Complex Networks:

Overview, Basic Models, Community Detection

David P. Feldman

College of the Atlantic

and

Santa Fe Institute

dave@hornacek.coa.edu

http://hornacek.coa.edu/dave/

More slides and notes available at:

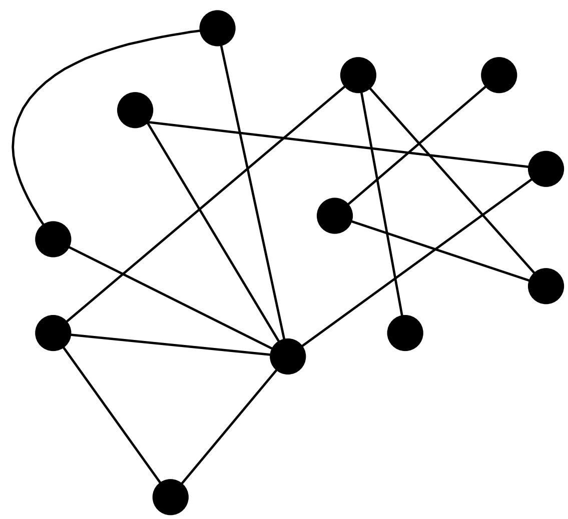
http://hornacek.coa.edu/dave/Teaching/Networks.08/

Outline

1. Basic definitions and concepts
2. Three classic, important, and fun models
   (a) Erdös-Rényi random graphs
   (b) Watts-Strogatz small world model
   (c) Barabasi-Albert preferential attachment
3. Higher-order structure: community detection
4. Quick introduction to network workbench.

What is a Network?

1. A collection of **nodes**
2. A collection of **edges** connecting nodes



- A network model (usually) treats all nodes and links the same
- In a picture of a network, the spatial location of nodes is (usually) arbitrary
- Networks are abstractions of connection and relation
- Network models emphasize topology and connection and not the function or individuality of nodes.

Network Questions: Structural

Given a network, there are a number of structural questions we may ask:

1. How many connections does the average node have?
2. Are some nodes more connected than others?
3. Is the entire network connected?
4. On average, how many links are there between nodes?
5. Are there clusters or groupings within which the connections are particularly strong?
6. *What is the best way to characterize a complex network?*
7. *How can we tell if two networks are "different"?*
8. *Are there useful ways of classifying or categorizing networks?*

Network Questions: Dynamics *of*

Things are the way they are because they got that way. (Richard Levins.)

1. How can we model the growth of networks?
2. What are the important features of networks that our models should capture?
3. Are there “universal” models of network growth? What details matter and what details don’t?
4. *To what extent are these models appropriate null models for statistical inference?*
5. *What’s the deal with power laws, anyway?*

Network Questions: Dynamics *on*

1. How do diseases/computer viruses/innovations/rumors/revolutions propagate on networks?
2. What properties of networks are relevant to the answer of the above question?
3. If you wanted to prevent (or encourage) spread of something on a network, what should you do?
4. What types of networks are robust to random attack or failure?
5. What types of networks are robust to directed attack?
6. *How are dynamics of and dynamics on coupled?*

Network Questions: Algorithms

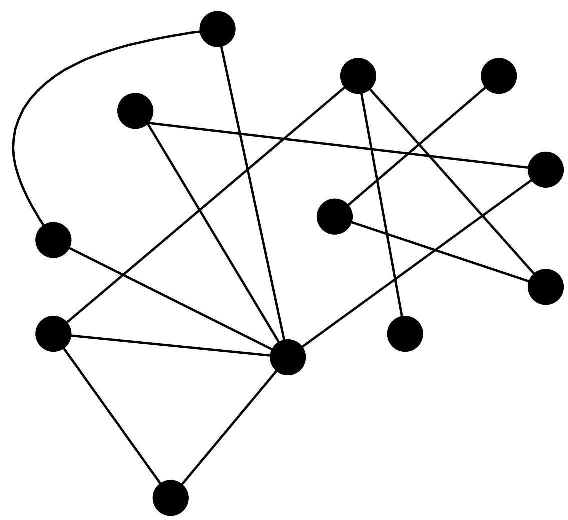
1. What types of networks are searchable or navigable?
2. What are good ways to visualize complex networks?
3. How does google page rank work?
4. If the internet were to double in size, would it still work?

There are also many domain-specific questions:

1. Are networks a sensible way to think about gene regulation or protein interactions or food webs?
2. What can social networks tell us about how people interact and form communities and make friends and enemies?
3. Lots and lots of other theoretical and methodological questions...
4. What else can be viewed as a network? Many applications await.

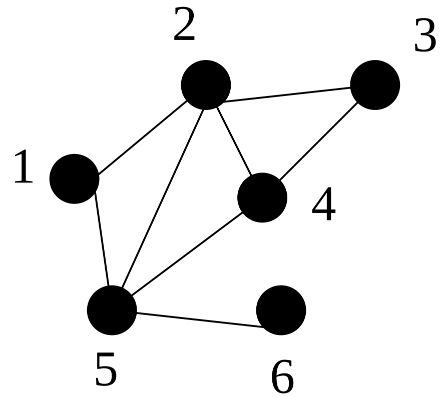
What is a Network?

1. A collection of **nodes**
2. A collection of **edges** connecting nodes



- Let $N =$ number of nodes.
- Let $M =$ number of links or edges.
- Networks are also knows as graphs, particularly among mathematicians.

Network Representation: Adjacency Matrix

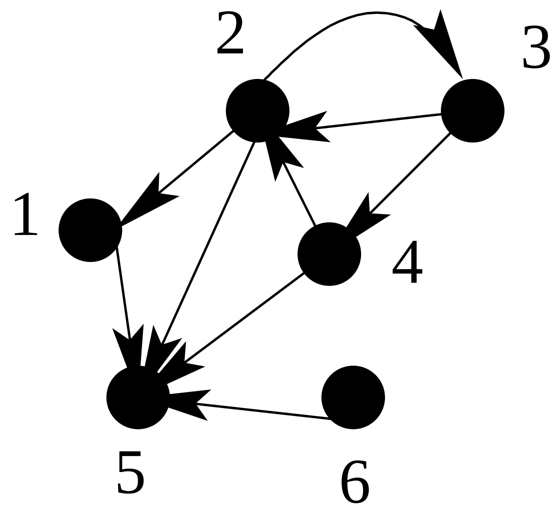


- Adjacency matrix A: $A_{ij} = 1$ if there is a link between nodes i and j. Otherwise $A_{ij} = 0$.
- For the graph shown above:

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix} \tag{1}$$

- Note that A is symmetric

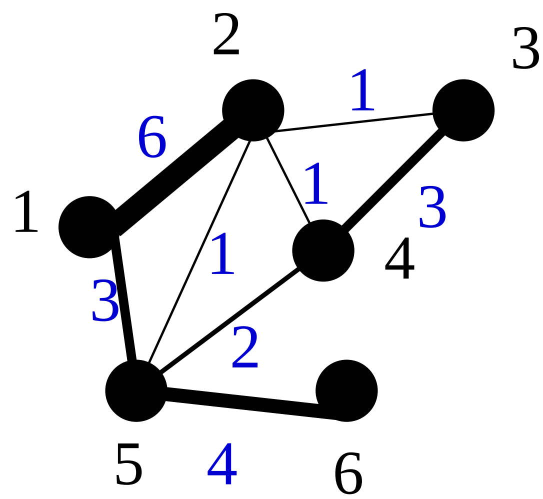
Variation: Directed Network



- Links have direction. Adjacency matrix is no longer symmetric.
-

$$A = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix} \quad (2)$$

Variation: Weighted Network



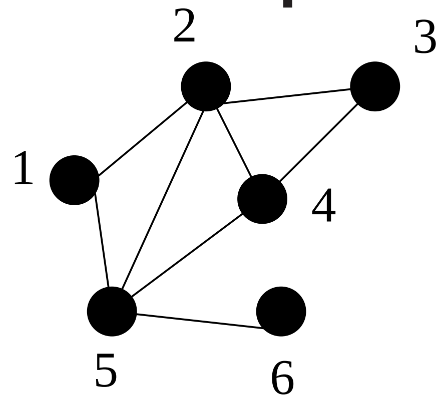
- Links have weights, indicating differint strengths of connection. Adjacency matrix is no longer all 1's and 0's
-

$$A = \begin{pmatrix} 0 & 6 & 0 & 0 & 3 & 0 \\ 6 & 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 3 & 0 & 0 \\ 0 & 1 & 3 & 0 & 2 & 0 \\ 1 & 3 & 0 & 2 & 0 & 4 \\ 0 & 0 & 0 & 0 & 4 & 0 \end{pmatrix} \tag{3}$$

Basic Network Properties

- Given a network, what are some useful ways of describing its connectivity, organization, structure, etc?
- I will present a few basic and quite standard definitions.
- I will talk about regular networks, but most of the quantities generalize fairly naturally to directed and/or weighted networks.

Basic Network Properties: Degree

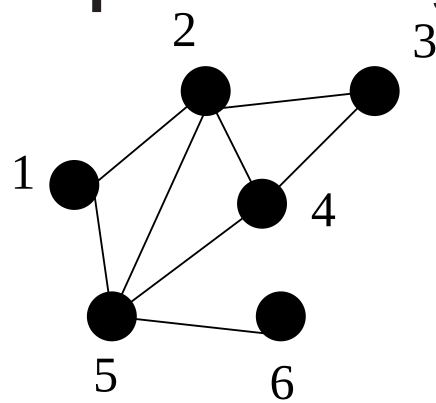


- The *degree* k of a node is the number of links connected to it.
- The degree is sometimes called *coordination number* and denoted with z. This is mostly a physics convention.
- Ex: $k_1 = 2$, $k_2 = 4$, $k_6 = 1$.
- Often we are interested in the average degree of all the nodes.
- This is often denoted k or $\langle k \rangle$. The latter is called "the expectation value of k."
- For this graph, $k = 2.67$.
- There is a hard and an easy way to calculate k.

Basic Network Properties: Degree Distribution

- We are usually interested in more than just the average degree.
- Are some nodes more connected than others? How much variance is there about the mean degree?
- For that matter, is the notion of an average degree or variance even meaningful?
- These questions can be addressed by looking at the *degree distribution*.
- $P(k)$ is the probability that a randomly chosen node has degree k.

Basic Network Properties: Degree Distribution



- The *degree* k of a node is the number of links connected to it.
- The degree is sometimes called *coordination number* and denoted with z. This is mostly a physics convention.
- Ex: $k_1 = 2$, $k_2 = 4$, $k_6 = 1$.
- Often we are interested in the average degree of all the nodes.
- This is often denoted k or $\langle k \rangle$. The latter is called "the expectation value of k."
- For this graph, $k = 2.67$.
- There is a hard and an easy way to calculate k.

Basic Network Properties: Degree Distribution

- We are usually interested in more than just the average degree.
- Are some nodes more connected than others? How much variance is there about the mean degree?
- For that matter, is the notion of an average degree or variance even meaningful?
- These questions can be addressed by looking at the *degree distribution*.
- $P(k)$ is the probability that a randomly chosen node has degree k.

Basic Network Properties: Distance and Diameter

- Basic Network Properties: Distance and Diameter
- Distance d_{ij} between nodes i and j
- $d_{ij} =$ # of links along shortest path connecting i and j .
- This is also denoted $d(i, j)$ or (i, j).
- The mean distance ℓ is the average of the d_{ij}'s.
- ℓ may be thought of as a measure of the size of the network.
- The diameter d of a graph is defined as the largest d_{ij}.
- The diameter is another measure of the size of the network.
- A network is said to have the *small world* property if grows no faster than the log of the number of nodes: $\sim \log(N)$.

Basic Network Properties: Clustering

- To what extent are your friends friends with each other?
- The cluster coefficient C_i is defined as the fraction of your friends that are friends with each other.
- I.e.,

$$C_i = \frac{\#\text{ of friendships among i's friends}}{\#\text{ of potential friendships among i's friends}}$$

- Note: There are different definitions of clustering which are not identical.

Which Nodes are Important?

- Which nodes are the most important in a network?
- What different roles might nodes play?
- Measures of importance of a node are often called centrality.
- **Degree Centrality:** Key Idea: An important node is involved in many interactions.
  - The degree centrality of a node is simply its degree.
  - Thus, under this line of reasoning, the most important node is the one with the most connections.

Which Nodes are Important?

- **Closeness Centrality:** Key Idea: An important node is close to lots of other nodes.
  - Measure how far a node is from other nodes.
  - Nodes in the “middle” have higher centrality.

- **Betweenness Centrality:** Key Idea: An important node connects lots of other nodes. I.e., an important node will be on a high proportion of paths between other nodes.

- To calculate $C_b(i)$, the betweenness centrality for node i:
  1. Consider all pairs of nodes j, k $\neq$ i.
  2. Determine the shortest path between all such j, k .
  3. Then $C_b(i)$ = fraction of those paths which go through i.

Highly Schematic Picture of Order and Disorder

ORDER **DISORDER**

Crystal Structures

Exact Symmetries

Group Theory

Abstract Algebra

Regular Graphs, Lattices

Ideal Gases

Tossing Coins (IID Processes)

Unpredictability

Chaos, Mixing, etc.

Erdos−Renyi Model, Random Graphs

- There are well understood mathematical techniques for studying the extremes of order and disorder.
- Intermediate regions are harder. Often one starts at one extreme and then perturbs or expands off that extreme to get approximate solutions.

Networks and Graphs after Erdős and Rényi

- A fair amount of work in sociology, social networks, economics, etc.
- Also work on computer and technological networks, engineering, etc.
- Then, in 1998, Duncan Watts and Steven Strogatz publish Collective Dynamics of ’Small-world” networks, Nature 393:440–442.
- This paper sparks a remarkable surge of interest in networks.
- Watts and Strogatz’s paper has been cited over 10,000 times.
- In 1999, Barabasi and Albert (re)-discover power laws in networks.
- Their paper, Emergence of Scaling in Random Networks, Science 286:509 has now been cited over 7,700 times.

Why this Sudden Surge in Networks Research?

In my opinion, this is due to a number of factors.

- Electronic data became available that wasn't available before.
- Advances in computing power.
- The idea of networks resonates with increased attention to connection, links, globalization, etc.
- Watts and Strogatz's model was very elegant and simple mathematically, and captured the imagination of a great many people.
- Once physicists became aware of networks, it was quickly realized that they were very well suited to a physics style of analysis.
- Arguably, there wasn't that much interesting and exciting going on in other areas of physics.

The Erdős Rényi Model

- The Model:
  1. Start with N nodes.
  2. Connect each pair of nodes with probability p.
- Questions:
  - Is the graph connected?
  - What is the degree distribution?
  - What is the size of the graph?
  - What is the clustering coefficient?
- Why might we care?
  - In science, we frequently need to ask, Could this have happened randomly, by chance?
  - In order to answer this question, we need to know about random graphs.

E-R Analysis: Degree Distribution is Poisson

- How many links does a node have? Each node gets $N-1$ potential links, and each chance yields a link with probability p.
- Thus, the degree distribution $\mathrm{P}(k)$ is:

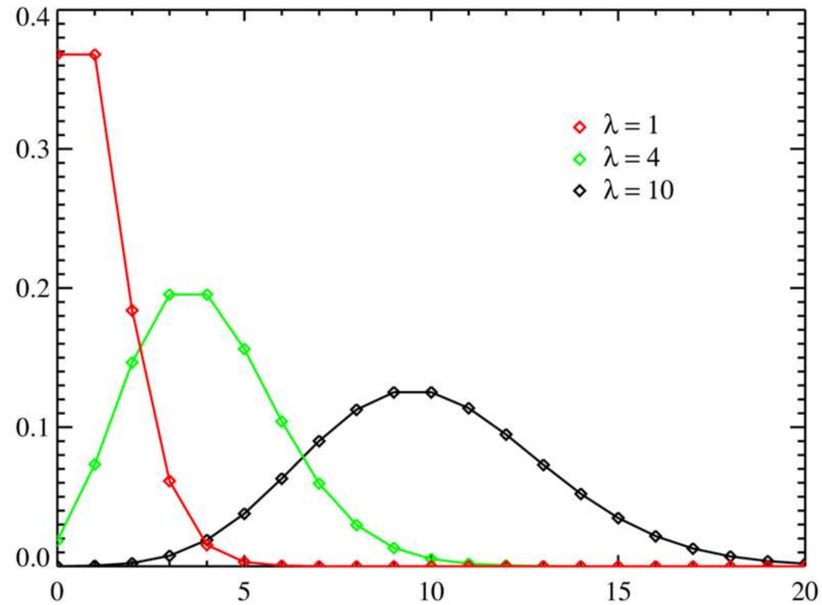
$$\mathrm{P}(k) = \binom{N-1}{k} p^{N-1}(1-p)^{N-1-k} \; . \tag{4}$$

- For large N, this equation becomes well approximated by:

$$\mathrm{P}(k) \approx \frac{z^k e^{-z}}{k!} \; , \tag{5}$$

- Where $z = p(n-1)$ is the mean degree.
- This is known as the Poisson distribution. It arises in many different applications, not just networks.

Poisson Distribution Properties



- Plot of Poisson Distribution $\mathrm{P}(x) = \frac{\lambda^k e^{-\lambda}}{k!}$ for three different λ values.
- Figure Source: `http://en.wikipedia.org/wiki/Image:Poisson_distribution_PMF.png`.
- Variance $=$ Mean $=$ λ.
- The distribution $P(k)$ decays *extremely* rapidly as k gets large—much faster than exponential!

ER Analysis: Clustering Coefficient

- The cluster coefficient is the fraction of your friends that are friends.
- Link probabilities in the ER model are independent.
- Thus, the probability that your friends are friends is just p.
- Hence, $C = p$.
- Conclusion: **Erdős-Rényi** graphs have small clustering coefficients.
- Almost all real-world graphs have clustering coefficients larger than would be expected for comparable ER graphs.

ER Analysis: Characteristic Path Length

- Let $z = np$ be the average degree.
- The number of nodes a distance d from any node is approximately x^d.
- How big must d be so that it includes all of the nodes in the graph? This value of d is ℓ, the characteristic path length:

$$z^\ell = n \quad \longrightarrow \quad \ell = \frac{\log n}{\log z} = \frac{\log n}{\log p + \log n} . \tag{6}$$

- Thus, ER graphs are "small-world," since ℓ grows logarithmically with n.
- Many real-world graphs have the small-world property.

ER Analysis: Is the Graph Connected?

- Roughly speaking, the graph undergoes a phrase transition as p is increased from being a collection of small connected fragments to a graph which has a giant connected component.
- A giant connected component is a connected component that is proportional to n in the large n limit.
- The critical parameter at which this occurs is, not surprisingly, $z = 1$.

ER Analysis: Connectivity Phase Transition

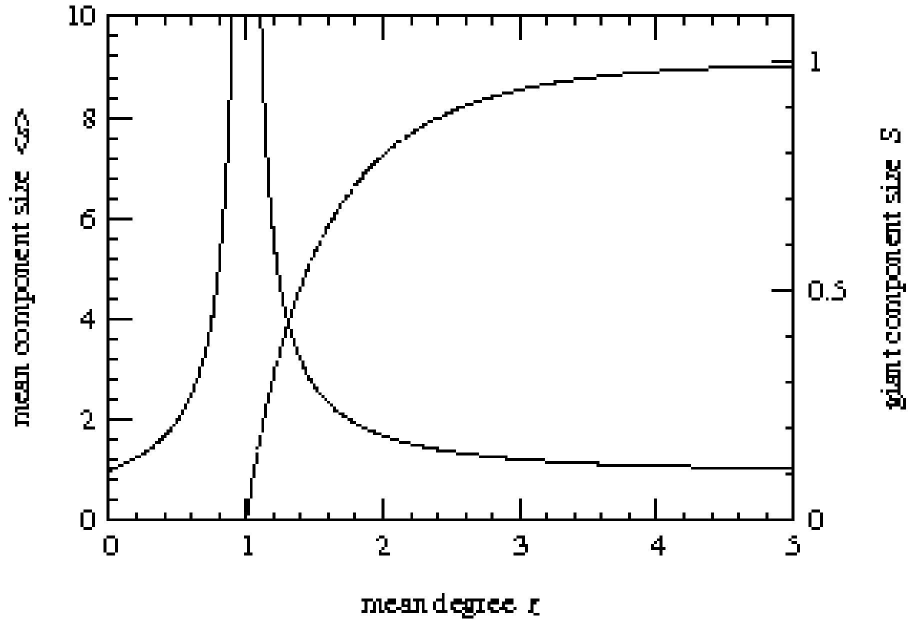


FIG. 10 The mean component size (solid line), excluding the giant component if there is one, and the giant component size (dotted line), for the Poisson random graph, Eqs. (20) and (21).

- Figure Source: M.E.J. Newman, The Structure and Function of Complex Networks, SIAM Reviews, 45(2):167–256, 2003.

Summary of Properties of Erdős-Rényi Model

- Degree distribution is Poisson
- Very low clustering
- Highly connected, “Small-world”
- Connectivity properties change discontinuously

Erdős-Rényi Model Conclusions

- Simple, tractable model of random graphs
- Not a realistic model, but a useful “straw man” or null model
- Does capture the small-world feature common in real-world networks
- Also has discontinuous changes, suggesting that other, more realistic models, might also have sharp thresholds
- Gives us intuition about what to expect from more complicated and realistic models

The Erdős Rényi Model

1. Start with N nodes.

2. Connect each pair of nodes with probability p.

- The mean degree is $z = Np$

- Note that there are a number of different ways to consider the large N limit.

- Often, we want N to get large while keeping z constant.

- In science, we frequently need to ask, Could this have happened randomly, by chance?

- In order to answer this question, we need to know about random graphs.

Summary of Properties of Erdős-Rényi Model

- Degree distribution is Poisson:

$$P(k) = \frac{z^k e^{-z}}{k!} \, . \tag{7}$$

- Very low clustering:

$$C = \frac{z}{N} \, . \tag{8}$$

- Highly connected, “Small-world”:

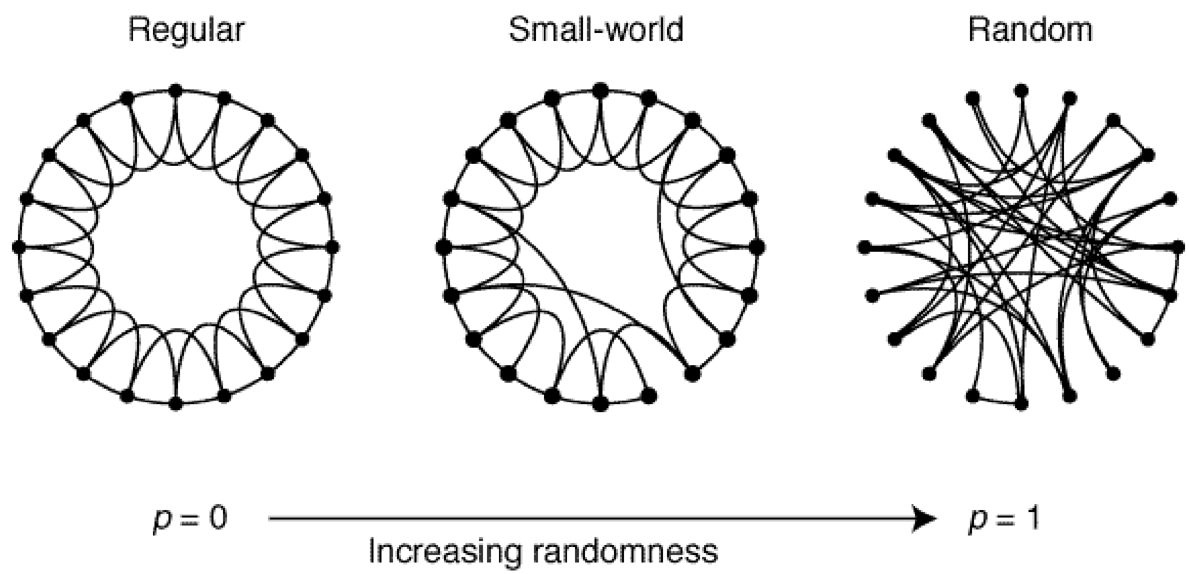
$$\ell \approx \log N \, . \tag{9}$$

- Connectivity properties change discontinuously as p is varied.

The Small-World Model

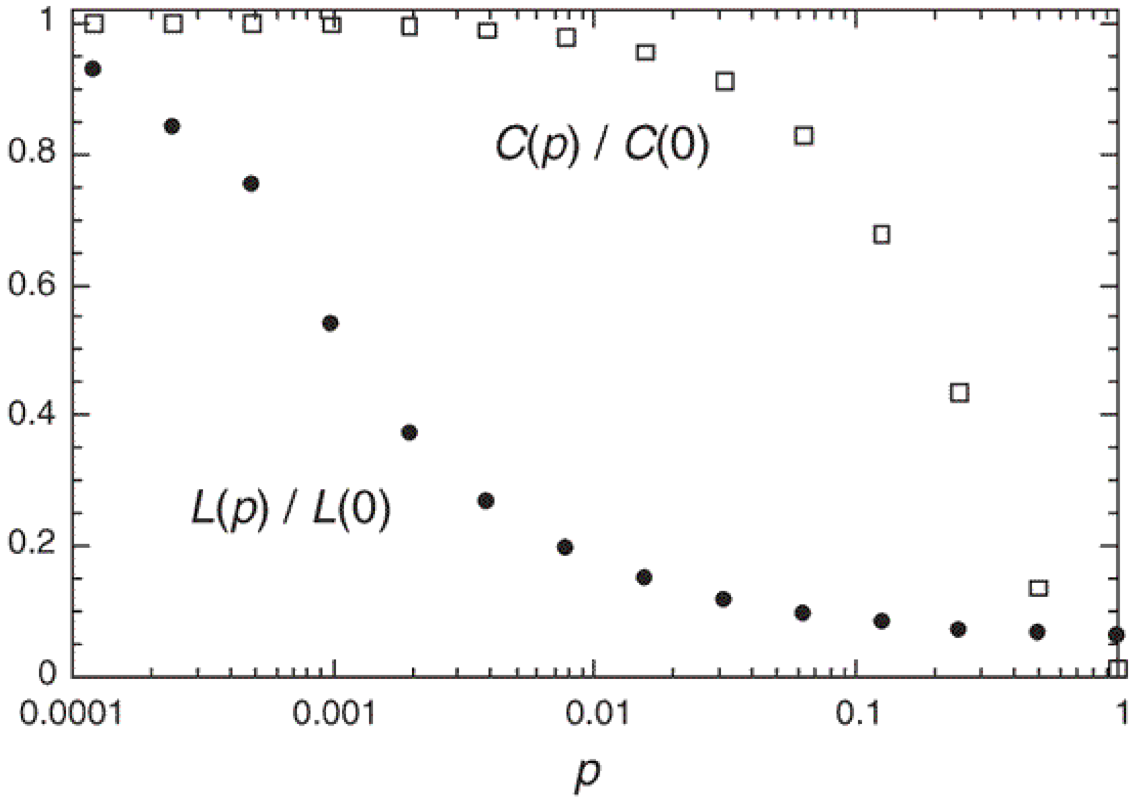
- The model:
  1. Begin with a regular lattice. Usually this is a one-dimensional ring, where each node has a few neighbors.
  2. Go through the regular lattice and consider each link.
  3. With probability p, rewire the link by random rewiring
- Initial question:
  1. How do C and ℓ vary with p?
- Watts and Strogatz, Nature 393:440–2. 1998.
- See also Newman, Models of the Small World," Journal of Statistical Physics 101:819-841. 2000.

Watts-Strogatz Model



- As p is increased the model moves from a regular graph, through intermediate graphs, to a random graph at $p = 1$.
- Figure source `http://www.nature.com/nature/journal/v393/n6684/fig_tab/393440a0_F1.html`

Watts-Strogatz Model: Basic Results



- There is a large intermediate region which shows "small-world" behavior: small ℓ but large C.
- Note the log scale on the horizontal axis.
- Figure source `http://www.nature.com/nature/journal/v393/n6684/fig_tab/393440a0_F2.html`

Watts-Strogatz: Preliminary Conclusions

1. The WS model shows a transition from a large-world to a small-world.
2. Disease models which have a probabilistic susceptibility to infection exhibit a sharp transition between epidemic and non-epidemic behavior.
3. Dynamical systems on small-world graphs exhibit behavior which is qualitatively different from behavior on regular graphs.
4. Many graphs show additional features (e.g., long-tailed degree distributions) which are not accounted for by the WS and similar models.
5. Nevertheless, the WS model qualitatively captures the small-world feature of many networks, and is a useful, albeit quite basic, model for a social network.

- Adapted from conclusions in Newman's 2003 review article.

Networks Growth and Dynamics

- “Things are the way they are because they got that way.” (Richard Levins)
- The Watts-Strogatz model sheds light on a static network.
- The WS network is *not* intended to be a model of how networks actually grow.
- The WS does, however, capture some aspects of some already-formed, real-world networks.
- What models are there for network growth, and what do they tell us?

Empirical Observation: Power-Law Degree Distribution

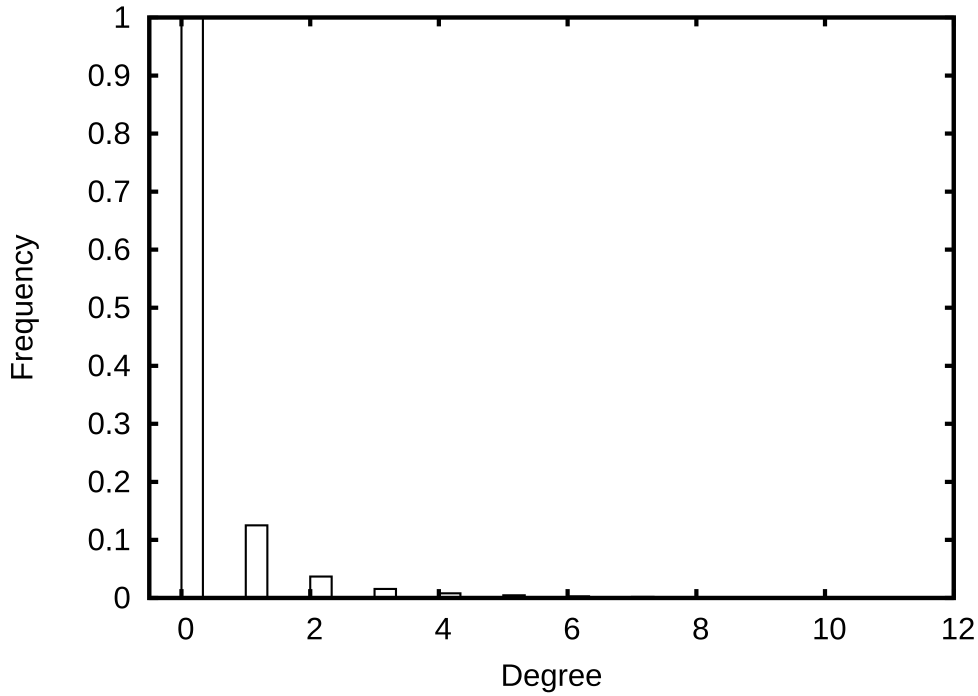
- It is well fairly established that many networks have a power-law degree distribution.
- A power-law distribution is one of the following form:

$$P(k) = ck^{-\alpha} \,, \tag{10}$$

where α is a positive constant, usually between 1 and 3.

- c is a constant that is adjusted to normalize the distribution.
- Often a distribution is called a power-law distribution even if it doesn't have exactly the above form, so long as it has this form for large k.
- What's usually of interest in these types of distribution is their large-k behavior.

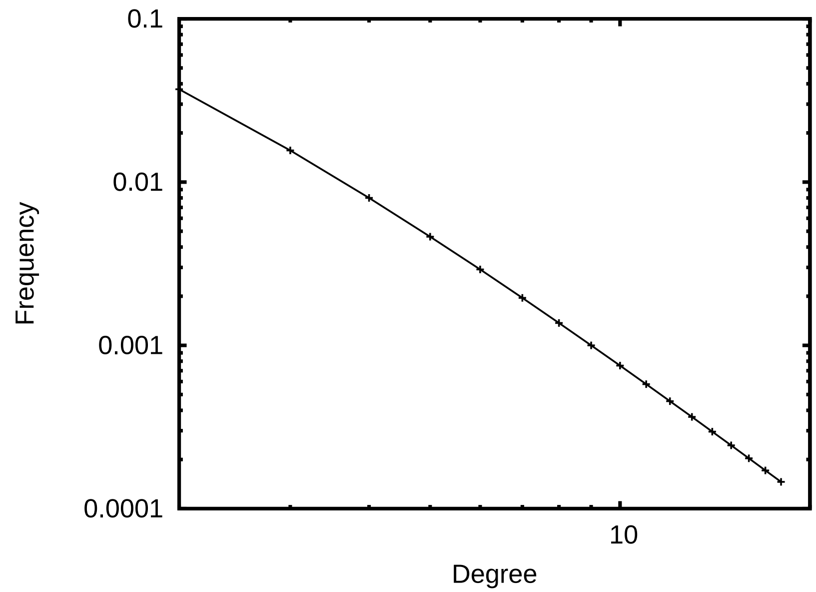
What do Power Law Distributions Look Like?



- Above is an (unnormalized) power law distribution for $\alpha = 3$.
- Note the relatively rapid decay of the degree k.
- However, the decay is much, much, much slower Poisson. For large k,

$$k^{-\alpha} >> \frac{1}{k!} \, . \tag{11}$$

What do Power Law Distributions Look Like?



- A power-law distribution is linear on a log-log plot.
- To see this, take the log of both sides of Eq. (11):

$$\log p \;=\; \log c - \alpha \log k \;. \qquad (12)$$

- The slope of the line is the exponent α.
- Note: This is resoundingly not the best way to estimate α. Van will talk about this.

Why Might we Care About Power Laws?

- Van will talk about this. For now, just a few initial thoughts.
- Power laws are very different from Poisson and Gaussian distributions.
- The probability of extreme events is much, much larger for PLs than Gaussians and Poissons.
- Power laws are “scale-free” or fractal
- This suggests that a common mechanism may be responsible for the behavior across all scales.
- I.e., there is a single explanation that explains both poorly and highly connected nodes.
- More generally, some think that power laws are “deep” and indicate some special type of organization or simplicity.
- Power laws—“the long tail”—have become a powerful metaphor or stylized fact.

How might Power Laws Form?

- Is there a model of network growth that exhibits a power law degree distribution?
- Yes. In 1999, Barabási and Albert (re)introduce a model for growth that produces power laws.
- Barabási and Albert, ”Emergence of scaling in random networks”, Science, 286:509-512, October 15, 1999.
- There is quite a long pre-history to this model. It turns out that their basic idea goes back to at least 1925, and their model is a special case of other models that had been previously published.
- There has also been much follow-up work and some significant critique of this model
- More on pre- and post-history later.

Rich-get-richer

- There is a class of growth models—not just for networks—based on the following idea.
- Nodes with more links are more likely to get more links.
- This idea goes by many different names:
  - Cumulative advantage (Simon)
  - Rich get richer
  - Preferential attachment (Barabási and Albert)
  - Matthew effect (Merton)
  - Yule process
- Matthew 25:29. “For unto every one that hath shall be given, and he shall have abundance: but from him that hath not shall be taken away even that which he hath.”

The "Barábasi-Albert" Model

- As noted above, there are lots of variations and precursors to this model.
- Here is the simplest version of the model:
  1. Nodes are added to the network one at a time. Each node makes m links to existing nodes.
  2. The nodes are randomly connected to the existing nodes, with a probability that is proportional to the number of links that node has.
- It turns out that this model has

$$P(k) \approx k^{-3}$$

  for large k.
- Variants on this model can produce other power laws.

Summary Thus Far

Basic Network Properties:

- Path lengths
- Degree distribution
- Cluster coefficient

Three Models:

- Erdős-Rényi random graphs
- Watts-Strogatz small-world model
- Preferential attachment models & scale-free degree distributions

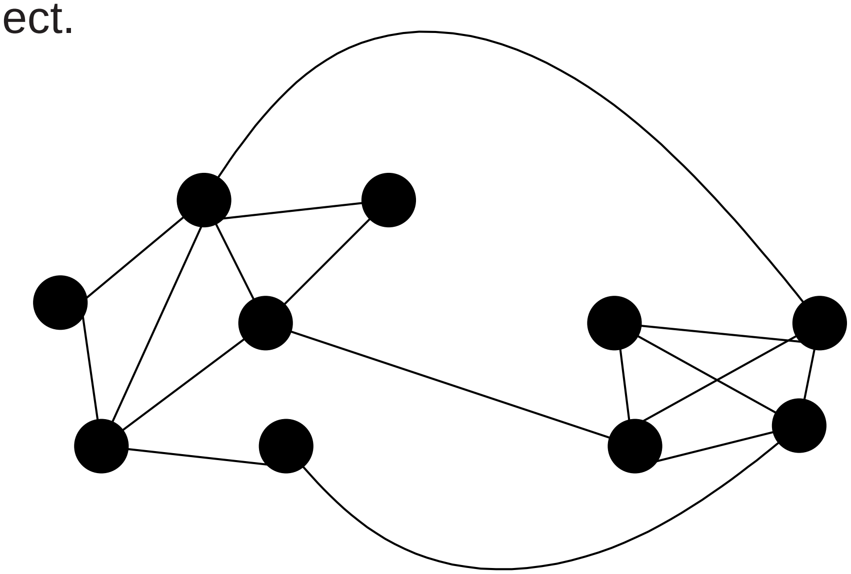
Note: “Scale-free network” is an improper term. It is only the degree distribution which is scale free, not the entire network.

Community Discovery and Higher Order Structure

- How can we say more about the structure of large networks?
- Are the nodes clustered in any way?
- What might these clusters mean?
- How can we discover these clusters?
- More generally, how can we think about higher order structures in networks? I.e., structures involving more than a single node.
- And what might these higher order structures mean physically or biologically or economically, etc.?
- All of these questions are areas of active research.

What Makes a Community?

- Suppose we suspect that a network is made of two communities. Can we test this?
- A group is a community if there are more within-community connections than one would expect.



- How can we quantify this?
- Note: Communities are also sometimes referred to as modules.

How can we Specify Communities?

- We are interested in discovering community structure.
- A community can be thought of as a *partition* of the network. Each node is assigned to one and only one community.
- Usually, we don't want to specify the number of communities in advance.
- That is, we want to discover the optimal number of communities *and* the optimal placement of individual nodes into communities.
- This is an extremely difficult computational task. Trying out all possible community specifications wildly unfeasible.

Community Discovery

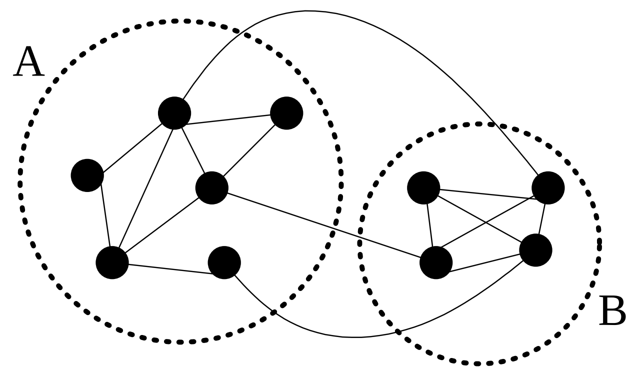
- This is an area of active research. There is not a standard algorithm, nor is there general agreement about which algorithm is the best.
- Nevertheless, there are some nice and commonly used methods for community discovery.
- These methods are not necessarily well understood(?) and may be unreliable.
- There are two components to any community discovery algorithm:
  1. A metric: some way of measuring how good a potential community partition is. This, is the thing that the algorithm tries to maximize.
  2. A search method: some way of generating candidate community partitions which are then evaluated according to the metric.

Modularity

- One commonly used metric for the quality of a community partition is the *modularity* Q.
- The larger the Q-value, the better the community partition.
- Let A be the adjacency matrix for the network.
- Let n_c be the number of communities in our partition.
- Define an $n_c \times n_c$ matrix E whose elements e_{ij} are the fraction of total links starting at a node in community i and ending in community j
- Let $a_i = \sum_j e_{ij}$ be the fraction of links connected to i.
- Then

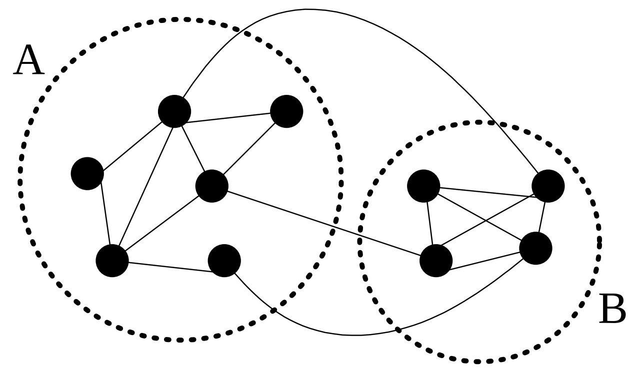
$$Q = \sum_i (e_{ii} - a_i^2) \ . \qquad (13)$$

Modularity: A Measure of Community-ness



- Suppose we think there are two communities, A and B.
- Divide the links into two types: between-community and within-community.
- For this network, there are 8 links within A, 6 within B, and 3 between A and B.
- There are 17 total links.
- So $\frac{8}{17}$ of the links are within community A.
- Is this a lot? How many would we expect?

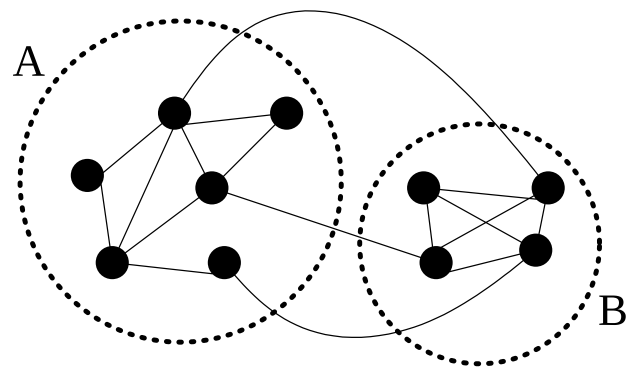
Modularity: Continued



- 8 links within A, 6 within B, and 3 between A and B, and 17 total links.
- $\frac{8}{17}$ of the links are within community A. Is this a lot?
- Of the 17 total links, 11 connect to A.
- If no community structure, then the communities edges link to are independent.
- So, if we draw a link at random, what is the chance it connects A to A?

$$\text{Prob of connection to A} \times \text{Prob of connecting to A} = \frac{11}{17} \times \frac{11}{17}$$

Modularity: Continued

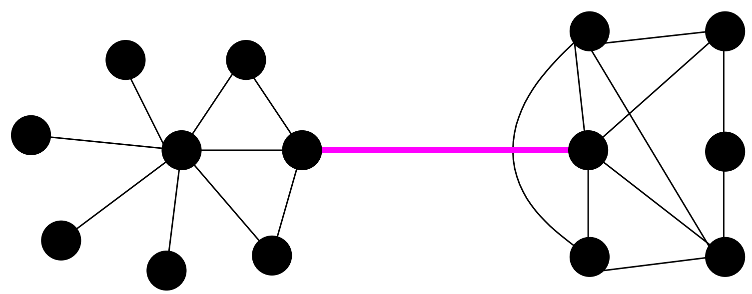


- **Modularity** is defined as the fraction of within-community links minus the number of within community links one would expect if the links were random.
- For community A: $\frac{8}{17} - \frac{11^2}{17^2}$.
- For community B: $\frac{6}{17} - \frac{9^2}{17^2}$.
- Adding these together, we get the modularity of the network. In this case, modularity $= 0.12$.
- **Modularity is a measure of the strength of a set of communities. The bigger the number, the stronger the community structure.**

Modularity: Conclusion

- There are a number of community discovery algorithms which are based on the modularity Q.
- These algorithms generate a series of candidate community partitions and evaluates the Q for each.
- The partition that has the largest Q is then chosen.
- I'm not sure the statistical properties of Q are well understood, or if this is the ideal metric for community-ness.
- These algorithms are conceptually simple, have been widely used, and have produced reasonable and interesting results.

Girvan-Newman Betweenness Algorithm



- The betweenness is a property of an edge.
- Betweenness measures how important an edge is in connecting other members of the network.
- To calculate betweenness, consider all possible pairs of edges.
- Find the shortest path connecting each pair.
- The betweenness of an edge is the number of shortest paths running along that edge
- See, e.g., Finding and evaluating community structure in networks, M. E. J. Newman and M. Girvan, Phys. Rev. E 69, 026113 (2004). for discussion of betweenness and modularity.

Girvan-Newman Betweenness Algorithm

- Central Idea: Edges with high betweenness separate communities.
  1. Calculate all betwennesses
  2. Remove the edge with the highest betweenness
  3. Repeat until all nodes are in their own community
- As one does this process, the network fractures into successively smaller and smaller, disconnected components
- Consider each disconnected component as a community
- Each successive splitting generates a new candidate community partition, one with one more community than before.
- Evaluate the modularity Q for each partition
- Choose the partition with the largest Q

Girvan-Newman Betweenness Algorithm

- Note: This is an example of a divisive algorithm.
- This algorithm gives good results for networks with known community structure: e.g., college American football conferences, Zachary's karate club.
- This is an $\mathcal{O}(m^2 n)$ algorithm. $n =$ number of nodes, $m =$ number of links.
- For sparse networks, $m \sim n$, so this is $\mathcal{O}(n^3)$.
- Not feasible for very large networks.
- This algorithm gives not just a single, optimal-Q community structure, but a full hierarchy, or dendogram.

Newman Fast Algorithm

- See M. E. J. Newman, Phys. Rev. E 69, 066133 (2004), and Aaron Clauset, M. E. J. Newman, and Cristopher Moore, Phys. Rev. E 70, 066111 (2004).
- This is an agglomerative algorithm.
- Initially, every node is its own community.
- Merge the two communities that lead to the largest increase in Q.
- Repeat, until all have been merged together.
- Like the previous algorithm, this yields a dendogram.
- The partition with the highest Q is chosen as optimal.
- This algorithm is $\mathcal{O}(md\log n)$, where d is the dendogram depth.
- For sparse, hierarchical networks, the order is $\mathcal{O}(n\log^2 n)$.
- This is much faster than the Girvan-Newman algorithm.

General Community Discovery/Data Mining Thoughts

- Discovering communities when you have good reason to believe that communities are present is a hard problem.
- But what if you're not certain communities are present?
- Most algorithms will still find communities.
- There is an almost irresistible temptation to give meaning to these communities.
- Is there some notion of statistical significance for communities?
- There are at least two issues: the significance of the overall community structure, and the significance of the placement of individual nodes.

Cautionary Notes on Modularity Maximization

- The performance of modularity maximization in practical contexts. B. H. Good, Y.-A. de Montjoye and A. Clauset. Physical Review E 81, 046106 (2010) `http://arxiv.org/abs/0910.0165`.
- The optimal partition may not coincide with the intuitive best parition.
- There are exponentially many, structually distinct partitions that have modularities very close to the maximum.
- Thus, heuristic methods can quickly find good (high Q) communities, but different heuristics find very different community structures.
- S. Fortunato and M. Barthélemy, Proc. Natl. Acad. Sci. USA 104, 36 (2007).
- “We find that modularity optimization may fail to identify modules smaller than a scale which depends on the total size of the network and on the degree of interconnectedness of the modules, even in cases where modules are unambiguously defined.”

Now What?

- Modularity-maximization as a technique for community detection may be doomed.
- This calls into question the notion of modularity.
- Much work needs to be done on statistical inference and data mining for network data.
- It is relatively easy to fool oneself into seeing thing that aren't there when analyzing networks. (This is the case with almost anything, not just networks.)
- For networks, can we (should we) be more careful and scientific, and not just descriptive and empirical?