Lecture Notes: Social Networks: Models, Algorithms, and Applications
Lecture 3: Jan 24, 2012
Scribes: Geoffrey Fairchild and Jason Fries

1 Degree Distributions

Last time, we discussed some graph-theoretic terminology. Specifically, we discussed degree distributions. Recall the definition of degree distribution:

Definition 1 *For a graph $G = (V, E)$ and for $k = 1, 2, ...,$ the degree distribution of G is $p_k =$ fraction of vertices with degree k. In other words, if we pick a vertex $v \in V$ uniformly at random, $P[deg(v) = k] = p_k$.*

We specifically discussed the power law degree distribution which has degree distribution $p_k = C \cdot k^{-\alpha}$ where C is a constant and $\alpha > 1$. While not all degree distributions will be power law, many of the degree distributions of observed networks will be heavy-tailed or long-tailed distributions. A heavy-tailed distribution is a distribution that is "heavier" than the exponential distribution. Here, "heavier" means that the tail is much longer than the exponential distribution (which drops off to zero rapidly). The exponential distribution takes the form $f(x, \lambda) = \lambda e^{-\lambda x}$ where λ is the *rate parameter*. Taking the log of the exponential distribution, we have $\log(f(x, \lambda)) = \log(\lambda) - \log(\lambda x)$.

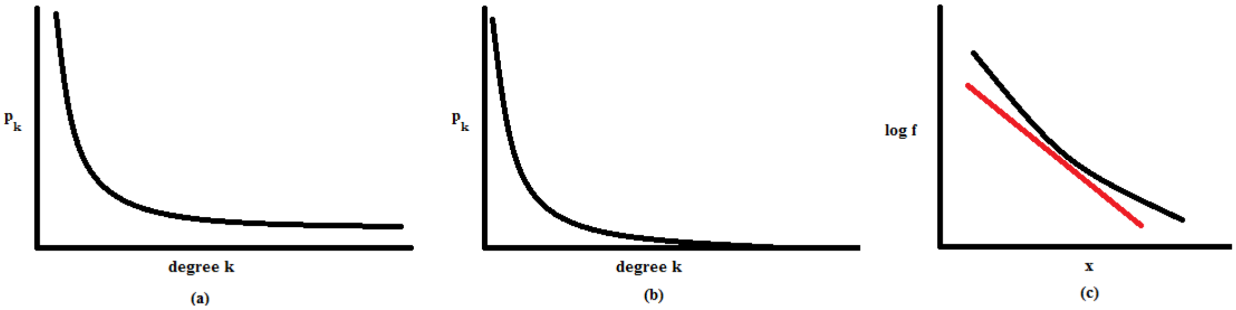

Figure 1: (a) is a typical long-tailed distribution observed in many networks. In general, there are few vertices with high degree while most vertices have low degree. (b) is a typical exponential distribution. Note the difference between the long-tailed distribution and the exponential distribution: the exponential distribution nears zero much faster. (c) is a log plot showing the difference between the exponential and long-tailed distributions. The exponential distribution, shown in red, is linear with slope $-\lambda$ on this scale, while the long-tailed distribution, shown in black, is not linear.

2 Clustering Coefficient

One property of a graph is the *clustering coefficient*. The clustering coefficient is the measure of the the extent to which one's friends are also friends of each other. This measure has become popular due to a 1998 paper in Nature by Watts and Strogatz [4]. This property is sometimes called the *local clustering coefficient*.

Definition 2 *Given a graph $G = (V, E)$ and a vertex $v \in V$, the clustering coefficient of v is $cc_1(v) = \frac{\text{number of pairs of neighbors connected by edges}}{\text{number of pairs of neighbors}}$. To compute the clustering coefficient for a graph G, simple average $cc_1(v)$ for all $v \in V$.*

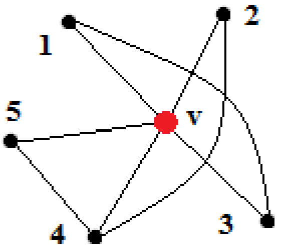

Figure 2: Here is a sample graph which we can use to illustrate the calculation of $cc_1(v)$. 3 of v's neighbors are connected to each other (`1-3`, `2-4`, and `4-5`). There are a total of $\binom{5}{2}$ pairs of neighbors. Thus, $cc_1(v) = \frac{3}{\binom{5}{2}} = \frac{3}{\frac{5!}{3!\cdot 2!}} = \frac{3}{10} = 0.3$.

An alternative clustering coefficient calculation, sometimes referred to as the *global clustering coefficient* or *transitivity*, was proposed by Newman, Strogatz, and Watts in 2001 [2].

Definition 3 *Given a graph $G = (V, E)$, the clustering coefficient of G is $cc_2(G) = \frac{\text{number of closed 2-paths}}{\text{number of 2-paths}}$.*

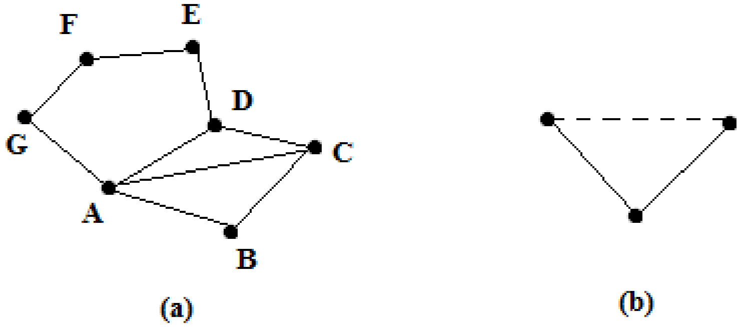

Figure 3: (a) is used to show several examples of 2-paths: `A-B-C`, `A-D-E`, and `C-B-A`. Note that `A-B-C` and `C-B-A` are considered different 2-paths and would thus both be present in the calculation of $cc_2(G)$. Examples of closed 2-paths are `A-B-C`, `A-C-B`, and `A-C-D`. (b) clarifies the notion of a closed 2-path. A closed 2-path is a 2-path where the end nodes are connected, forming a triangle. The global clustering coefficient is therefore asking this question: what fraction of situations presented in (b) contain the 3rd edge?

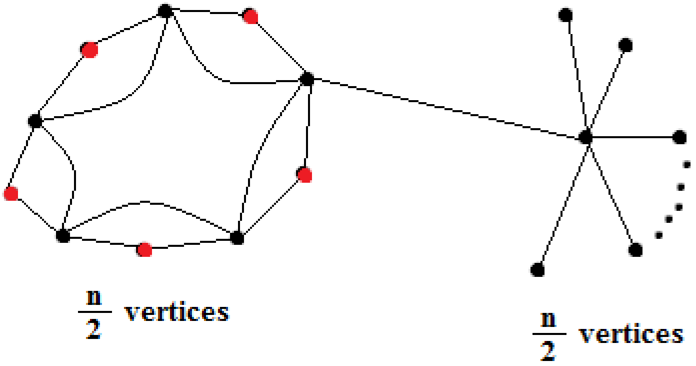

Figure 4: This graph demonstrates how the local clustering coefficient (cc_1) can differ from the transitive clustering coefficient (cc_2). The left circular cluster contains $\frac{n}{2}$ of the vertices in the entire graph. Each red node $r \in V$ has $cc_1(r) = 1$ because each red node only has two neighbors which are always connected to each other. Because the red nodes comprise half of the left circular cluster, we have $\frac{n}{4}$ nodes in the entire graph with $cc_1 = 1$. This then tells us that $cc_1(G) \geq \frac{1}{4}$. The number of closed 2-paths is $\frac{n}{4}$ because each red node is part of a triangle. However, the right star cluster (which has $\frac{n}{2}$ nodes) contains many 2-paths that aren't closed. Therefore, $cc_2(G) = \frac{\frac{n}{4}}{...+\binom{\frac{n}{2}}{2}} \approx \frac{\frac{n}{4}}{...+n^2} \approx \frac{1}{n}$. Thus, $\lim_{n\to\infty} cc_2 = 0$.

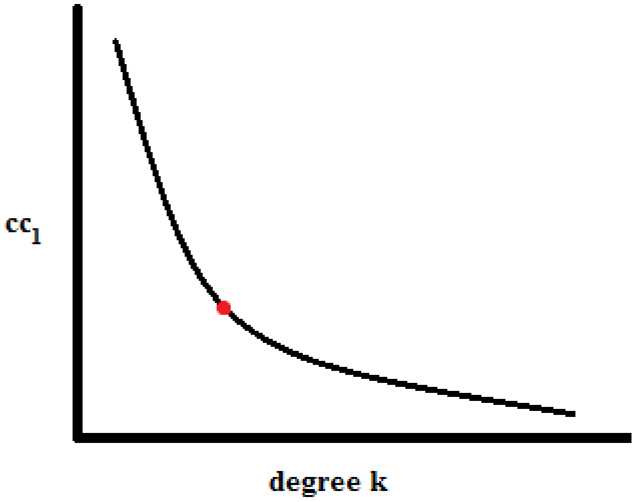

Figure 5: This graph shows the primary complaint about the local clustering coefficient. Each point on this line (such as the red dot) represents the average cc_1, averaged over all degree k vertices. Vertices with low degree can inflate the value of cc_1 for a graph (e.g., in the case of Figure 4).

	N	average degree	cc_1	cc_1 of corresponding random graph
actors network	225226	61	0.79	0.00027
power grid	4941	2.67	0.080	0.005
C. elegans	282	14	0.28	0.05

Table 1: Comparing observed networks against "corresponding" random graphs (defined formally in the notes from 1/26/12), we see that observed networks tend to have much higher clustering coefficients. In general, cc_1 of the random graph is at least one order of magnitude smaller than the value for the observed graph.

Why is the value of cc_1 so much higher for the observed graph? Sociologists offer several possible explanations:

- **homophily** – The tendency of individuals to connect to other "similar" individuals. Note that this is an external effect; the effect is not present because of the edge between the two people.
  - socioeconomic class
  - coworkers
  - etc.
- **network effects** – The fact that nodes `u` and `w` both know `v` (but not each other) could imply some "trust" between `u` and `w`. Additionally, the fact that `u` and `w` don't know each other induces a "latent stress" on `v` which produces some "incentive" for `v` to work towards a `u-w` edge.

3 "Giant" Components

Definition 4 *A giant component in a graph is a maximal subgraph in which each vertex pair is connected by a path.*

Most observed networks have a "giant" component containing $\geq 90\%$ of the vertices. In a study of a network created by 449913 actors in 2000, 440971 actors ($\sim$98%) were part of the giant component. Why does this happen? Figure 6 helps explain this phenomenon.

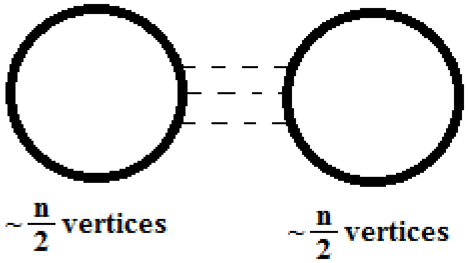

Figure 6: Suppose we have two components of a graph, each with $\sim \frac{n}{2}$ nodes. There are a total of $\sim \frac{n^2}{4}$ possible edges in this graph, so the probability of having an edge that connects these two components is very high. It is more likely that two separate components will exist without an edge connecting them only if one of the components is much smaller than the other.

4 Types of Networks

There are a variety of “observable” networks around us. Different types of networks, and even networks within the same genre, can be radically different in structure.

- biological networks
  - metabolic networks
  - neural networks
  - protein-protein interaction networks
  - ecological networks (predator-prey networks)
- technological networks
  - internet graph
  - power grid
  - telephone networks
  - transportation networks
- information networks
  - world wide web graph
  - peer-to-peer graphs (e.g., BitTorrent)
- social networks
  - “Southern Women Study” [1]
  - prominent families in renaissance Florence (1400-1434) [3]
- social+information networks
  - Facebook
  - Twitter
  - email

Social network analysis and data isn’t new. In 1941, sociologists Davis et al. published a book containing their “Southern Women Study” [1]. They used newspaper information to construct a bipartite graph connecting people to places/events like what’s shown in Figure 7. Using information like this, it would be trivial to extract the underlying people-people graph. Padgett et al. analyzed renaissance-era data about the Medici family in Florence, Italy [3]. Using marriage records, patronage connections, and other similar data, they set out to explain the rise of the Medici family.

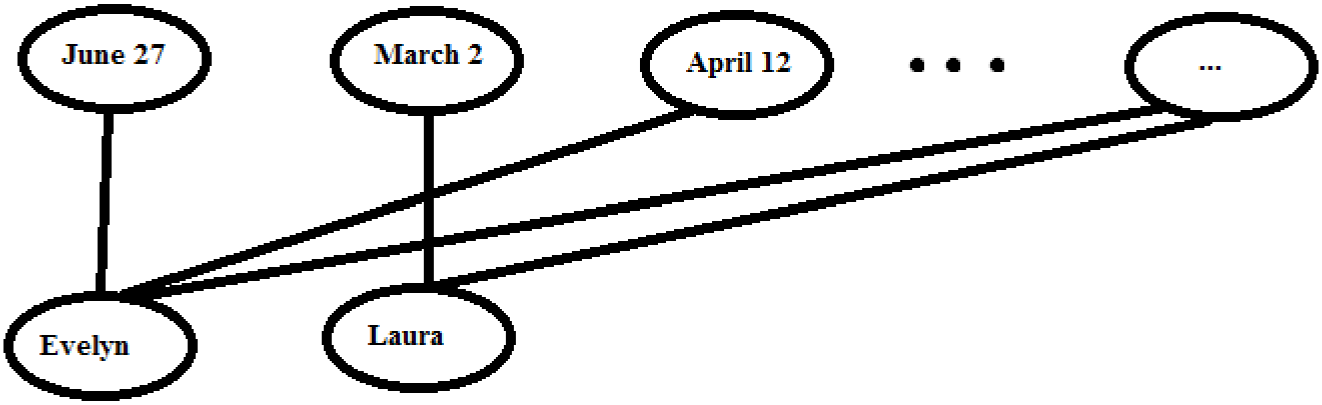

Figure 7: This is an example of the people-places bipartite graph described in the "Southern Women Study".

References

[1] Allison Davis, Burleigh B. Gardner, and Mary R. Gardner. *Deep South: A Social Anthropological Study of Caste and Class*. University of Chicago Press, 1941.

[2] Mark E. Newman, Steven H. Strogatz, and Duncan J. Watts. Random graphs with arbitrary degree distributions and their applications. *Physical Review E*, 64:026118, 2001.

[3] John F. Padgett and Christopher K. Ansell. Robust action and the rise of the medici, 1400-1434. *American Journal of Sociology*, 98:1259–1319, 1993.

[4] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442, 1998.