

Random-walk sampling, Facebook social network

You are given four large networks in Pajek format (edge list and LNA formats are also available).

- [Internet overlay map](#) (75,885 nodes)
- [Facebook social network](#) (63,392 nodes)
- [Enron e-mail network](#) (84,384 nodes)
- [Google web graph](#) (855,802 nodes)

Later on you will be studying also two very large samples of Facebook social network with millions of nodes and links. Due to the size of these networks, they are available only in edge list format.

- [1st Facebook network](#) (8,217,272 nodes)
- [2nd Facebook network](#) (7,698,354 nodes)

I. Estimation by random-walk sampling

1. Try to implement a simple random-walk sampling and apply it to the largest connected components of four large networks above until you sample 15% of the nodes (with repetitions). You should use the same routine as in Homework #2.
2. Let s be the number of sampled nodes and k_1, \dots, k_s their degree sequence. Estimate the average degree in the complete network $\langle k \rangle$ using a biased average

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

and also the corrected estimate presented in lectures

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

Compare both estimates to the true average degree $\langle k \rangle$ and discuss the results.

II. Sampling Facebook social network

The two samples of Facebook social network above were generated by random node selection technique called *rejection sampling* and by breadth-first search approach called *snowball sampling*. 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 obvious from their structure.