Lecture 7: Quantitative Network Fundamentals II

Selected Social Science Metrics

Degree Distributions and Power Laws

February 25, 2010

Social Network Analysis

- Many structural metrics have been invented and used by Social Scientists studying social networks over the past 70+ years. The Journal *Social Networks* is the research front
- These are well-covered in Wasserman and Faust –*Social Network Analysis* (1994) The following slides cover a ***few*** selected examples in ***one*** area *from that book*. The purpose is to give some feel for the application of such metrics which attempt to measure structural properties of direct interest for social network analysis

- We should also note that transitivity (clustering) and almost all other metrics discussed in this lecture were familiar to and used by social network scientists ***before*** the recent upsurge in activity over the past 10 years.

Transitivity or Clustering coefficient, C

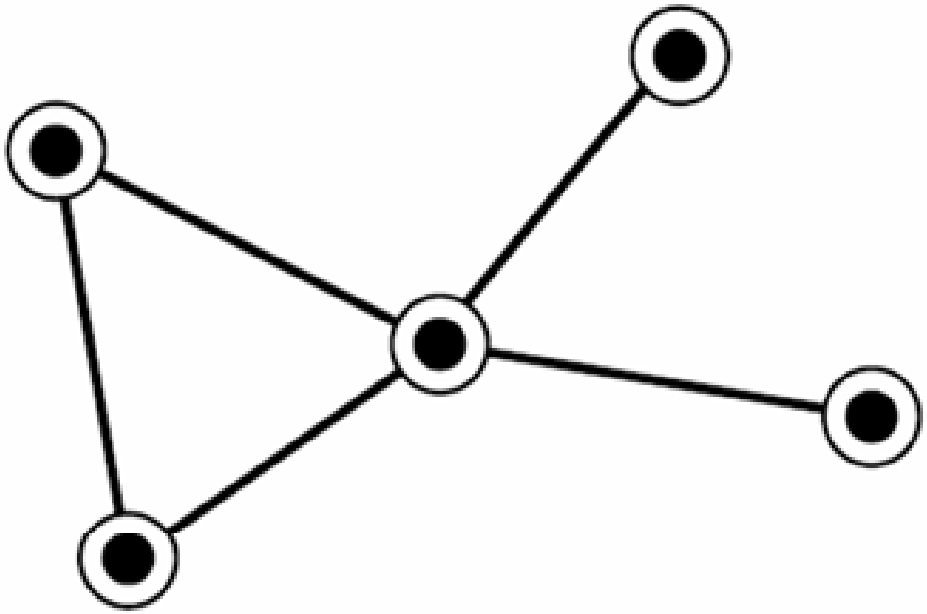
- Measures quantitatively the degree to which nodes which each have relationships with a common node are likely to have a direct relationship.

$$C_1 = \frac{3\ x\ number\ of\ triangles\ in\ network}{number\ of\ connected\ triples\ of\ nodes}$$

$$C_2 = \frac{1}{n}\sum_i C_i;$$

$$C_i = \frac{number\ of\ triangles\ connected\ to\ node\ i}{number\ of\ triples\ centered\ on\ i}$$

Example calculation of transitivity coefficients



This network has one triangle and eight connected triples, and therefore has a clustering coefficient, C_1 of 3 x 1/8 = 3/8 The individual vertices have local clustering coefficients, of 1, 1, 1/6, 0 and 0, for a mean value, C_2 = 13/30.

Source: M. E. J. Newman, *The Structure and Function of Complex Networks*, SIAM Review, Vol. 45, No. 2, pp . 167–256, 2003 Society for Industrial and Applied Mat

Transitivity or Clustering coefficient, $\tau\tau$

- (Almost) always > than expected from random networks thus offering some support for earlier assertions that real networks have some non-random "structure"
  - Thus, assessing clustering is a quick check whether you have a random graph where C = <k>/n. Indeed the size dependence of transitivity can be useful to calculate

Structural Typology (lecture 1)

- Totally regular
  - Grids/crystals
  - Pure Trees
  - Layered trees
  - Star graphs
- Deterministic methods used

- Real things
  - The ones we are interested in
- New methods or adaptations of existing methods needed

- Less regular
  - -"Hub and spokes"
  - -"Small Worlds"
  - -Communities
  - -Clusters
  - -Motifs

- No internal structure
  - Perfect gases
  - Crowds of people
  - Classical economics with invisible hand
- Stochastic methods used

Transitivity or Clustering coefficient, τ -continued

- (Almost) always > than expected from random networks thus offering some support for earlier assertions that real networks have some non-random "structure"
  - Thus, assessing clustering is a quick check whether you have a random graph where C = <k>/n. Indeed the size dependence of transitivity can be useful to calculate
- Higher order clusters (groups of n related nodes) also of interest but no clean way (yet) to separate lower order and higher order tendencies. Moreover, Whitney showed in lecture 4 that methods for calculating higher order clustering in large networks is unknown territory.
- In directed graphs, n=2 effects (the proportion of nodes that point at each other) can be of interest and is labeled **reciprocity**. This is an important social network attribute.

Centrality

- Numerous metrics exist in the Social Networks Literature for assessing the "centrality" of a social network.
  - Centrality metrics attempt to characterize the level of "centralization" of control or action on this network
  - One application is to assess how important **a given actor** (node) is in the network (ranking of nodes according to link information)
  - Another application is to assess **overall** how much of the control of the network is controlled by the "more important" actors (*group or network centrality*)
  - The relative importance of single channels/links and groups of links has also been of interest.
- We will look at a several of the social science defined metrics and explore the definitions by looking at "ideal toy graphs": Team (family or full) graphs, Circle (or line) graphs and Star graphs.

Degree Centrality

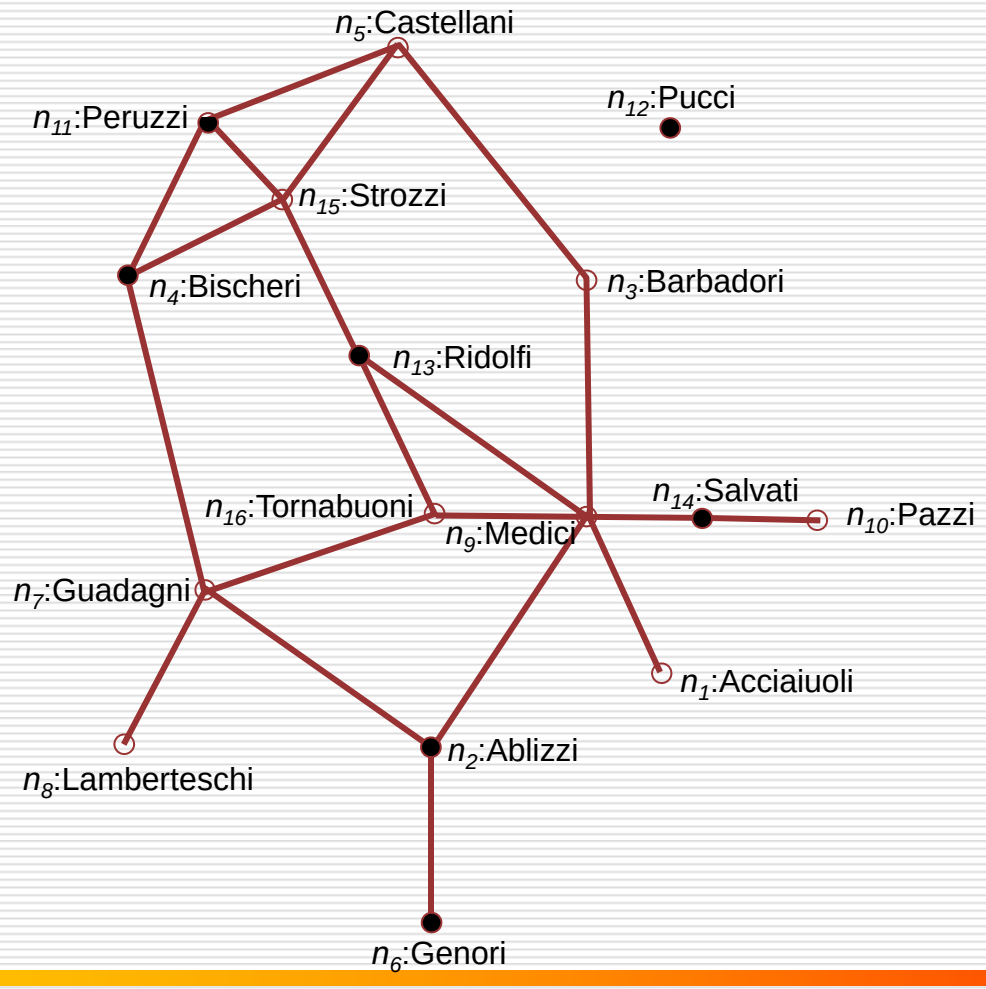
- Actor (can be individual, group or organization depending on what is being studied). The actor in the example we will use is a "Family". *Most central is the node with the most links*.

$$C_D^{'}(n_i) = \frac{\sum_j x_{ij}}{n-1}$$

- Group (all actors in network)
  - = 1 for a star graph
  - = 0 for a circle graph or "team"
  - = 1/(n-1) for line graph

$$C_D = \frac{\sum_{i=1}^{n} [C_D^{\max}(n) - C_D(n_i)]}{[(n-2)]}$$

Padgett's Florentine Families: 15th Century Marriage Relations



Florentine Families Centrality Metrics I: Degree

	$C'_D(n_i)$	$C'_C(n_i)$	$C'_B(n_i)$	$C'_I(n_i)$
Acciaiuoli	0.071			
Ablizzi	0.214			
Barbadori	0.143			
Bischeri	0.214			
Castellani	0.214			
Genori	0.071			
Guadagni	0.286			
Lamberteschi	0.071			
Medici	0.429			
Pazzi	0.071			
Peruzzi	0.214			
Pucci	---			
Ridolfi	0.214			
Salvati	0.143			
Strozzi	0.286			
Tornabuoni	0.214			
Centralization	**0.257**			

Closeness Centrality

- Actor
  - *Closest is shortest* (geodesic) *distance* from other nodes =1 for max closeness and 0 for min

$$C_C^{'}(n_i) = \frac{n-1}{\sum_{j-1}^{n} d(n_i, n_j)}$$

- Group
  - = 0 for circle graph or full network
  - = 1 for star graph
  - 0.277 for line (7 nodes)
  - can estimate several ways including dispersion

$$C_C - \frac{\sum_{i=1}^{n} [C_C^{i\max} - C_C^{'}(n_i)]}{(n-2)(n-1)/(2n-3)}$$

Florentine Families Centrality Metrics II Closeness

	$C'_D(n_i)$	$C'_C(n_i)$	$C'_B(n_i)$	$C'_I(n_i)$
Acciaiuoli	0.071	0.368		
Ablizzi	0.214	0.483		
Barbadori	0.143	0.438		
Bischeri	0.214	0.400		
Castellani	0.214	0.389		
Genori	0.071	0.333		
Guadagni	0.286	0.467		
Lamberteschi	0.071	0.326		
Medici	0.429	0.560		
Pazzi	0.071	0.286		
Peruzzi	0.214	0.368		
Pucci	---	---		
Ridolfi	0.214	0.500		
Salvati	0.143	0.389		
Strozzi	0.286	0.438		
Tornabuoni	0.214	0.483		
Centralization	**0.257**	**0.322**		

Betweeness Centrality I

- Actor
  - Power or influence comes from being an intermediary
  - z is the number of geodesics between two points
- Group
  - =1 for star graph
  - =0 for circle
  - =0.311 for 7 node line
  - Tree Hierarchy = xx

$$C_B^{'}(n_i) = \frac{\sum_{j<k} z_{jk}(n_i)/z_{jk}}{[(n-1)(n-2)/2]}$$

$$C_B = \frac{\sum_{i=1}^{n} [C_B^{'\max}(n) - C_B^{'}(n_i)]}{n-1}$$

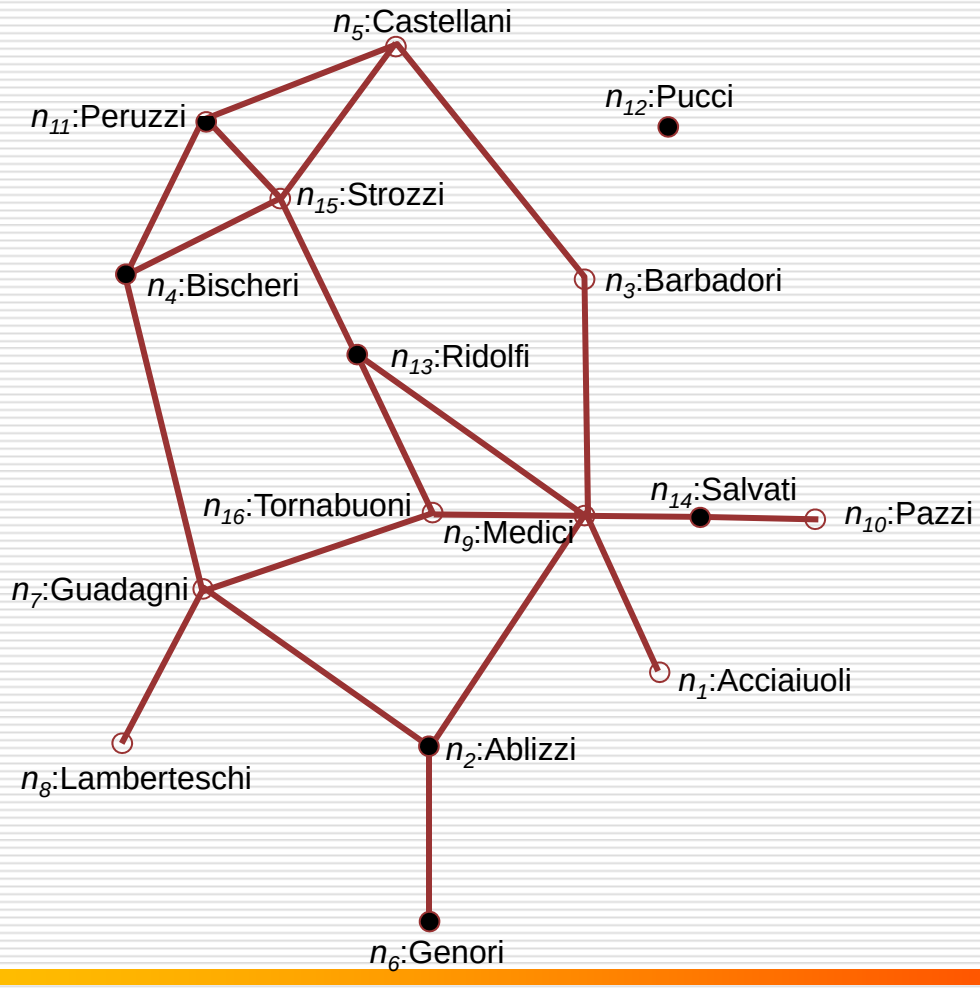
Definition of Hierarchy

- Hierarchy: A description of a group of elements (system?) where each element is graded or ranked and then arranged in a structure that separates elements according to rank which each descending rank being in some way subordinate to the next higher rank (this leads to a level number or node depth). Although hierarchy often describes power or authority relationships, it is also used in describing levels of abstraction and other system features. Flow and containment hierarchies have also been distinguished.
- Hierarchies can take on a variety of structures ranging from Pure layers to pure trees. (Moses next week)

Florentine Families Centrality Metrics:Betweeness

	$C'_D(n_i)$	$C'_C(n_i)$	$C'_B(n_i)$	$C'_I(n_i)$
Acciaiuoli	0.071	0.368	0.000	
Ablizzi	0.214	0.483	0.212	
Barbadori	0.143	0.438	0.093	
Bischeri	0.214	0.400	0.104	
Castellani	0.214	0.389	0.055	
Genori	0.071	0.333	0.000	
Guadagni	0.286	0.467	0.255	
Lamberteschi	0.071	0.326	0.000	
Medici	0.429	0.560	0.522	
Pazzi	0.071	0.286	0.000	
Peruzzi	0.214	0.368	0.022	
Pucci	---	---	---	
Ridolfi	0.214	0.500	0.114	
Salvati	0.143	0.389	0.143	
Strozzi	0.286	0.438	0.103	
Tornabuoni	0.214	0.483	0.092	
Centralization	**0.257**	**0.322**	**0.437**	

Padgett's Florentine Families: 15th Century Marriage Relations



Betweeness Centrality II

- Actor
  - Power or influence comes from being an intermediary
  - z is the number of geodesics between two points
- Group
  - =1 for star graph
  - =0 for circle
  - =0.311 for 7 node line

$$C_B^{'}(n_i) = \frac{\sum_{j<k} z_{jk}(n_i)/z_{jk}}{[(n-1)(n-2)/2]}$$

$$C_B = \frac{\sum_{i=1}^{n} [C_B^{'\max}(n) - C_B^{'}(n_i)]}{n-1}$$

- ***Betweeness Centrality*** has been most applied of the centrality metrics in Social Network Analysis (1994)

Information Centrality

- Actor
  - Estimates the information value of the connections
  - shorter distances are better but are not the only paths used
  - T is the trace, R a row sum and c an element in a matrix constructed from the sociomatrix with information content
  - Actor indices are ***proportions of total "information" flow controlled by a single actor*** and sums to 1 in network
- No group index (1997)

$$C_I(n_i) = \frac{1}{c_{ii} + (T - 2R)/n}$$

$$C_I'(n_i) = \frac{C_I(n_i)}{\sum_i C_I(n_i)}$$

Florentine Families Centrality Metrics

	$C'_D(n_i)$	$C'_C(n_i)$	$C'_B(n_i)$	$C'_I(n_i)$
Acciaiuoli	0.071	0.368	0.000	0.049
Ablizzi	0.214	0.483	0.212	0.074
Barbadori	0.143	0.438	0.093	0.068
Bischeri	0.214	0.400	0.104	0.074
Castellani	0.214	0.389	0.055	0.070
Genori	0.071	0.333	0.000	0.043
Guadagni	0.286	0.467	0.255	0.081
Lamberteschi	0.071	0.326	0.000	0.043
Medici	0.429	0.560	0.522	0.095
Pazzi	0.071	0.286	0.000	0.033
Peruzzi	0.214	0.368	0.022	0.069
Pucci	---	---	---	---
Ridolfi	0.214	0.500	0.114	0.080
Salvati	0.143	0.389	0.143	0.050
Strozzi	0.286	0.438	0.103	0.070
Tornabuoni	0.214	0.483	0.092	0.080
Centralization	**0.257**	**0.322**	**0.437**	**---**

Eigenvector Centrality-UCINET

- UCINET-help, help topics, index (on toolbar), eigenvector centrality
- Given an adjacency matrix A, the centrality of vertex i (denoted ci), is given by ci =aSAijcj where a is a parameter. The centrality of each vertex is therefore determined by the centrality of the vertices it is connected to. The parameter á is required to give the equations a non-trivial solution and is therefore the reciprocal of an eigenvalue. It follows that the centralities will be the elements of the corresponding eigenvector. The normalized eigenvector centrality is the scaled eigenvector centrality divided by the maximum difference possible expressed as **a percentage**.

- For a given binary network with vertices v1....vn and maximum eigenvector centrality cmax, the network eigenvector centralization measure is S(cmax - c(vi)) divided by the maximum value possible, where c(vi) is the eigenvector centrality of vertex vi.

- This routine calculates these measures and some descriptive statistics based on these measures. This routine **only handles symmetric data** and in these circumstances the eigenvalues provide a measure of the accuracy of the centrality measure. To help interpretation the routine calculates all positive eigenvalues but only gives the eigenvector corresponding to the largest eigenvalue.

Eigenvector Centrality (from Newman and Brin and Page)

- Each node has a weight x_i that is defined to be proportional to the weights of all nodes that point to the node (i)
- And $x_i = \lambda^{-1} \sum_j A_{ij} x_j$

- And then λ x
- Thus the weights are an eigenvector of the adjacency matrix (A) with eigenvalue λ

Florentine Families Centrality Metrics (with Eigenvector Centrality)

	$C'_D(n_i)$	$C'_C(n_i)$	$C'_B(n_i)$	$C'_I(n_i)$	
Acciaiuoli	0.071	0.368	0.000	0.049	.19
Ablizzi	0.214	0.483	0.212	0.074	.35
Barbadori	0.143	0.438	0.093	0.068	.30
Bischeri	0.214	0.400	0.104	0.074	.40
Castellani	0.214	0.389	0.055	0.070	.37
Genori	0.071	0.333	0.000	0.043	.11
Guadagni	0.286	0.467	0.255	0.081	.41
Lamberteschi	0.071	0.326	0.000	0.043	.12
Medici	0.429	0.560	0.522	0.095	.61
Pazzi	0.071	0.286	0.000	0.033	.06
Peruzzi	0.214	0.368	0.022	0.069	.39
Pucci	---	---	---	---	0
Ridolfi	0.214	0.500	0.114	0.080	.48
Salvati	0.143	0.389	0.143	0.050	.20
Strozzi	0.286	0.438	0.103	0.070	.50
Tornabuoni	0.214	0.483	0.092	0.080	.46
Centralization	**0.257**	**0.322**	**0.437**	**---**	.43

Centrality II

- Numerous metrics exist in the Social Networks Literature for assessing the "centrality" of a social network.
  - Centrality metrics attempt to characterize the level of "centralization" of control or action on this network
  - One application is to assess how important a given actor (node) is in the network
  - Another application is to assess overall how much of the control of the network is controlled by the "more important" actors
  - The relative importance of single channels/links and groups of links has also been of interest.
- Centrality utility:
  - The calculation methods have been applied in search, navigation and community structure models but otherwise the "Network Science" Community does not utilize these measures. CM bias is that they are probably useful in social and other networks.
  - Hidden Hierarchy, robustness –communication and other meanings are all dependent on effects such as those defined and some of these measures (***betweenness and eigenvector centrality***) ***deserve more attention in modern network analysis***.

Related newer Centrality-like metrics

- Jon Kleinberg (Computer Science at Cornell) has done much of the leading work in search and navigation (more later).
- In some of his earliest work on this topic (1997-1999), he "invented" some useful new metrics for looking at important nodes (particularly on directed networks and probably most useful in the domain he was interested in-- the www)

- He looked for ways to find related sets of "Authorities" and "Hubs" and differentiated these from single "high in-degree nodes"

Prestige and Acquaintance Calculation

Authority: not only referred to by many nodes, but also by many Hubs. (measurement: ***prestige***)

Hub: not only refers to many nodes, but also to many Authorities.
(measurement: ***acquaintance***)

A : adjacency matrix
x_i : prestige (of node i)
y_i : acquaintance

$$Ax = \lambda y$$
$$A^T y = \mu x$$

Solve for x and y.

These metrics proved useful in directed citation networks (Mo-Han Hsieh thesis work on Internet Standards)

Lecture 7: Quantitative Aspects Of Networks III: Outline

- Some Social Network Concepts-intuition and calculation
  - clustering (transitivity)
  - centrality
    - degree, closeness, betweenness, information, eigen
  - prestige and acquaintance
- **degree distributions**
  - skew (and non-skew) distributions
  - fitting power laws to observed data
  - the normality of power laws
  - truncation
  - Structural implications and growth assumptions

Degree Distributions

- Define p_k as the fraction of nodes in a network with degree k. This is equivalent to the probability of randomly picking a node of degree k
- A plot of p_k can be formed by making a histogram of the degrees of the nodes. **This is the degree distribution of the network.**
- Histograms
  - Normal (and nearly so)
  - Skewed (and heavily skewed)
- Suggest some normal or nearly normal distributions..and some not likely to be normal

Degree Distributions II

- Define p_k as the fraction of nodes in a network with degree k. This is equivalent to the probability of randomly picking a node of degree k
- A plot of p_k can be formed by making a histogram of the degrees of the nodes. **This is the degree distribution of the network.**
- Histograms
  - Normal (and nearly so)
  - Skewed (and heavily skewed)
- Reasons for normal vs. skewed?

- Power law (skewed)

$$p_k \sim k^{-\alpha}$$

- Plot ln p_k vs. ln k, slope = α

Why might cumulative plot be superior?

Comparison of Models with Structural Metrics : Degree distribution

- Does the existence of a power law for degree distributions for networks indicate existence of a specific mechanism for formation?
  - ***No***, power laws are consistent with a wide variety of mechanisms for network formation (Newman, "Power laws, Pareto distributions and Zipf's law"2004/5)
- Does the existence of power laws for degree distributions for networks indicate the existence of a certain kind of structure for the network?
  - ***No***, power laws are consistent with a wide variety of networks having various structures and some without central hubs (Li et al)
  - Moreover, power laws are ***the equivalent of normal distributions at high variation*** (Samorodnitsky and Taqqu)

Central Limit theorem

- The central limit theorem states that given a distribution with a mean μ and variance σ^2, the sampling distribution of the mean approaches a normal distribution with a mean (μ) and a variance σ^2/N as N, the sample size, increases. The amazing and counter-intuitive thing about the central limit theorem is that no matter what the shape of the original distribution, the sampling distribution of the mean approaches a normal distribution. Furthermore, for most distributions, a normal distribution is approached very quickly as N increases.

Central Limit Theorem

The mean of a sequence of *n* iid random variables with

- **Finite** μ (and variance)

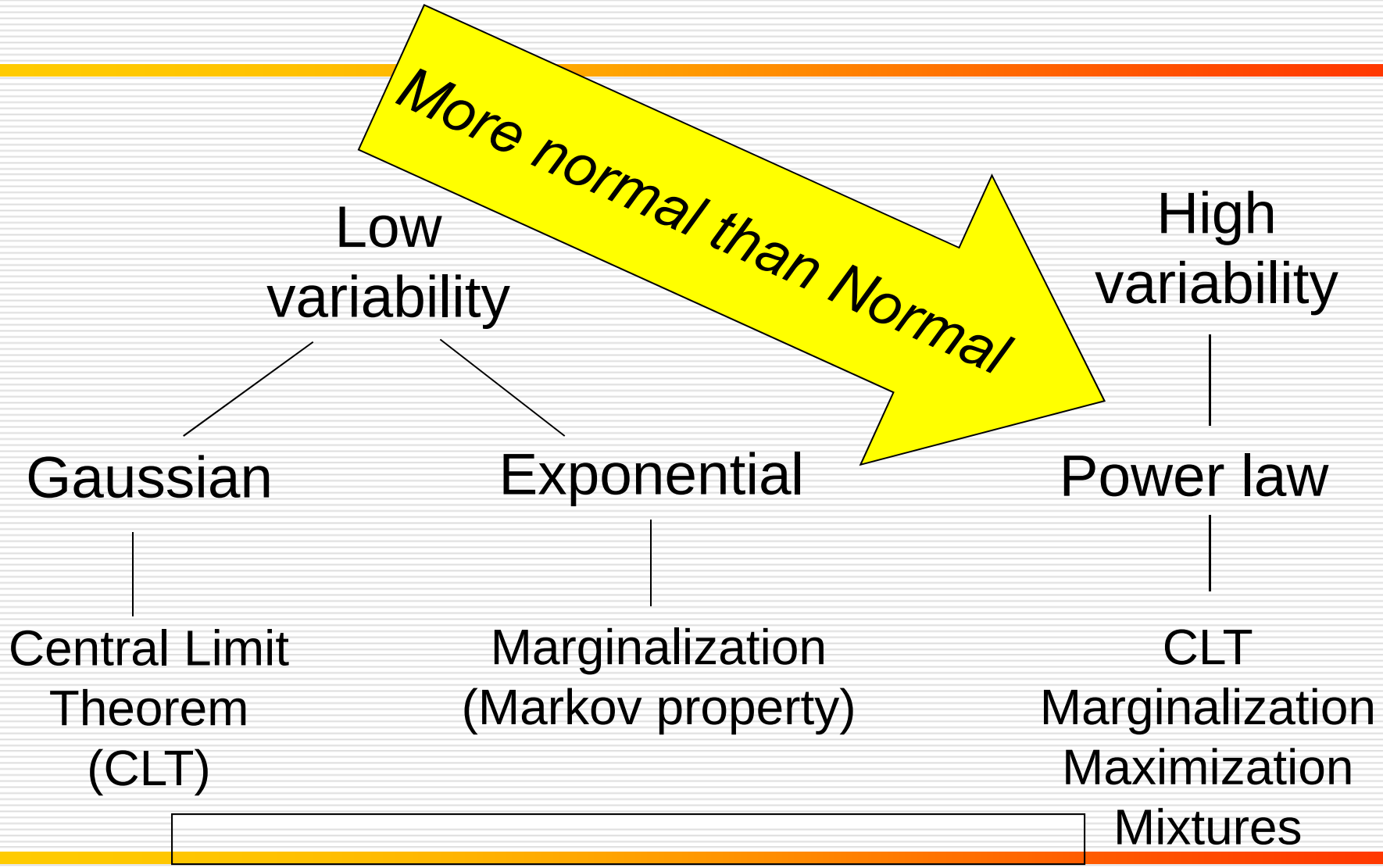
$$E\left(\left|x_i - E(x_i)\right|^{2+\delta}\right) < \infty \quad \delta > 0$$

approximates a normal distribution in the limit of a large *n*.

Marginal and Markov process defined

- Marginal probability- In a multivariate distribution, the probability of one variable, or function of several of these variables, taking a specific value (or falling in a range)
  - Metric: μ An **outer measure** on a product space, by restriction to one of the two factors: if α is an outer measure on X x Y , the marginal probability is a measure that satisfies
$$\alpha(A) = \mu(A\,x\,Y)$$
- Markov chain or process. A sequence of events, usually called states, the probability of each of which is dependent only the event immediately preceding it.

Power laws are ubiquitous



Comparison of Models with Structural Metrics : Degree distribution

- Does the existence of a power law for degree distributions for networks indicate existence of a specific mechanism for formation?
  - ***No***, power laws are consistent with a wide variety of mechanisms for network formation (Newman, "Power laws, Pareto distributions and Zipf's law"2004/5)
- Does the existence of power laws for degree distributions for networks indicate the existence of a certain kind of structure for the network?
  - ***No***, power laws are consistent with a wide variety of networks having various structures and some without central hubs (Li et al)
  - Moreover, power laws are the equivalent of normal distributions at high variation (Samorodnitsky and Taqqu)
- → **Power laws are *very* useful for representation and manipulation of data but are *not at all* indicative of structure or behavior (despite what you may read)**

Degree Distributions III

- Define p_k as the fraction of nodes in a network with degree k. This is equivalent to the probability of randomly picking a node of degree k
- A plot of p_k can be formed by making a histogram of the degrees of the vertices. **This is the degree distribution of the network.** Some distributions
- Random Graph- binomial

$$p_k \cong \frac{< k >^k e^{-<k>}}{k!}$$

(poisson at large n)

- Exponential

$$p_k \approx e^{\frac{-k}{\kappa}}$$

- Power Law

$$p_k \sim k^{-\alpha}$$

- lognormal

$$p(\ln k) \approx \exp(-\frac{(\ln k - \mu)^2}{2\sigma^2})$$

Degree Distributions IV

- Other Distributions
  - Power law with exponential cutoff is "common"
  - For bipartite graphs, there are two degree distributions, one for each type of node (multipartite one for each type of node)
  - For directed graphs, each node has an in-degree and an out-degree and the degree distribution becomes a function of two variables (j and k for in and out degrees). Since in and out degrees can be strongly correlated, the joint distribution also contains information about the network.
- Maximum Degree (Power Law) $k_{\max} \sim n^{1/(\alpha-1)}$

Network Metrics (from lectures 2, 3, 4 and now lecture 7)

- n, the number of nodes
- m, the number of links
- 2m/n is the average degree <k> as the number of links on a given node, k, is the degree.
- m/[(n)(n-1)] or <k>/(n-1)is the "sparseness" or normalized interconnection "density"
- Path length, *l*

$$l = \frac{1}{\frac{1}{2}n(n-1)} \sum_{i \geq j} d_{ij}$$

- Connectivity
- Clustering (2 definitions)
- Centrality (5 definitions + prestige and acquaintance)
- Degree Distribution
- Compare some systems (See Table 2 in Newman review article)

Networks structural characteristics: Preliminary summary of results

- Most measures-even simple ones- show that real systems (represented as networks) have "structure" (linking regularities beyond random).
- Real system architectures will not be describable by a single structural metric or feature. One must consider, size, sparseness, degree distribution, transitivity (and probably centrality and others) ***simultaneously*** in order to ***begin*** to understand a specific complex system and its similarities/differences from other complex systems.
- Although there are numerous metrics available, these are not necessarily (or even likely) to be the simplest or best to describe the systems we are interested in compactly.
- ***However***, invention of new characteristics without fully understanding and exploring existing metrics is most likely to introduce unnecessary confusion rather than enlightenment (the 2 clustering metrics is an example)

References for Lecture 7

- **Overall key references**
  - Wasserman and Faust, *Social Network Analysis: Methods and Applications*, Cambridge University Press (1994)
  - M. E. J. Newman, "The structure and function of complex networks" *SIAM Review* vol. **45**, 167-256 (2003)
  - J. Scott, *Social Network Analysis: A Handbook* Sage Publications (2000)
- For Centrality related
  - W & F (above) plus UCINET help and Hanneman book
  - Jon. M. Kleinberg "Authoriative Sources in a Hyperlinked Environment" Journal of the ACM, Vol. **46,** no. 5, 1999,pp 604-632
- For Power Laws
  - M. E. J. Newman, "Power Laws, Pareto Distrubutions and Zipf's law, cond-mat/0412004v2
  - Samorodnitsky,G. and Taqqu, S., *Stable Non-Gaussian Random Processes: Stochastic Processes with Infinite Variance*, Chapman and Hall, London, (1994)
  - A Barabasi and R. Albert, "The Emergence of Scaling Laws in Random Networks", *Science* **286**, pp 509-512 (1999)
  - Amaral, L. A. N., Scala, A., Bertelemy, M. and Stanley, H. E. "Classes of Small World Networks", Proc. Nat. Acad. Sci. **97**, 11149-52 (2000)

ESD.342 Network Representations of Complex Engineering Systems

Spring 2010