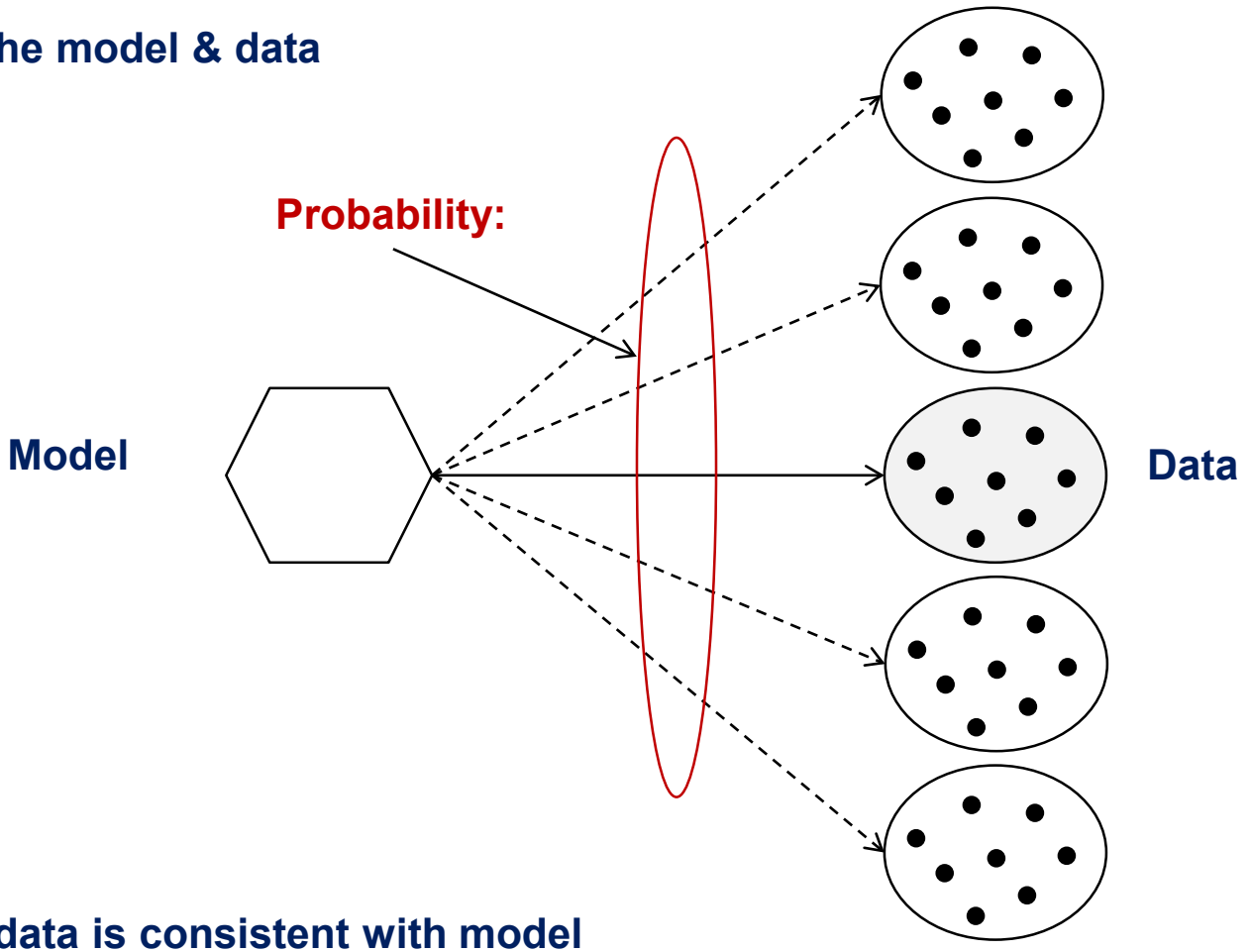


Which class does the new measurement belong to ?

Role of Probability

Probability:

Connection between the model & data



Hypothesis Testing:

Deciding whether the data is consistent with model

Conditional Probability

Why is conditional probability important ?:

- **It allows evidence to be taken into account in inference**
- **Example: What is the probability that you have cancer given a positive test result?**

Conditional Probability:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \quad \Rightarrow \quad P(A \cap B) = P(A \mid B)P(B)$$

Independence:

$$P(A \mid B) = P(A) \quad \Rightarrow \quad P(A \cap B) = P(A)P(B)$$

Bayes' Theorem

Bayes' Theorem follows from conditional probability:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \qquad \& \qquad P(B \mid A) = \frac{P(A \cap B)}{P(A)}$$

$$\therefore \quad P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Significance:

- **Bayes' Theorem allows inversion of conditional probability**
- **Useful for those cases where gathering data for P(A|B) may be difficult while P(B|A) is readily available**

Conditional Probability, an example

Example: **Consider a cancer test in a population with the following statistics:**

$$P(C \mid P) = ?$$
$$P(H \mid N) = ?$$

	Cancer (C) (1%)	Healthy(H) (99%)
Positive (P)	80%	5%
Negative(N)	20%	95%

$$P(C \mid P) = \frac{P(P \mid C)P(C)}{P(P)}$$

$$P(P) = P(P \mid C)P(C) + P(P \mid H)P(H)$$

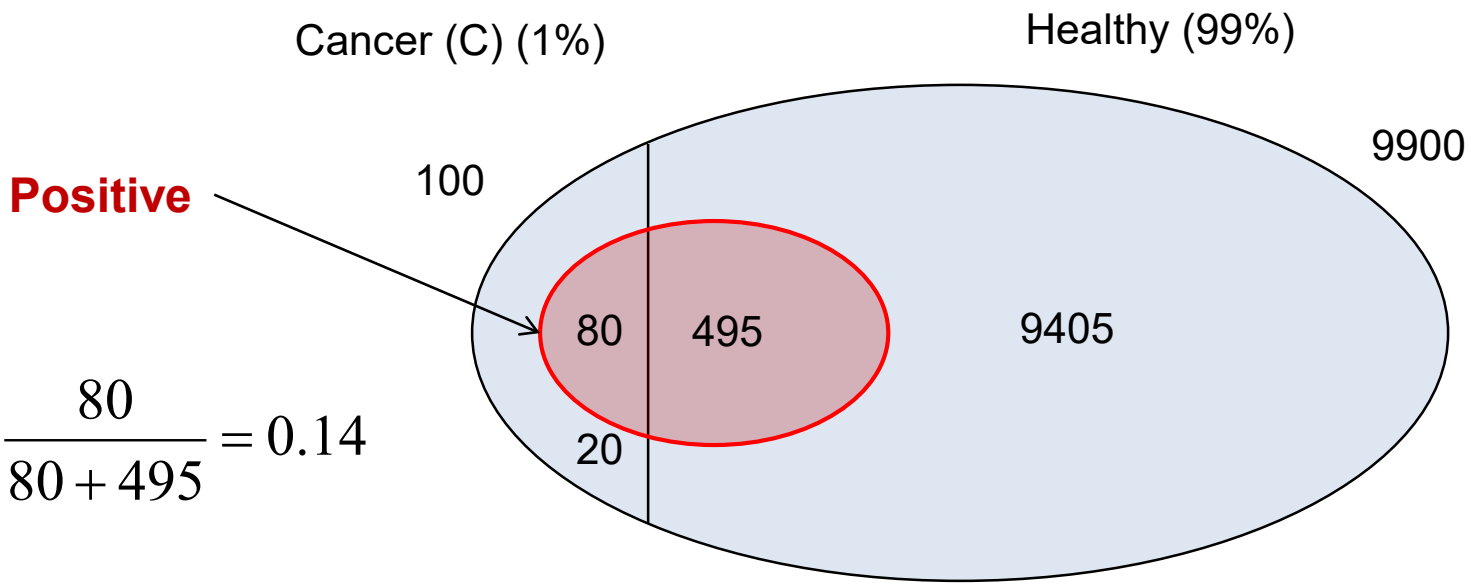
$$P(P) = 0.8 * 0.01 + 0.05 * 0.99 = 0.0575$$

$$P(C \mid P) = \frac{0.8 * 0.01}{0.0575} = 0.14 \quad \Rightarrow$$

Low probability of cancer despite positive test!

Conditional Probability, an example

How does it work? : Consider a population of 10000:



$$P(C \mid P) = \frac{80}{80 + 495} = 0.14$$

Interesting reading:

http://opinionator.blogs.nytimes.com/2010/04/25/chances-are/?_r=0

*I.J. Good, **"When batterer turns murderer,"** Nature, Vol. 375 (1995), p. 541.*
*I.J. Good, **"When batterer becomes murderer,"** Nature, Vol. 381 (1996), p. 481.*

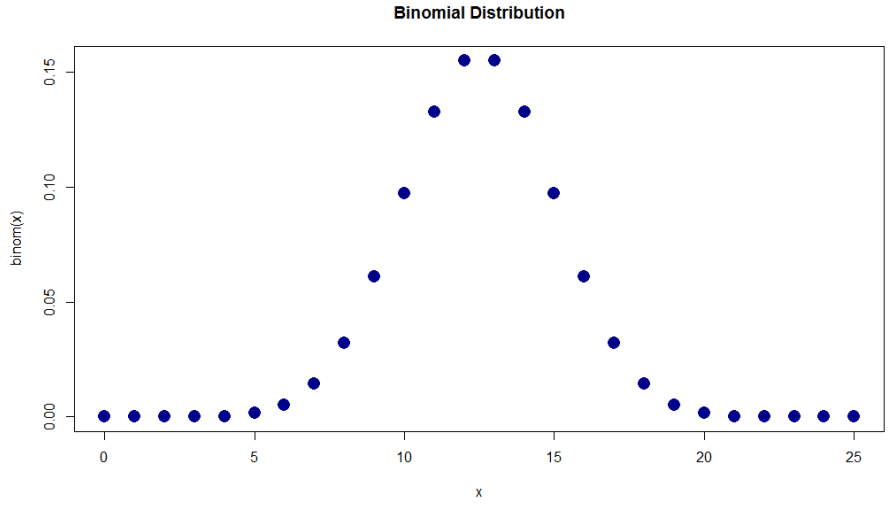
Discrete and Continuous Probability Distributions

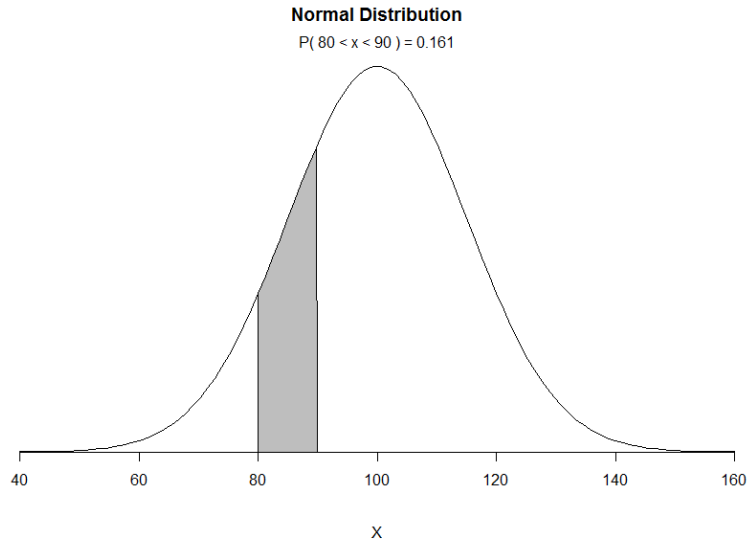
Probability Distribution:

- **Defines probability of all possible outcomes for a random variable**
- **Can be discrete (distribution) or continuous (density), has to be normalized to 1**

$$\sum_{i=1}^{N} P(k_i) = 1$$

$$\int_{-\infty}^{\infty} dx f(x) = 1$$





Basic Properties of Distributions

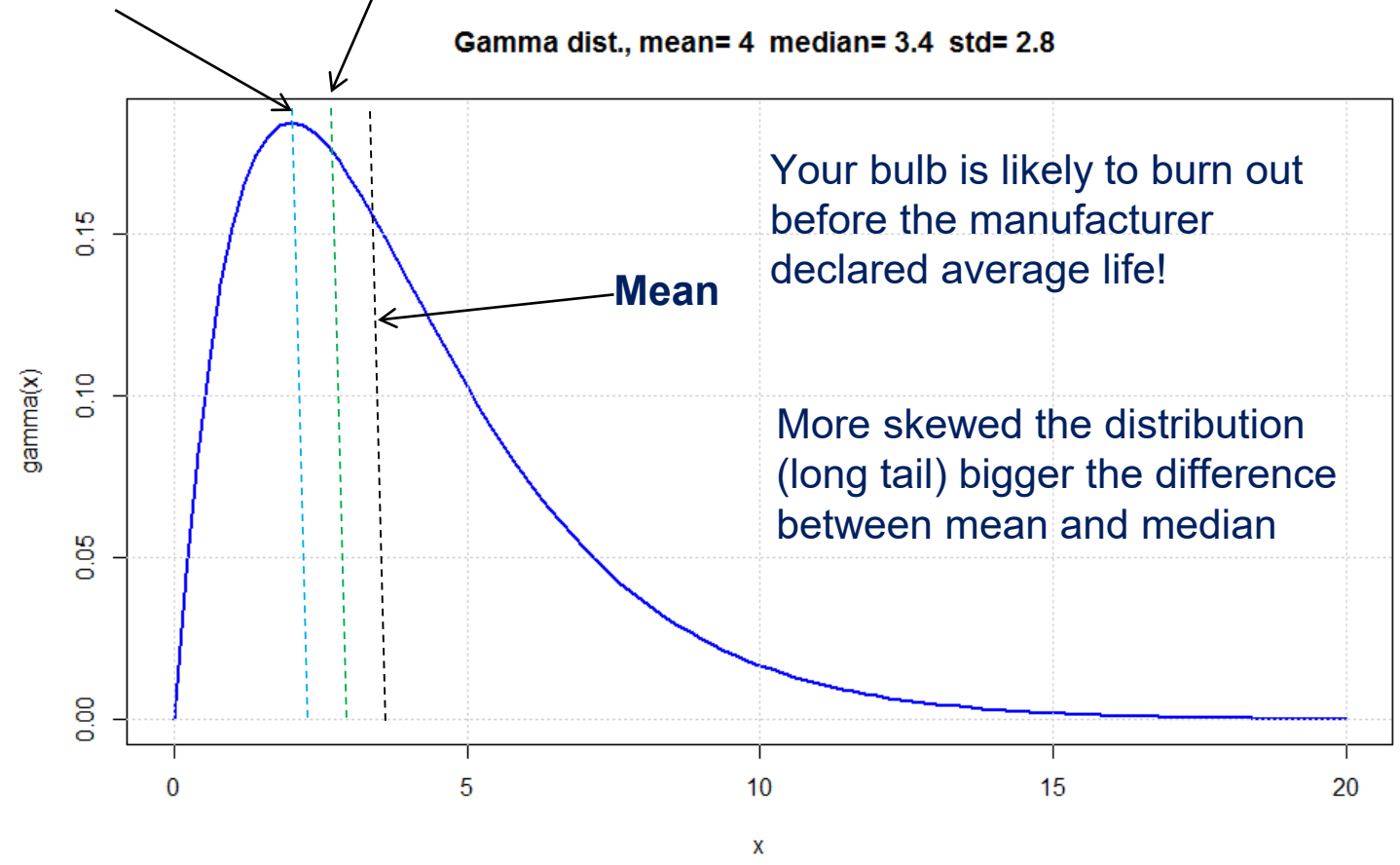
	Discrete:	Continuous:
Probability	$P(k_a \le k \le k_b) = \sum_{i=a}^{b} P(k_i)$	$P(a < x < b) = \int_a^b dx f(x)$
Mean	$\mu = \sum_{i=1}^{N} k_i P(k_i)$	$\mu = \int_{-\infty}^{\infty} dx x f(x)$
Variance	$\sigma^2 = \sum_{i=1}^{N} (k_i - \mu)^2 P(k_i)$	$\sigma^2 = \int_{-\infty}^{\infty} dx (x - \mu)^2 f(x)$

Properties, continued

Median

$$P(x < m) = \int_{-\infty}^{m} dx f(x) = 0.5$$

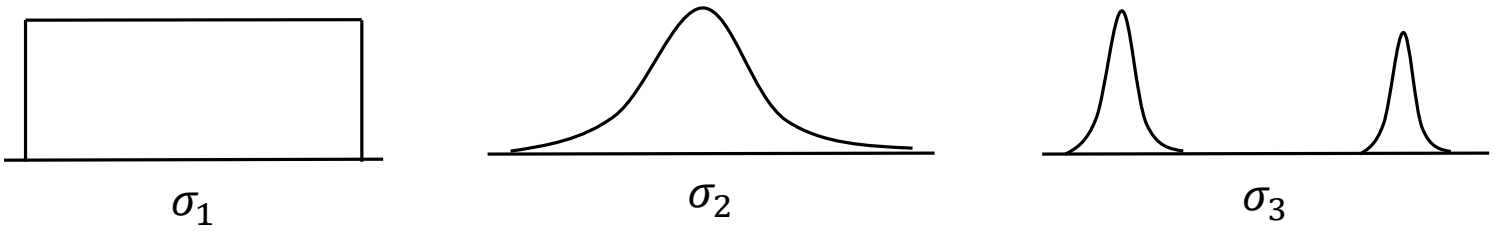
Mode



Entropy of a Distribution

Entropy: **A measure of randomness**

Discrete: $S = -\sum_{i=1}^{N} p_i \log(p_i)$ **Continuous:** $S = -\int dx f(x) \log(f(x))$



σ_1 σ_2 σ_3

Variance: $\sigma_3 > \sigma_1 > \sigma_2$ **Entropy:** $S_1 > S_2 > S_3$

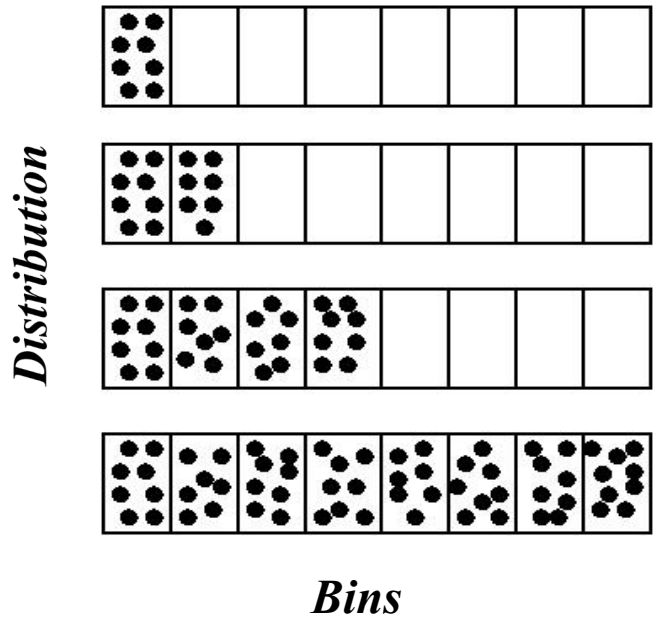
Observation:

- **Variance is not always a good measure of randomness.**

Entropy as a measure of missing information

Entropy is also equivalent to information needed to describe a random variable

$$S = -\sum_i p_i \log(p_i)$$



$$S = -1\log(1) = 0 \qquad \textit{Bits: -}$$

$$S = -2\frac{1}{2}\log(\frac{1}{2}) = 1 \qquad \textit{Bits: 1}$$

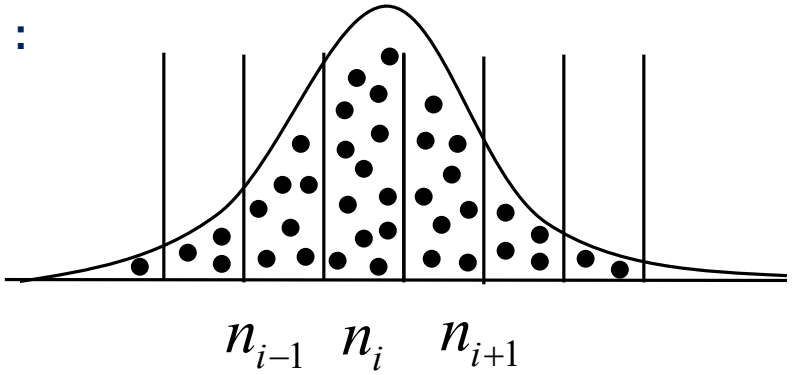
$$S = -4\frac{1}{4}\log(\frac{1}{4}) = 2 \qquad \textit{Bits: 10}$$

$$S = -8\frac{1}{8}\log(\frac{1}{8}) = 3 \qquad \textit{Bits: 011}$$

Where does the Entropy expression come from?

Number of ways N measurements can be allocated :

$$W = \frac{N!}{n_1! n_2! \cdots n_n!}$$



Entropy is defined as:

$$S \equiv \frac{1}{N}\log(W) \quad \Rightarrow \quad S \equiv \frac{1}{N}\log(N!) - \frac{1}{N}\sum_i \log(n_i!)$$

Stirling's approximation: $\log(N!) = N\log(N) - N$

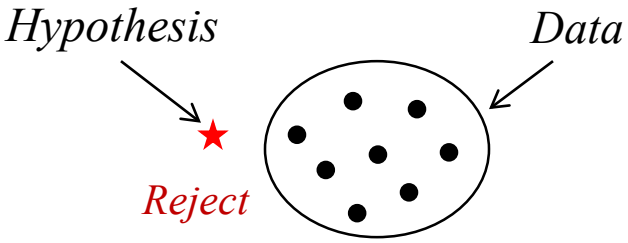
$$S \equiv \frac{1}{N}\left(N\log(N) - N\right) - \frac{1}{N}\sum_i \left(n_i \log(n_i) - n_i\right)$$

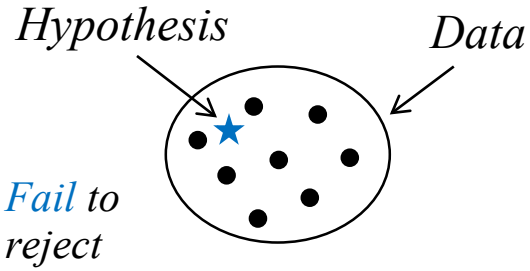
$$p_i \equiv \frac{n_i}{N} \quad \Rightarrow \quad \boxed{S = -\sum_i p_i \log(p_i)}$$

Two Approaches to Statistical Analysis

Frequentist:

- **Starts with a specific hypothesis**
- **Rejects hypothesis if data can not be explained by it**

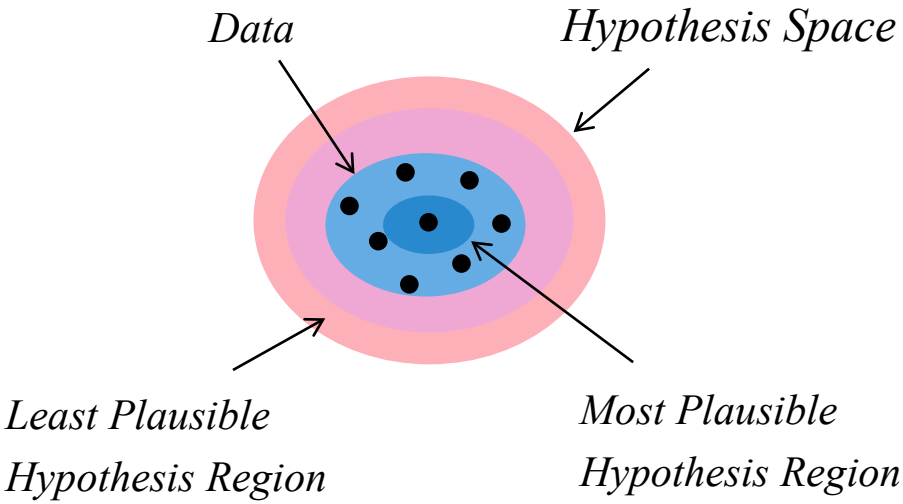




Bayesian:

- **Treats hypothesis as a random variable**
- **Determines plausibility of all possible hypothesis given data using Bayes' Eq.**

$$P(f \mid Data) = \frac{P(Data \mid f) * P(f)}{P(Data)}$$



Bayesian Approach

Bayesian:

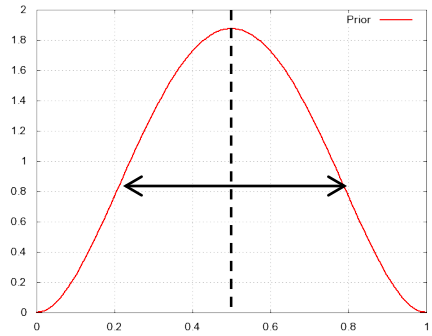
- **Fuse measured data with prior information**

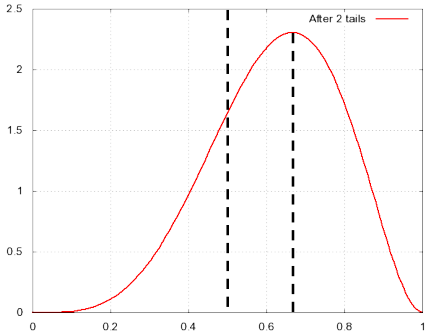
$$P(f \mid Data) = \frac{P(Data \mid f) * P(f)}{P(Data)}$$

Observations:

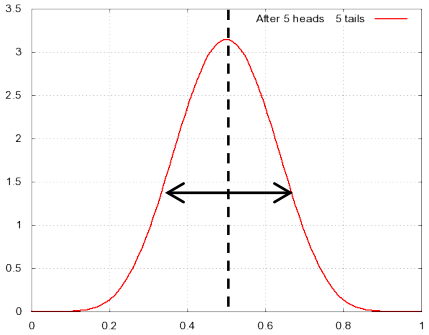
- **Bayesian approach allows a systematic representation of prior & uncertainty**
- **Prior information is key when data is limited**
- A **Fusion** of **Bayesian** & **Frequentist** approaches: Use **prior information** if supported by **prior data**.

Prior :





After 10 tosses:



Bayesian Analysis of Binomial Distribution

Binomial distribution:

$$f(k \mid \pi) = C(n,k)\pi^k (1-\pi)^{n-k}$$

Bayes rule :

$$f(\pi \mid k) = \frac{f(k \mid \pi) f(\pi)}{f(k)}$$

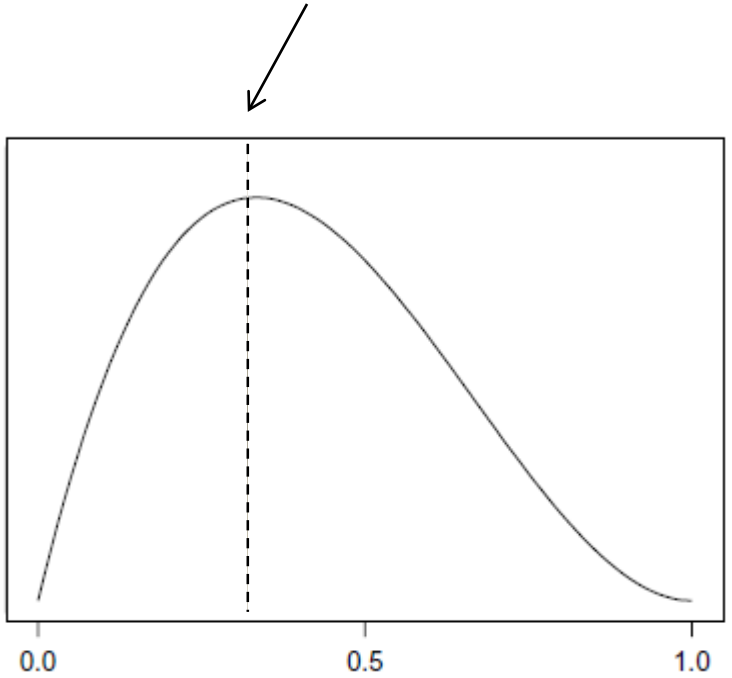
Example:

$$f(\pi) = 1$$

$$n = 3 \quad k = 1$$

$$f(\pi \mid 2) \propto \pi(1-\pi)^2$$

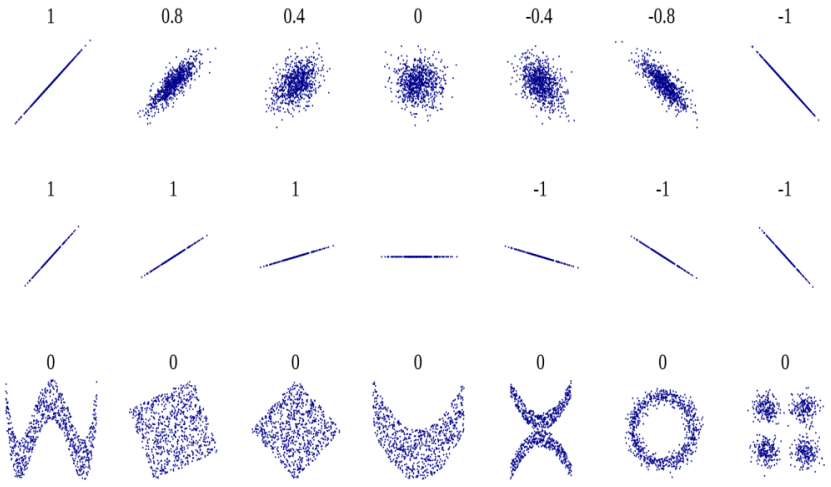
Probability of success is skewed towards lower values because only 1 out of 3 was successful:



Correlation of Numerical Data

Correlation:

- **Information one variable provides about another variable**



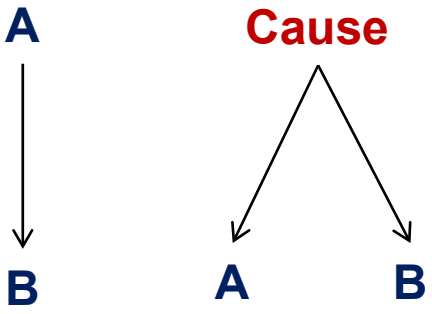
https://commons.wikimedia.org/wiki/File%3ACorrelation_examples2.svg

Pearson Correlation:

$$Correlation(X_i X_j) = \frac{Cov(X_i X_j)}{\sqrt{Var(X_i) Var(X_j)}}$$

Observations:

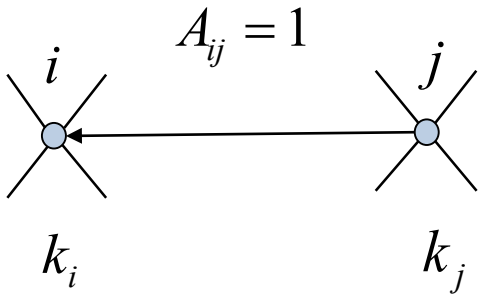
- **Measures linear relationship between two numerical variables**
- **Lack of linear correlation does not imply lack of correlation in general**
- **It is not a measure of causation**



Degree Distribution

Degree

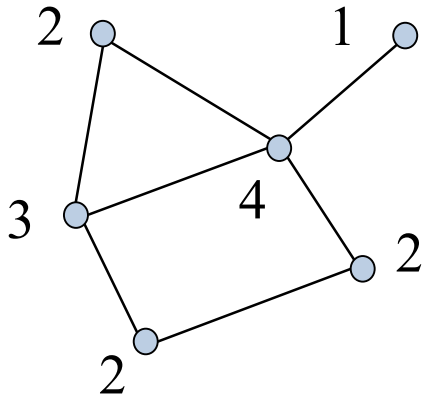
A measure of local connections:



Out-degree: $\delta_j^+ = I^T A = \sum_i A_{ij}$

$$\mathrm{I} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

In-degree: $\delta_i^- = AI = \sum_j A_{ij}$



Directionality:

- **Out: Maybe a sign of access and connection**
- **In: Maybe a sign of popularity or authority**
- **Undirected: Measure of well connectedness**

Degree Histogram, PDF and Cumulative Distributions

Degree distribution can be noisy

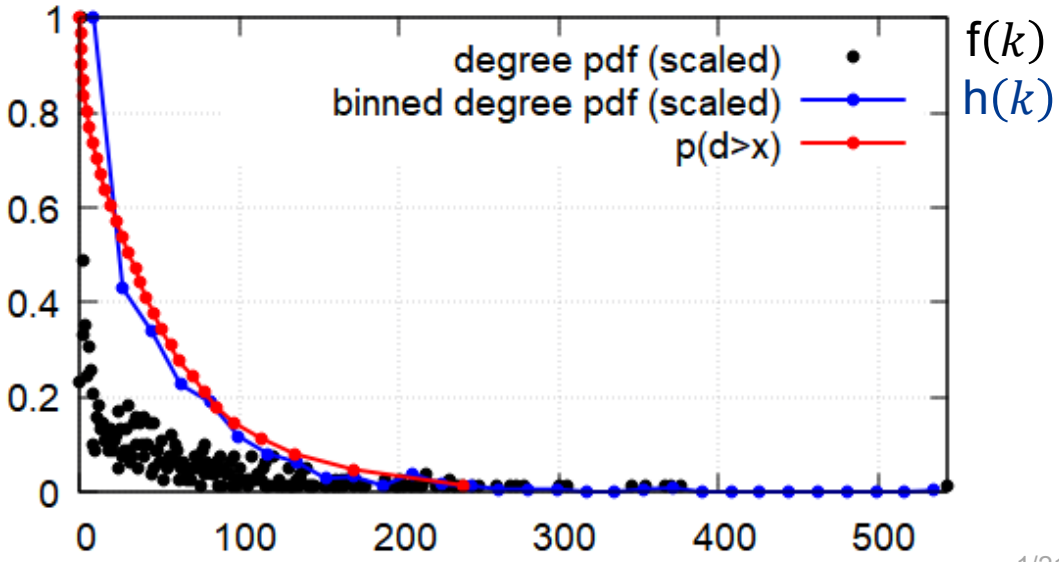
$f(k)$: *number of nodes with degree* k

One way it can be smoothed is by defining a histogram.

$h(k)$: *number of nodes with degree between* $(k, k+\Delta)$

Another option is to instead consider a cumulative distribution

$p(k > x)$ $\quad$ $p(k) = \frac{f(k)}{\sum_i f(i)}$

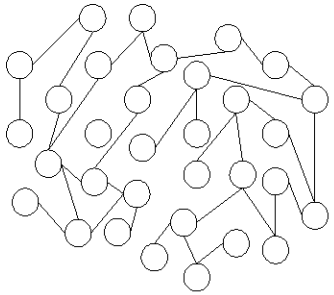


Exponential vs Scale Free Networks

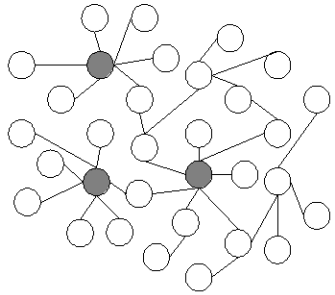
Exponential:

$$P(k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

High degree nodes are suppressed



(a) Random network

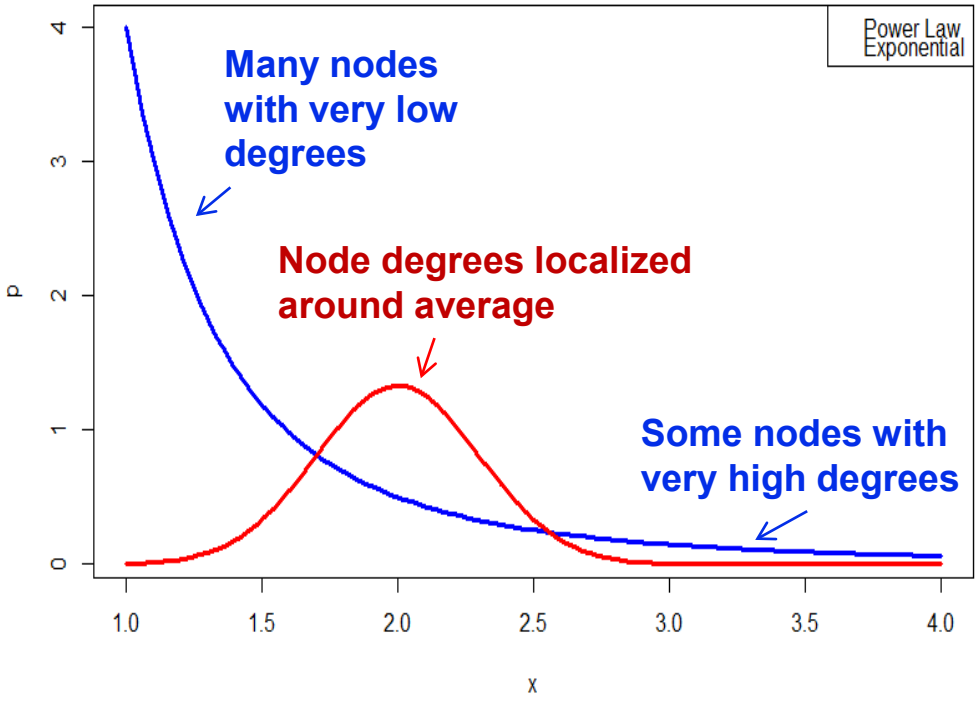


(b) Scale-free network

Scale free:

$$f(ax) = bf(x)$$

$$P(k) = \frac{\lambda}{k^{\alpha}}$$



"Long tail" allows very high degree nodes to exist

Smaller alpha leads to longer tail

Scaling properties of Real World Networks

TABLE II. The scaling exponents characterizing the degree distribution of several scale-free networks, for which $P(k)$ follows a power law (2). We indicate the size of the network, its average degree $\langle k \rangle$, and the cutoff κ for the power-law scaling. For directed networks we list separately the indegree (γ_{in}) and outdegree (γ_{out}) exponents, while for the undirected networks, marked with an asterisk (*), these values are identical. The columns l_{real}, l_{rand}, and l_{pow} compare the average path lengths of real networks with power-law degree distribution and the predictions of random-graph theory (17) and of Newman, Strogatz, and Watts (2001) [also see Eq. (63) above], as discussed in Sec. V. The numbers in the last column are keyed to the symbols in Figs. 8 and 9.

Network	Size	$\langle k \rangle$	κ	γ_{out}	γ_{in}	ℓ_{real}	ℓ_{rand}	ℓ_{pow}	Reference	Nr.
WWW	325 729	4.51	900	2.45	2.1	11.2	8.32	4.77	Albert, Jeong, and Barabási 1999	1
WWW	4×10^7	7		2.38	2.1				Kumar *et al.*, 1999	2
WWW	2×10^8	7.5	4000	2.72	2.1	16	8.85	7.61	Broder *et al.*, 2000	3
WWW, site	260 000				1.94				Huberman and Adamic, 2000	4
Internet, domain*	3015–4389	3.42–3.76	30–40	2.1–2.2	2.1–2.2	4	6.3	5.2	Faloutsos, 1999	5
Internet, router*	3888	2.57	30	2.48	2.48	12.15	8.75	7.67	Faloutsos, 1999	6
Internet, router*	150 000	2.66	60	2.4	2.4	11	12.8	7.47	Govindan, 2000	7
Movie actors*	212 250	28.78	900	2.3	2.3	4.54	3.65	4.01	Barabási and Albert, 1999	8
Co-authors, SPIRES*	56 627	173	1100	1.2	1.2	4	2.12	1.95	Newman, 2001b	9
Co-authors, neuro.*	209 293	11.54	400	2.1	2.1	6	5.01	3.86	Barabási *et al.*, 2001	10
Co-authors, math.*	70 975	3.9	120	2.5	2.5	9.5	8.2	6.53	Barabási *et al.*, 2001	11
Sexual contacts*	2810			3.4	3.4				Liljeros *et al.*, 2001	12
Metabolic, *E. coli*	778	7.4	110	2.2	2.2	3.2	3.32	2.89	Jeong *et al.*, 2000	13
Protein, *S. cerev.**	1870	2.39		2.4	2.4				Jeong, Mason, *et al.*, 2001	14
Ythan estuary*	134	8.7	35	1.05	1.05	2.43	2.26	1.71	Montoya and Solé, 2000	14
Silwood Park*	154	4.75	27	1.13	1.13	3.4	3.23	2	Montoya and Solé, 2000	16
Citation	783 339	8.57			3				Redner, 1998	17
Phone call	53×10^6	3.16		2.1	2.1				Aiello *et al.*, 2000	18
Words, co-occurrence*	460 902	70.13		2.7	2.7				Ferrer i Cancho and Solé, 2001	19
Words, synonyms*	22 311	13.48		2.8	2.8				Yook *et al.*, 2001b	20

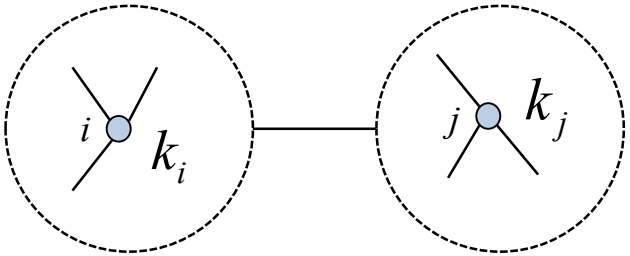
R. Albert and A.-L. Barabási: Statistical mechanics of complex networks

Role of Degree in Probability of Connection Between Nodes

Why is it important: Important to measure significance of a link between nodes

Probability of a connection:

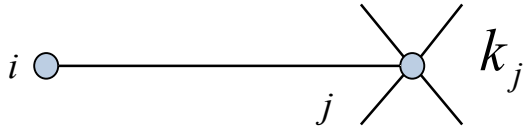
$$p_{ij} = ?$$



$$M \equiv \frac{1}{2}\sum_{i=1}^{N} k_i$$

Probability of attaching 1 link

$$p_1 = \frac{k_j}{2M} << 1$$



Probability of not attaching any one of links of *i* to *j*

$$p_0 = \left(1 - \frac{k_j}{2M}\right)^{k_i} \cong 1 - \frac{k_i k_j}{2M} \quad \Rightarrow \quad \boxed{p_{ij} = \frac{k_i k_j}{2M}}$$

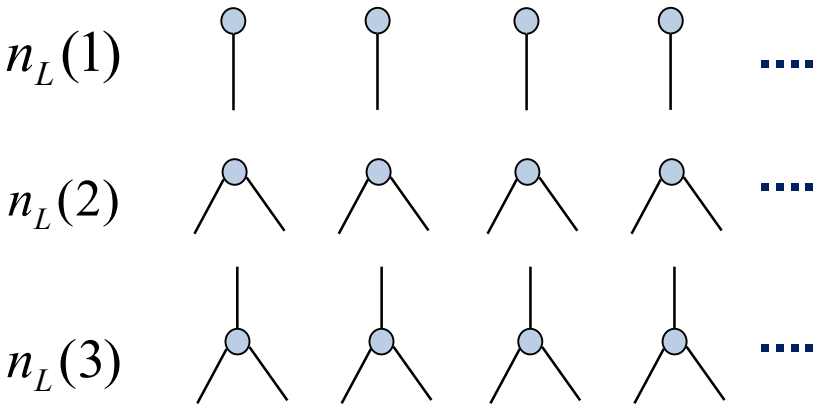
Question of the degree distribution of neighbors (or why we are not that popular)

Probability of connection:

Probability of having k links $P(k)$

# of nodes with k links $NP(k)$

# of links nodes with k links have $n_L(k) = NP(k)k$



Total # of links $2M = n_L = \sum_k n_L(k) = N\sum_k P(k)k$

Degree distribution of neighbors

What is the probability that a random link is attached to a node that has k links ?

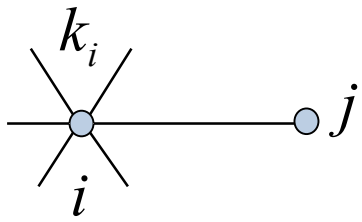
Answer is proportional to the number of links that belong to nodes with k neighbors

$$P_{ng}(k) = \frac{n_L(k)}{n_L} = \frac{kP(k)}{\sum_{k'} k' P(k')} \qquad P_{ng}(k) = \frac{kP(k)}{\langle k \rangle}$$

A Bayesian approach:

Probability that node i has k neighbors given i & j are connected

$$P(k_i \mid i \leftrightarrow j) = \frac{P(i \leftrightarrow j \mid k_i) P(k_i)}{P(i \leftrightarrow j)}$$



$$P(i \leftrightarrow j) = \frac{\langle k \rangle}{N-1} \qquad P(i \leftrightarrow j \mid k_i) = \frac{k_i}{N-1} \qquad P(k_i \mid i \leftrightarrow j) = \frac{k_i P(k_i)}{\langle k \rangle}$$

Average degree of neighbors

What is the average degree of a node's neighbors ? :

Answer is proportional to the number of links that belong to nodes with k neighbors!

$$\langle k \rangle_{ng} = \sum_k k P_{ng}(k) = \frac{1}{\langle k \rangle} \sum_k k\big(kP(k)\big)$$

Mean degree: $\mu = \langle k \rangle$

Standard deviation: $\sigma^2 = \langle k^2 \rangle - \langle k \rangle^2$

$$\langle k \rangle_{ng} = \frac{\langle k^2 \rangle}{\langle k \rangle}$$

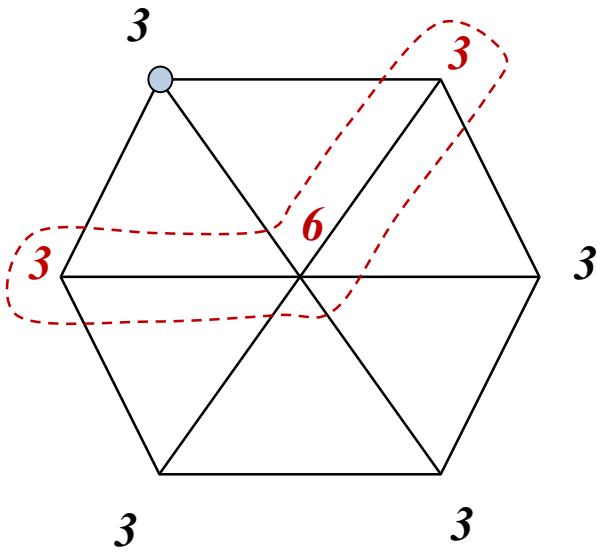
$$\langle k \rangle_{ng} = \frac{\mu^2 + \sigma^2}{\mu} = \mu + \frac{\sigma^2}{\mu} \geq \mu$$

Average degree of neighbors higher than the average degree of graph !

Average degree of neighbors

An example to illustrate why average degree of neighbors tend to be higher:

Number of neighbors:



Average degree:

$$\frac{6*3+6}{7} = 3.43$$

Average degree of neighbors of any node on the periphery:

$$\frac{3+6+3}{3} = 4 > 3.43 > 3$$

Observations:

- **Most nodes have lower than average degree**
- **Most nodes have neighbors whose average degree is higher than themselves**

Expected number of joint neighbors

Why is it important ?: A key measure for discovering hidden links

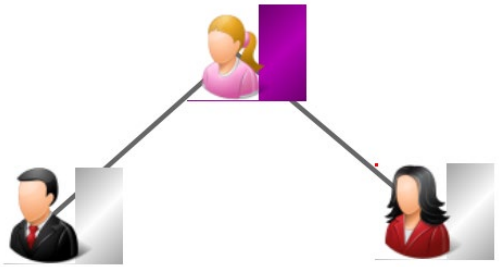
Number of Joint Neighbors:

$$n_{ij} = \sum_l p_{il\&jl} = \sum_l p_{il} p_{jl|il}$$

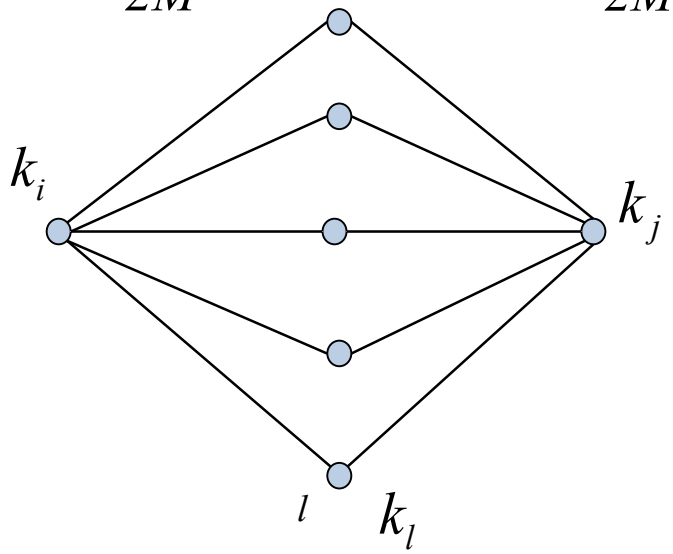
$$n_{ij} = \sum_l \frac{k_i k_l}{2M} \frac{k_j (k_l - 1)}{2M}$$

$$n_{ij} = \frac{k_i k_j}{2M} \sum_l \frac{k_l (k_l - 1)}{2M}$$

$$n_{ij} = p_{ij} \frac{\sum_l k_l (k_l - 1)}{\sum_l k_l} = p_{ij} \frac{\langle k^2 \rangle - \langle k \rangle}{\langle k \rangle}$$



$$p_{il} = \frac{k_i k_l}{2M} \qquad p_{jl|il} = \frac{k_j (k_l - 1)}{2M}$$



Clustering Coefficient

Clustering

Clustering:

- **A measure of community structure.**
- **Measures how connected a node's neighbors are.**

Why is it important ?

- **Most real-world networks tend to create tightly knit groups characterized by high clustering**
- **Clustering can be used to detect structural holes (missing links between neighbors)**
- **Local clustering can be regarded as a type of centrality measure (weakness or strength depending on analytic question)**

Frequency of Triangles

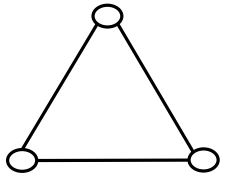
Random graph with N vertices & average degree of λ

Probability two vertices are connected

$$\frac{\lambda}{N-1}$$

Probability of three random vertices forming a triangle

$$\left(\frac{\lambda}{N-1}\right)^3$$



Total expected number of triangles

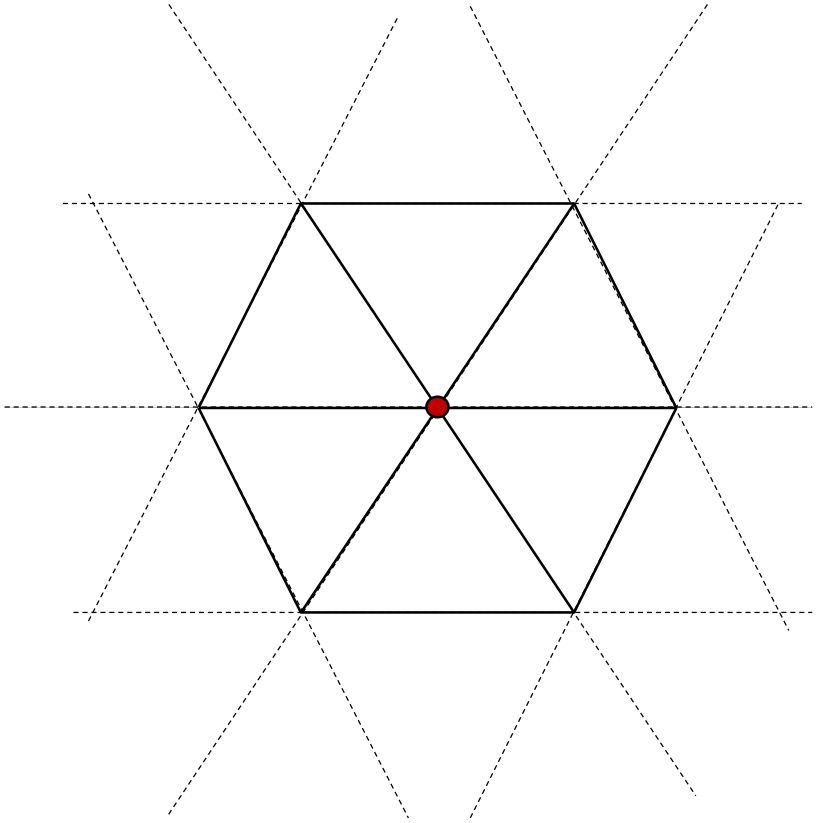
$$C(N,3)\left(\frac{\lambda}{N-1}\right)^3$$

- **Independent of N!**
- **On the order of λ^3**

Clustering coefficient

Clustering is about triangles:

$$C = \frac{\text{Links between neighbors}}{\text{Possible links between neighbors}}$$



Clustering of a hexagon grid:

$$C_{Grid} = \frac{6}{C(6,2)} = \frac{6}{6*5/2} = 0.4$$

Clustering of a comparable random graph:

$$C_{ER} = \frac{\lambda}{N} \to 0 \qquad C_{ER} << C_{Grid}$$

How to build a graph with high clustering ?

Clustering Properties of Real World Graphs

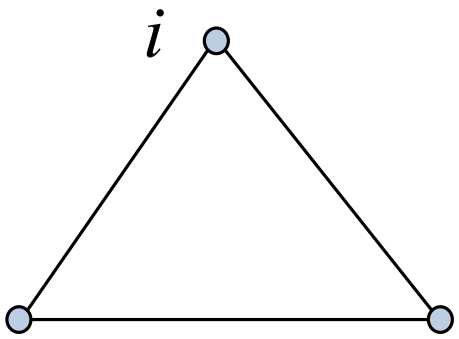
TABLE I. The general characteristics of several real networks. For each network we have indicated the number of nodes, the average degree $\langle k \rangle$, the average path length ℓ, and the clustering coefficient C. For a comparison we have included the average path length ℓ_{rand} and clustering coefficient C_{rand} of a random graph of the same size and average degree. The numbers in the last column are keyed to the symbols in Figs. 8 and 9.

Network	Size	$\langle k \rangle$	ℓ	ℓ_{rand}	C	C_{rand}	Reference	Nr.
WWW, site level, undir.	153 127	35.21	3.1	3.35	0.1078	0.00023	Adamic, 1999	1
Internet, domain level	3015–6209	3.52–4.11	3.7–3.76	6.36–6.18	0.18–0.3	0.001	Yook *et al.*, 2001a, Pastor-Satorras *et al.*, 2001	2
Movie actors	225 226	61	3.65	2.99	0.79	0.00027	Watts and Strogatz, 1998	3
LANL co-authorship	52 909	9.7	5.9	4.79	0.43	1.8×10^{-4}	Newman, 2001a, 2001b, 2001c	4
MEDLINE co-authorship	1 520 251	18.1	4.6	4.91	0.066	1.1×10^{-5}	Newman, 2001a, 2001b, 2001c	5
SPIRES co-authorship	56 627	173	4.0	2.12	0.726	0.003	Newman, 2001a, 2001b, 2001c	6
NCSTRL co-authorship	11 994	3.59	9.7	7.34	0.496	3×10^{-4}	Newman, 2001a, 2001b, 2001c	7
Math. co-authorship	70 975	3.9	9.5	8.2	0.59	5.4×10^{-5}	Barabási *et al.*, 2001	8
Neurosci. co-authorship	209 293	11.5	6	5.01	0.76	5.5×10^{-5}	Barabási *et al.*, 2001	9
E. coli, substrate graph	282	7.35	2.9	3.04	0.32	0.026	Wagner and Fell, 2000	10
E. coli, reaction graph	315	28.3	2.62	1.98	0.59	0.09	Wagner and Fell, 2000	11
Ythan estuary food web	134	8.7	2.43	2.26	0.22	0.06	Montoya and Solé, 2000	12
Silwood Park food web	154	4.75	3.40	3.23	0.15	0.03	Montoya and Solé, 2000	13
Words, co-occurrence	460.902	70.13	2.67	3.03	0.437	0.0001	Ferrer i Cancho and Solé, 2001	14
Words, synonyms	22 311	13.48	4.5	3.84	0.7	0.0006	Yook *et al.*, 2001b	15
Power grid	4941	2.67	18.7	12.4	0.08	0.005	Watts and Strogatz, 1998	16
C. Elegans	282	14	2.65	2.25	0.28	0.05	Watts and Strogatz, 1998	17

R. Albert and A.-L. Barabási: Statistical mechanics of complex networks

Calculating clustering coefficient using Adjacency Matrix

Number of triangles with one corner at i is given by A_{ii}^3



$d_i = Degree\ of\ node\ i$

For an undirected graph each triangle is counted twice in clockwise as well as counterclockwise directions. So number of triangles *a given vertex i* participates in is given by

$$T_i = A_{ii}^3/2$$

Clustering coefficient of the vertex is given by dividing number of triangles by the total number of possible triangles:

$$C_i = \frac{2T_i}{d_i(d_i - 1)}$$

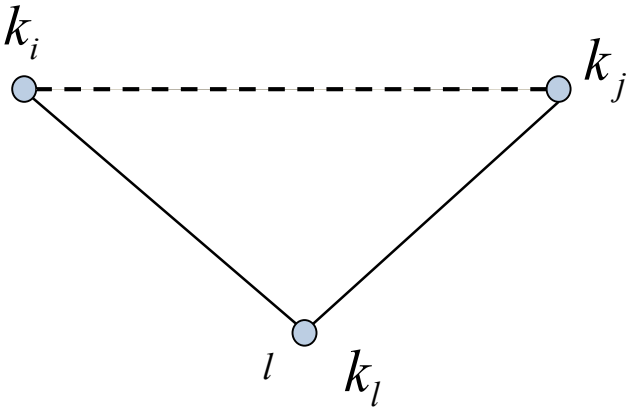
Expected Value of Clustering coefficient for a given degree distribution.

Clustering coefficient of a node:

$$C_l = \sum_{k_i k_j} P(k_i + 1 \mid il) P(k_j + 1 \mid jl) p_{ij}$$

← Ratio of links between neighbors to total possible links

$$P(k_i \mid il) = P_{ng}(k_i) = \frac{k_i P(k_i)}{\langle k \rangle}$$



$$C_l = \sum_{k_i k_j} P_{ng}(k_i + 1) P_{ng}(k_j + 1) p_{ij}$$

$$C_l = \sum_{k_i k_j} \left[\frac{(k_i + 1) P(k_i + 1)}{\langle k \rangle} \right] \left[\frac{(k_j + 1) P(k_j + 1)}{\langle k \rangle} \right] \frac{k_i k_j}{2M}$$

Expected value of clustering coefficient for a given degree distribution

Clustering coefficient :

$$C = \frac{1}{2M}\left[\sum_k k \frac{(k+1)P(k+1)}{\langle k \rangle}\right]^2$$

$$C = \frac{1}{2M\langle k \rangle^2}\left[\sum_{k=0} k(k+1)P(k+1)\right]^2$$

$$C = \frac{1}{2M\langle k \rangle^2}\left[\sum_{k=0} (k-1)(k)P(k)\right]^2 = \frac{1}{2M\langle k \rangle^2}\left[\sum_{k=0} (k^2-k)P(k)\right]^2$$

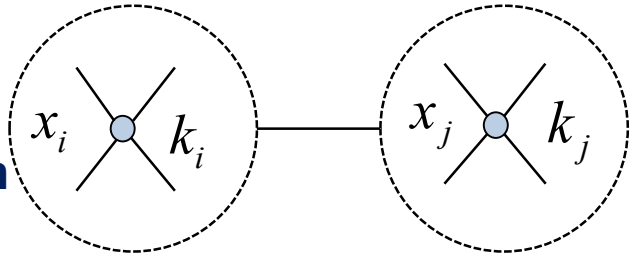
$$C = \frac{\left(\langle k^2 \rangle - \langle k \rangle\right)^2}{2M\langle k \rangle^2} = \frac{\left(\langle k^2 \rangle - \langle k \rangle\right)^2}{N\langle k \rangle^3}$$

$$\sigma^2 = \langle k^2 \rangle - \langle k \rangle^2$$

Assortativity

Node Degree Correlations

Correlating attributes of neighbors:



Iterate over edges and collect variables at both ends.

$$Cov(X_i X_j) = \frac{\sum_{ij} A_{ij}(x_i - \mu)(x_j - \mu)}{\sum_{ij} A_{ij}}$$

Weight is necessary because x_i enters correlation calculation k_i times.

$$Mean(X_i) \to \mu = \frac{1}{2M}\sum_{i=1}^{N} x_i k_i \qquad M \equiv \frac{1}{2}\sum_{i=1}^{N} k_i \qquad Var(X_i) \to \sigma^2 = \frac{1}{2M}\sum_{i=1}^{N}(x_i - \mu)^2 k_i$$

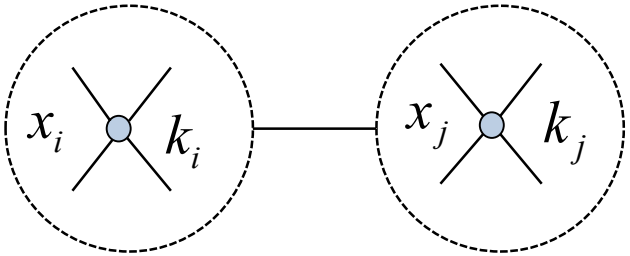
$$Correlation(X_i X_j) = \frac{Cov(X_i X_j)}{\sqrt{Var(X_i)Var(X_j)}}$$

Valid in general. For the subsequent discussion we will explore what this looks like when both are same variables,

Node Attribute Correlation

Correlating attributes of neighbors:

$$Corr(X_i X_j) = \frac{Cov(X_i X_j)}{\sqrt{Var(X_i)Var(X_j)}}$$



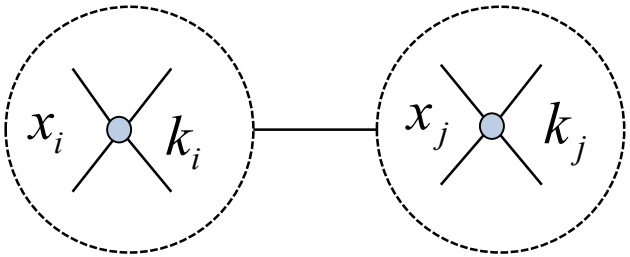
$$Corr(X_i X_j) = \frac{\sum_{ij} A_{ij}(x_i - \mu)(x_j - \mu)}{\sum_{i} k_i (x_i - \mu)(x_i - \mu)} = \frac{\sum_{ij} A_{ij}(x_i x_j - \mu x_i - \mu x_j + \mu^2)}{\sum_{ij} \delta_{ij} k_i (x_i - \mu)(x_j - \mu)}$$

$$\mu = \frac{1}{2M}\sum_{i=1}^{N} x_i k_i$$

$$Corr(X_i X_j) = \frac{\sum_{ij} A_{ij} x_i x_j - \mu \sum_{i} k_i x_i - \mu \sum_{j} k_j x_j + 2M\mu^2}{\sum_{ij} \delta_{ij} k_i x_i x_j - \sum_{ij} \delta_{ij} k_i x_i \mu - \sum_{ij} \delta_{ij} k_j \mu x_j + \sum_{ij} \delta_{ij} k_i \mu^2} = \frac{\sum_{ij} A_{ij} x_i x_j - 2M\mu^2}{\sum_{ij} \delta_{ij} k_i x_i x_j - 2M\mu^2}$$

Node Attribute Correlation

Correlating attributes of neighbors:



$$Corr(X_i X_j) = \frac{\sum_{ij} A_{ij} x_i x_j - 2M\left(\frac{1}{2M}\sum_{ij} k_i x_i\right)^2}{\sum_{ij} \delta_{ij} k_i x_i x_j - 2M\left(\frac{1}{2M}\sum_{ij} k_i x_i\right)^2}$$

$$Corr(X_i X_j) = \frac{\sum_{ij} x_i x_j \left[A_{ij} - \frac{1}{2M} k_i k_j\right]}{\sum_{ij} x_i x_j \left[\delta_{ij} - \frac{1}{2M} k_i k_j\right]}$$

These are not sparse matrices ! Not a good idea to construct this as a matrix in implementation

$$Corr(X_i X_j) = \frac{x^T A x - (x \cdot k)^2 / (2M)}{x^2 - (x \cdot k)^2 / (2M)}$$

Instead carry out calculation to simplify

Attribute Based Link Analysis

Special Case: Correlating degrees $x_i \to k_i$

$$Corr(k_i k_j) = \frac{\sum_{ij} k_i k_j \left[A_{ij} - \frac{1}{2M} k_i k_j \right]}{\sum_{ij} k_i k_j \left[\delta_{ij} - \frac{1}{2M} k_i k_j \right]}$$

$$S_1 \equiv \sum_i k_i \qquad S_2 \equiv \sum_i k_i^2$$

$$S_3 \equiv \sum_i k_i^3 \qquad S_e \equiv \sum_{ij} A_{ij} k_i k_j$$

$$Corr(k_i k_j) = \frac{S_1 S_e - S_2^2}{S_1 S_3 - S_2^2}$$

$Corr < 0$

Disassortative

- **Biological**
- **Cyber**

$Corr = 0$

Neutral

- **Random**

$Corr > 0$

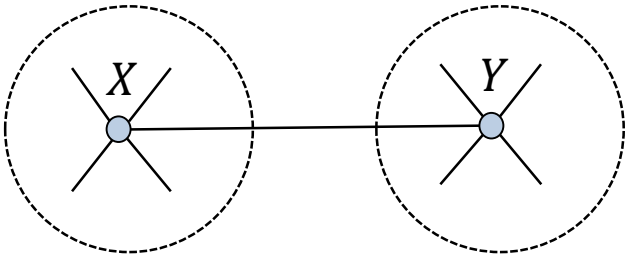
Assortative

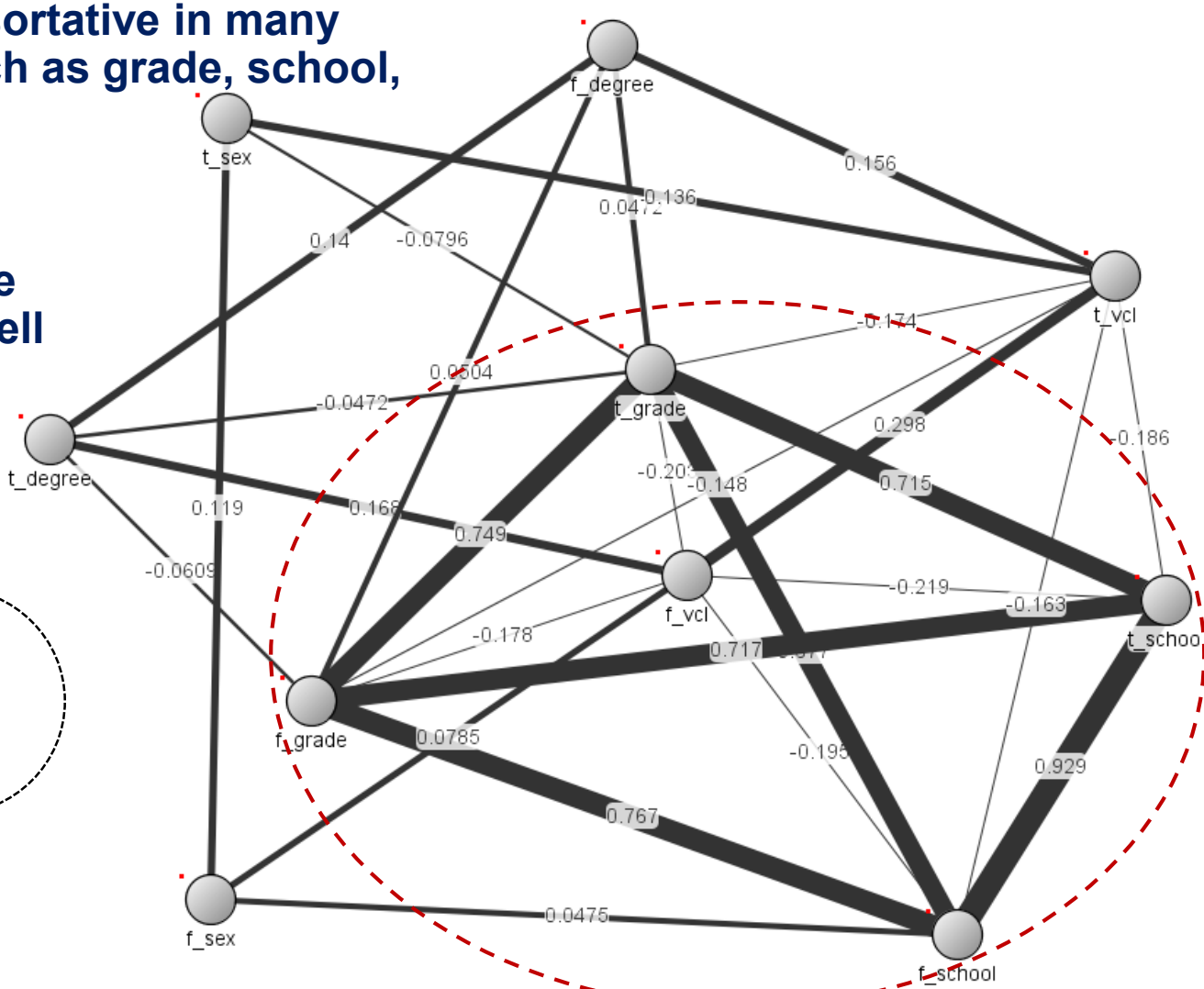
- **Social networks**

Students network

Students network is assortative in many attribute subspaces such as grade, school, degree, sex

It is possible to correlate different attributes as well ("from degree" and "to vertex clustering" for example):





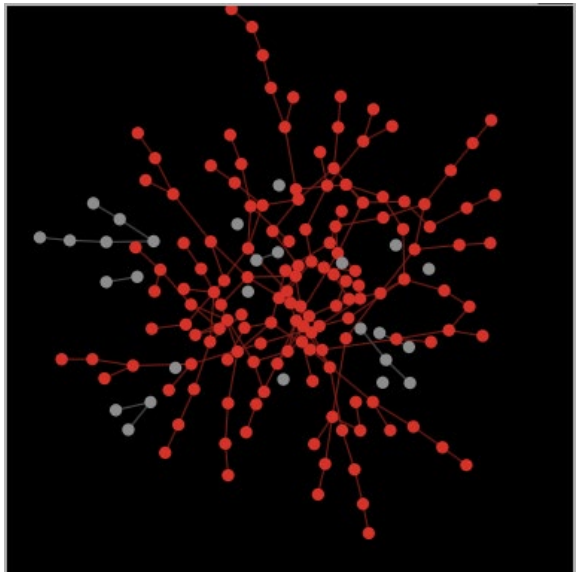
Understanding correlations of nodes that are linked is an important part of explaining relationships and can help with link inference (to be discussed later)

Impact of degree correlations in resiliency

It turns out assortativity makes a network resilient to attacks.

Following example (Netlogo 5.0.3) demonstrates impact of assortativity on resilience on a network of 200 nodes:

Assortativity = 1
After removing top 30 high degree nodes



Assortativity = -1
After removing top 10 high degree nodes