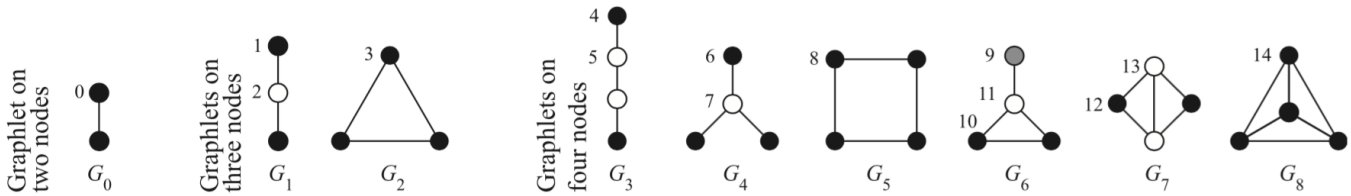
Graphlet degrees, random-walk sampling, network comparison

I. Graphlet degree distributions

You are given five networks in Pajek format.

- Java class dependency network (java.net)
- iMDB actors collaboration network (collaboration_imdb.net)
- Gnutella peer-to-peer sharing network (gnutella.net)
- Sample of Facebook social network (facebook.net)
- *nec* overlay map of the Internet (nec.net)

1. **(discuss)** Consider the orca algorithm by Hočevar and Demšar for computing graphlet orbit degrees k^i for graphlets with up to 4 nodes, $i \in \{0, \ldots, 14\}$.



2. **(homework)** Compute the graphlet orbit degrees k^i and plot the graphlet degree distributions p_k^i. How are the graphlet orbit degrees k^i distributed in different networks?

II. Estimation by random-walk sampling

You are given five networks in Pajek format.

- Java class dependency network (java.net)
- *nec* overlay map of the Internet (nec.net)
- Sample of Facebook social network (facebook.net)
- Enron e-mail communication network (enron.net)
- A small part of Google web graph (www_google.net)

1. **(code)** Represent the networks with simple undirected graphs and reduce them to their largest connected component.

2. **(code)** Implement a random-walk sampling and apply it to the networks until you sample 15% of the nodes (with repetitions). Let s be the number of sampled nodes and $k_1, \ldots, k_s$ their degree sequence. Estimate the average degree of the network $\langle k \rangle$ using a biased average

$$\frac{\sum_i k_i}{s}$$

and also the corrected estimate

$$\frac{s}{\sum_i k_i^{-1}}.$$

3. **(discuss)** Compare both estimates to the true average degree $\langle k \rangle$.

III. Sampling Facebook social network

You are given two large samples of Facebook social network with around ten million of nodes and links. Due to their size, the networks are available only in compressed edge list format.

- 1st sample of Facebook network (facebook_1.adj.zip)
- 2nd sample of Facebook network (facebook_2.adj.zip)

1. **(discuss)** The samples were generated by a uniform random node selection technique called *rejection sampling* and by the breadth-first search approach called *snowball sampling*.

2. **(homework)** Try to figure out which network sample is which. Since these are still very tiny samples of Facebook social network, the answer might not be immediately obvious from their structure.

IV. Networks and models comparison

You are given three social networks and three food web graphs in Pajek format.

- Zachary karate club network (karate_club.net)
- Davis southern women network (southern_women.net)
- Lusseau bottlenose dolphins network (dolphins.net)

- Little Rock Lake food web (foodweb_littlerock.net)
- Cypress Wetlands food web (dry) (foodweb_baydry.net)
- Cypress Wetlands food web (wet) (foodweb_baywet.net)

1. **(discuss)** Consider different approaches for comparing networks. These include comparing networks by different metrics or statistics, graph edit distance, graphlet degree distribution agreement, portrait divergence, D-measure etc. You can implement the approaches by yourself, browse your network library for existing implementations or use the code provided below.

    - Simplified D-measure: simplified_dmeasure.py

- Network portrait divergence: portrait_divergence.py
- Graphlet distribution agreement: graphlet_aggrement.py

Note that the last script requires a working installation of the orca algorithm for counting graphlet orbits.

2. **(code)** Compare the networks between each other and plot their dissimilarity or distances with a heat map. How similar are networks of different type? For instance, are social networks more similar to each other than to food webs? Does the answer depend on the selected measure of dissimilarity or distance?

3. **(code)** Compare the networks also to small synthetic graphs such as Erdös-Rényi random graphs, Barabási-Albert scale-free graphs and Watts-Strogatz small-world graphs with $n = 500$ nodes and $m = 1500$ edges. How similar are real networks to synthetic graphs? How do synthetic graphs compare between each other?