CSE 255 – Lecture 7

Data Mining and Predictive Analytics

Rich-club phenomena

Rich clubs

Rich nodes, and rich-clubs

- We’ve seen how real-world networks are characterized by “rich-get-richer” phenomena, and that networks tend to follow power-law distributions, i.e., a small number of nodes have high degree, while many nodes have low degree

- But what about the connectivity structure **between** high-degree nodes?
  - i.e., are nodes of high degree likely to connect to other nodes of high degree?

Rich clubs

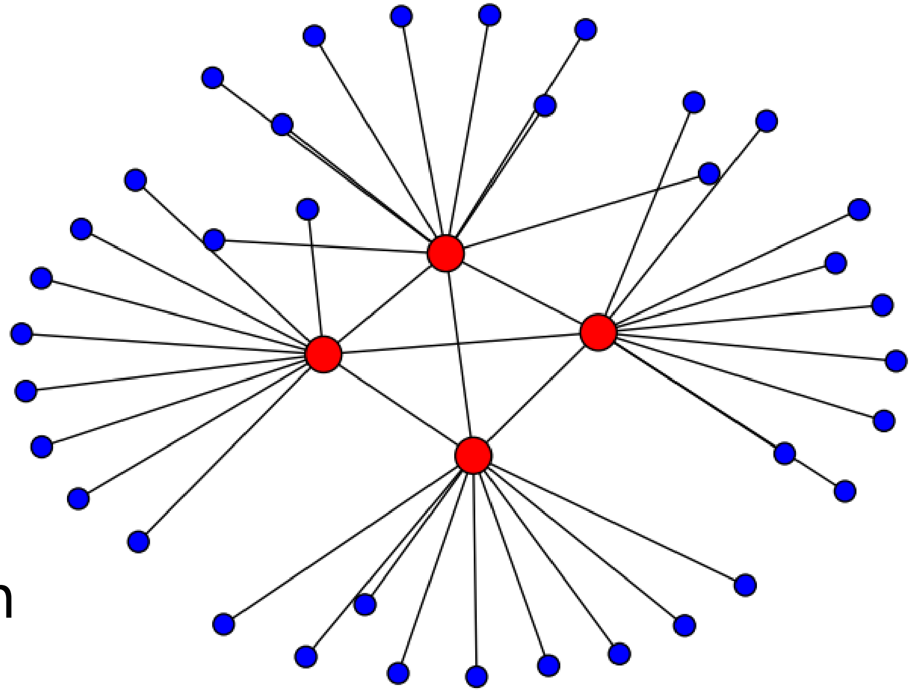
Rich nodes, and rich-clubs

Why is this characterization important?

- Is a power-grid robust to cascading failure? Maybe we would want "hubs" in such a network **not** to be connected to other hubs
- Do people who write many papers tend to co-author with others who write many papers, or do they form distinct authorship "hierarchies"?
- High-degree nodes in Protein-Protein-Interaction networks may perform regulatory functions. Whether high-degree nodes are interconnected may determine whether they regulate different functional modules

Rich clubs

e.g. (from Wikipedia)



red nodes form a "rich club"

Rich clubs

Let’s try to characterize this property

What fraction of edges appear between high-degree nodes?

#edges between nodes of degree > k

$$\phi(k) = \frac{2E_{>k}}{N_{>k}(N_{>k}-1)}$$

degree

#nodes of degree > k

Whether this function **increases** or **decreases** with *k* will somehow characterize whether edges tend to be “more prevalent” between high-degree nodes or not

Rich clubs

But! In most networks this won’t look particularly interesting – high-degree nodes are more likely to be connected to each other just by chance

So really, we need to compare the above distribution to that of a random model

Rich clubs

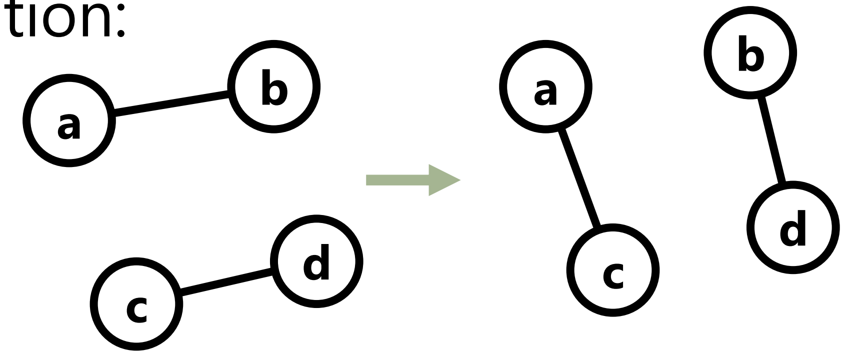
We need to compare the observed graph with a “random” graph that has the same degree distribution

- We’ve seen already how designing random graph models that preserve the correct statistics is **hard**
- But it turns out that generating a random sample with a **specific** degree distribution is **easy!**

Rich clubs

How to randomly sample graphs with a specific degree distribution:

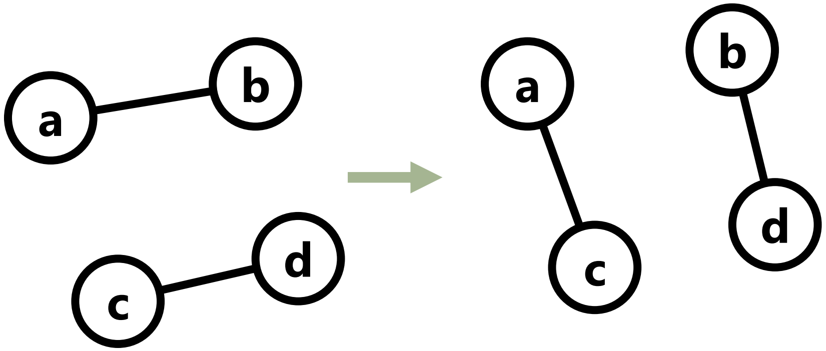
- Start with any graph with the correct degree distribution (e.g. the input graph we're trying to assess)
- Randomly select pairs of (non-adjacent) edges (a→b) and (c→d)
- Apply the following operation:
- Repeat many times



Rich clubs

How to randomly sample graphs with a specific degree distribution:

- This operation **preserves the degree distribution**

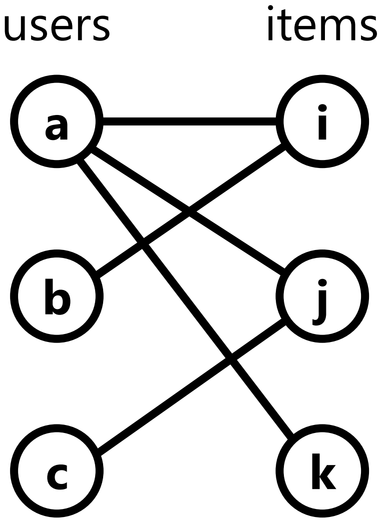


- If the operation is repeated enough times, then **any** graph with the desired degree distribution is equally likely to be observed (i.e., we're sampling uniformly; see Milo et al.)

Rich clubs

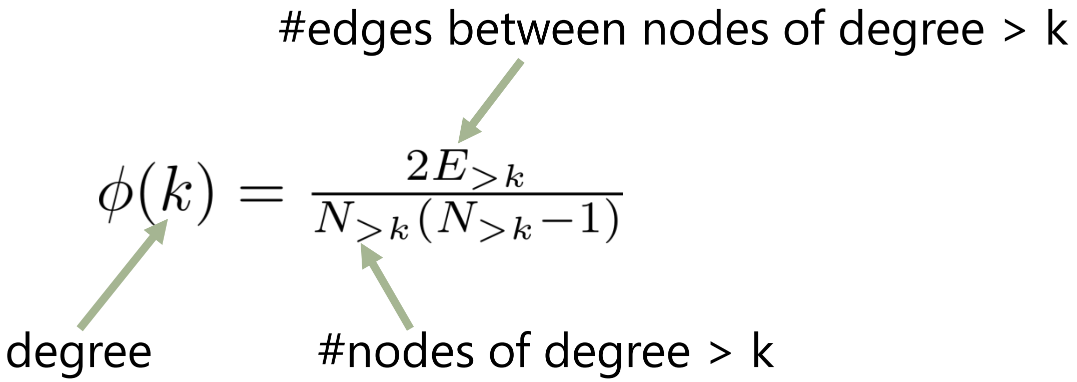
(aside: this is also useful when training recommender systems!)

e.g. when training a system to predict likely purchases for each user we wouldn’t want the system to be biased toward “power users”. To get around this, we may generate a “negative set” for training purposes that has the same degree distribution as the positive set. We can do this by applying the previous procedure to the bipartite purchasing graph



Rich clubs

We can now compare the statistics of the randomly sampled graph to the statistics of the observed graph



#edges between nodes of degree > k

$$\phi(k) = \frac{2E_{>k}}{N_{>k}(N_{>k}-1)}$$

degree

#nodes of degree > k

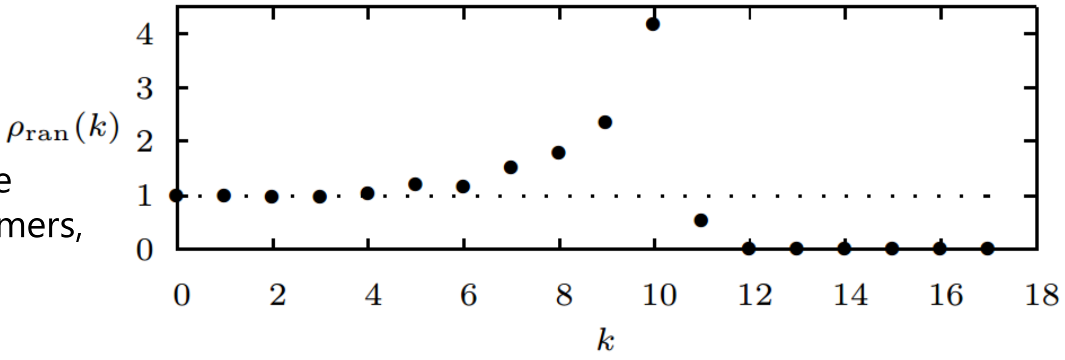
i.e., we compute

$$\phi_{\text{obs}}(k)/\phi_{\text{rand}}(k)$$

Some results

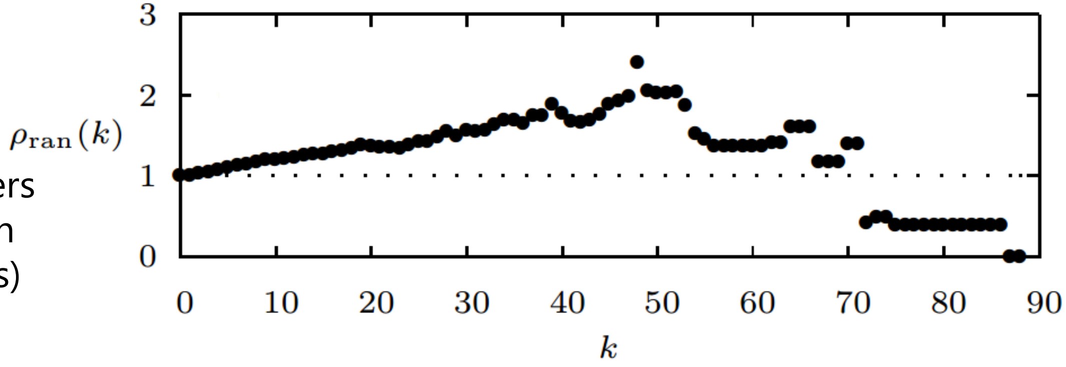
Power grid
(Watts & Strogatz, 1998)

(high-degree plants/transformers tend to be connected to other high-degree plants/transformers, except for the most connected plants)



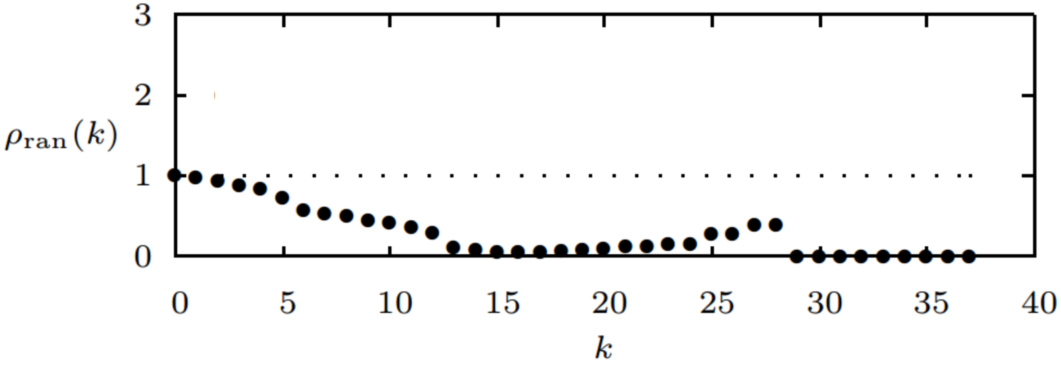
Scientific collaborations
(Colizza et al. 2006)

(highly-connected authors tend to write papers with other highly connected authors, though perhaps not for the **most** connected authors)



PPI network
(Jeong et al. 2001)

(highly-interactive proteins (e.g. regulators) tend **not** to interact with each other)

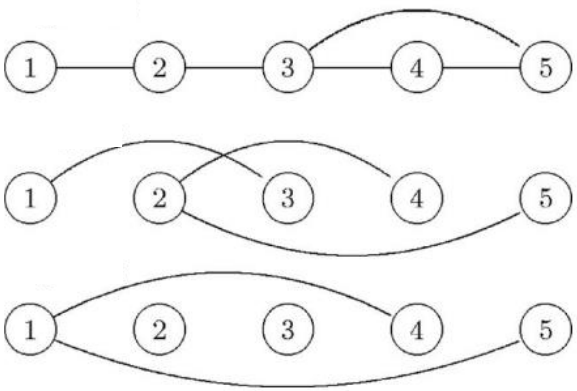


Rich clubs

What about network **hierarchies?**

- So far we've looked at the likelihood that high-degree nodes are connected to other high-degree nodes
- But what about the probability that high-degree nodes have **paths** to other high degree nodes
- e.g. do high-degree nodes tend to be two-hops away from other high-degree nodes?

To test this we construct "higher-order" networks (by taking powers of the adjacency matrix)



original network

2nd-order network

3rd-order network

Rich clubs

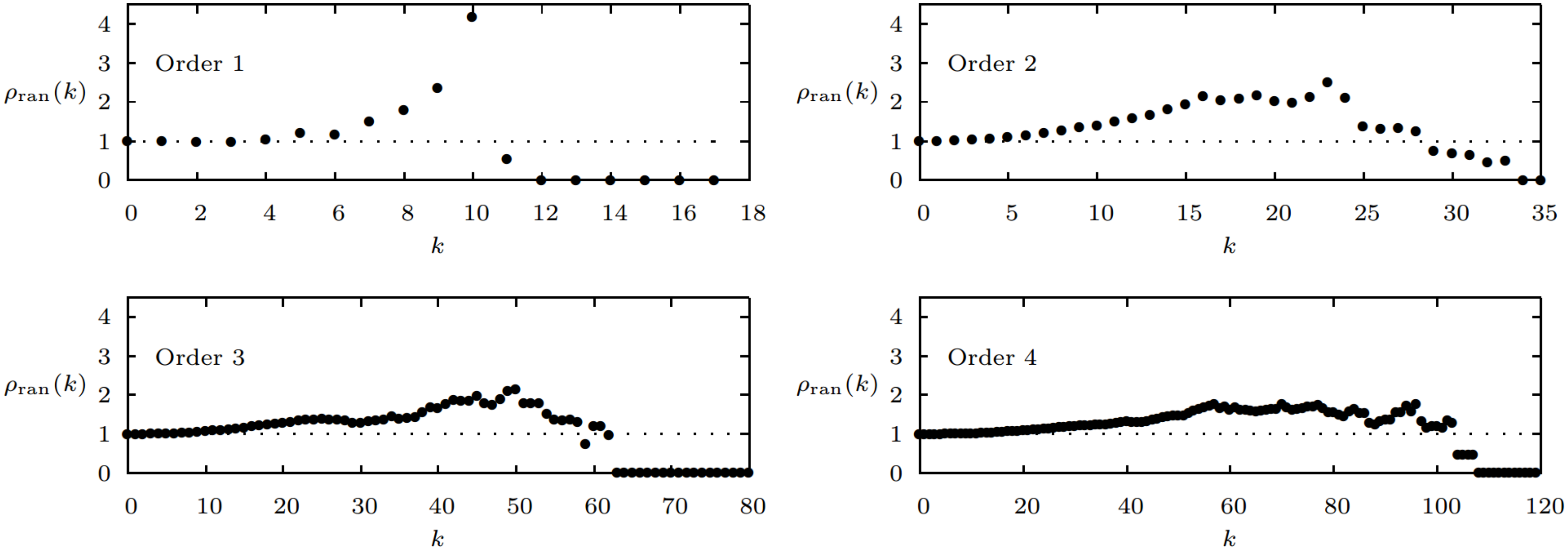
Rich-clubs in higher-order networks

Why (part 2)?

- Perhaps in citation networks there are authors who act as “bridges” between the highest degree nodes, e.g. people who operate at the intersection between two areas
- Perhaps we can better characterize robustness to failure in a power grid, e.g. the distances between critical points of failure
- Or in a PPI networks, while we saw that “regulators” don’t tend to interact, perhaps some proteins act as “intermediaries” between regulators

Rich clubs

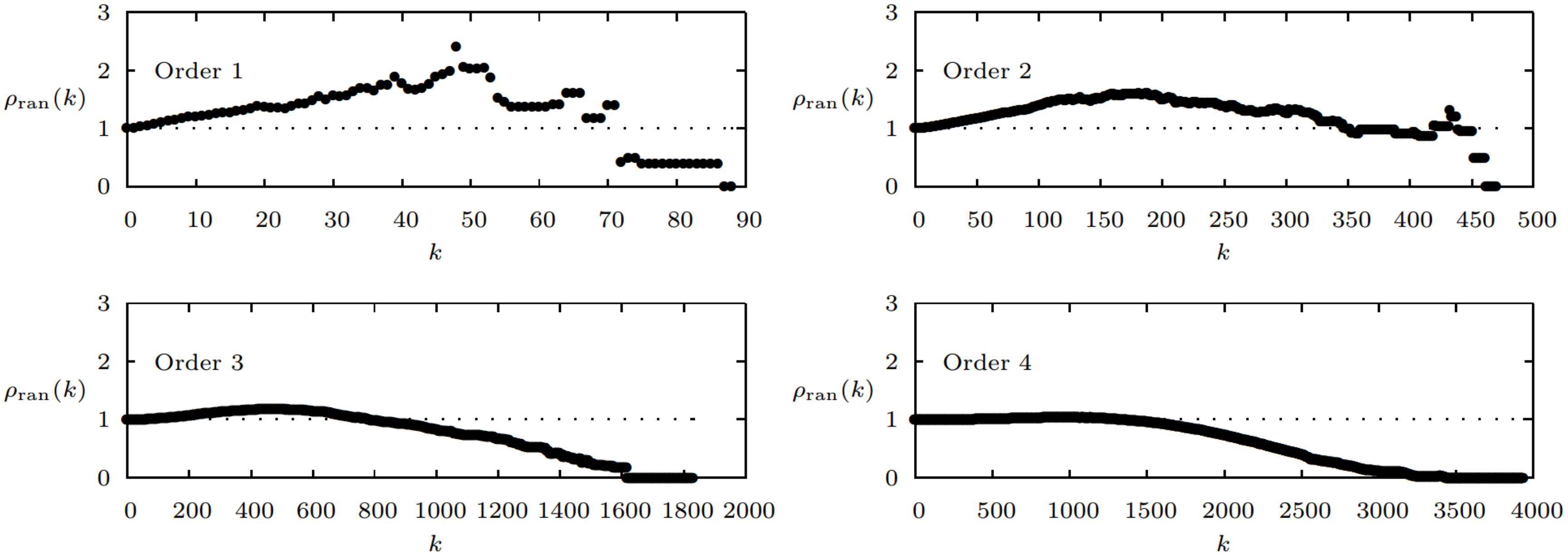
Rich-clubs in higher-order networks



Power grids: roughly the same sort of behavior exists across all orders of the network

Rich clubs

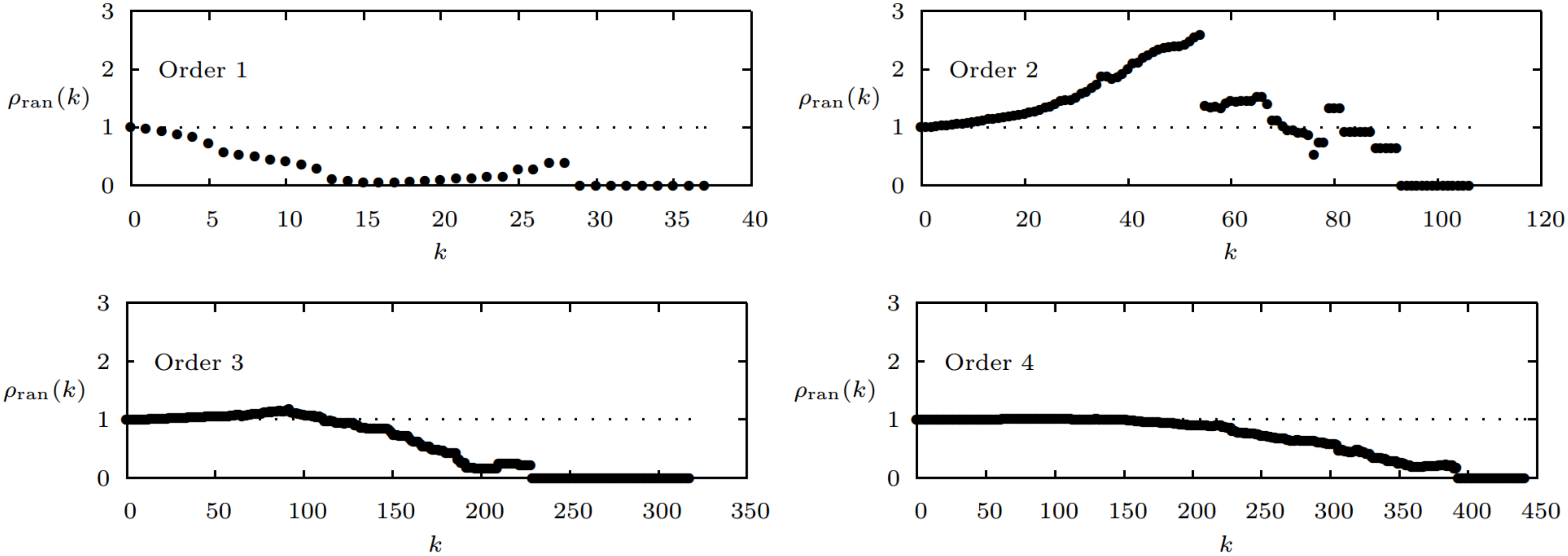
Rich-clubs in higher-order networks



Citation networks: rich-club behavior gradually disappears in higher-order networks

Rich clubs

Rich-clubs in higher-order networks



PPI networks: the phenomenon behaves **non-monotonically:** there are few interacting hubs, but hubs seem to have intermediaries

Rich clubs

Morals of the story:

- The rich-club phenomenon can reveal important structural properties of a network, such as the roles of high-degree nodes and the networks resilience against cascading failures
- To understand whether high-degree nodes connect to each other, we need to compare against the right random model
- Rich-clubs in **higher-order** networks can further reveal the extent to which high-degree nodes may be connected via intermediaries

Questions?

The **final homework** is available on the course webpage:
http://cseweb.ucsd.edu/~jmcauley/cse255/homework9.pdf
If you want it back, swing by during office hours

Course evaluations are also online!