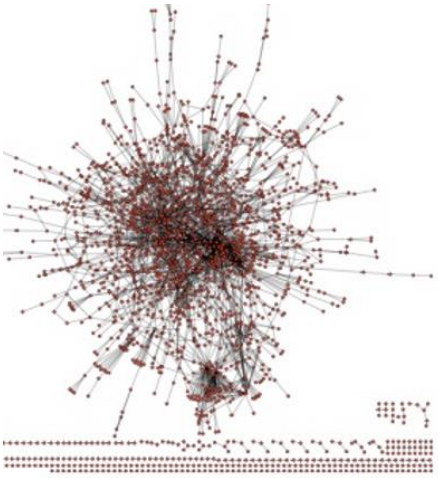
Case Study Title: Understanding and analyzing a protein-protein interaction network

Background:

Network science is an academic field which studies complex networks such as telecommunication networks, computer networks, biological networks, cognitive and semantic networks, and social networks, considering distinct elements or actors represented by nodes (or vertices) and the connections between the elements or actors as links (or edges). The field draws on theories and methods including graph theory from mathematics, statistical mechanics from physics, data mining and information visualization from computer science, inferential modeling from statistics, and social structure from sociology. The United States National Research Council defines network science as "the study of network representations of physical, biological, and social phenomena leading to predictive models of these phenomena."

As biological function emerges through interactions between a cell's molecular constituents, understanding cellular mechanisms requires a reasonably complete catalogue of all physical interactions between proteins. Despite major efforts in high-throughput mapping, the number of missing human protein-protein interactions (PPIs) exceeds the experimentally documented interactions. Consequently, computational tools are increasingly used to predict undetected, yet potentially biologically relevant interactions. One way to do so is to model protein-protein interactions as a network, followed by network analytics such as link prediction.

We will use the following example network for illustration.

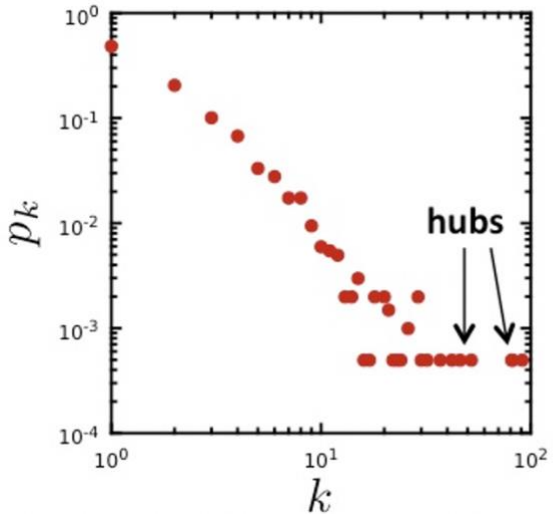


Undirected network
N=2,018 proteins as nodes
L=2,930 binding interactions as links.
Average degree <k>=2.90.

Not connected: 185 components
the largest (giant component) 1,647 nodes

What analytics can we run?

Let's start with degree distribution, including understanding what the hubs in the network are.

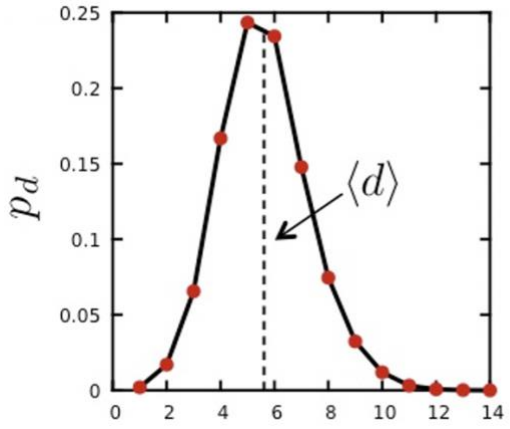


p_k is the probability that a node has degree **k**.

N_k = # nodes with degree k

$p_k = N_k / N$

Next, we should analyze the mean diameter of the connected components in this network:

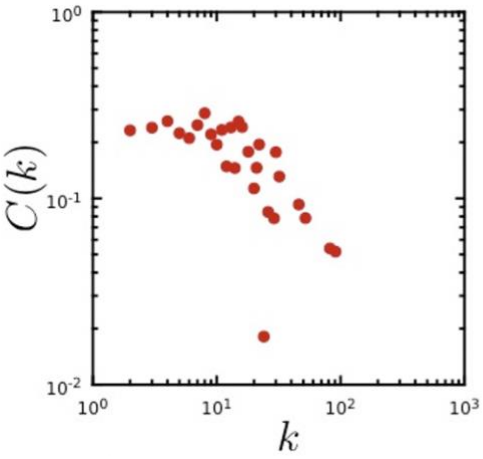


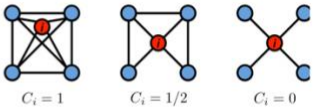
d_{max}=14

<d>=5.61

Here, the x axis is the diameter.

Finally, we can run a clustering coefficient analysis:





$$C_i = \frac{2e_i}{k_i(k_i - 1)}$$

<C>=0.12

Questions

i) What is an efficient algorithm to get the connected components in the network?

ii) What is the connection between the average in-degree, average-out-degree and number of edges and nodes in the network? Try to provide a formula if possible.

iii) Why do we refer to the nodes in the degree distribution figure in the 'What analytics can we run?' section as hubs? What is your intuitive understanding of what a hub is? Are there situations where high-degree nodes may not be hubs?

iv) The mean diameter in the example network is 5.61 and the maximum is 14. Try to interpret these findings.

v) We showed three kinds of analytics earlier. Are there other kinds of analytics you can suggest? For example, would an assortativity analysis help shed more light on the structure of the network?

vi) Suppose the nodes in the network had 'metadata' associated with them. Try to read about proteins to understand what this metadata could be. Suppose some of the nodes had labels from a

closed set but for others, the label was missing. How might you do machine learning given only the network structure and metadata to predict those missing labels?

vii) We mentioned link prediction as an important example earlier in the introduction. What heuristics or machine learning methods can you think of to do link prediction given only the PPI network? What if metadata was also present? Could you draw a simple architecture of how you could combine structural and metadata information (including text) to do link prediction? Be specific.

vii) Based on the analytics presented earlier, comment on how the protein-protein interaction network is similar to, or different from, social networks.