Modeling the Variance of Network Populations with Mixed Kronecker Product Graph Models

Sebastian Moreno, Jennifer Neville, Sergey Kirshner and S.V.N. Vishwanathan

Natural variability of real networks

- We investigate the natural variability of network datasets using public Purdue Facebook wall communications.
- From this, we constructed 25 sample networks by first selecting a time point at random and then collecting 1024 nodes sequentially from the communication activity, along with all edges among the sampled set over the next 60 days.
- For each of the 25 sampled networks, we plot the degree, cluster coefficient, and hop plot distributions.

Background: KPGMs

- KPGM is a fractal model, which uses independent Bernoulli rvs for each edge, which reduces variance in the generated graphs.
- To address this, we developed a tied-KPGM (KPGM), which ties the Bernoulli trials for each Kronecker multiplication, resulting in increased clustering and variance.
- This can be viewed as a hierarchy of Bernoulli trials, each applied to smaller blocks of the matrix, where the final edge probability is determined from the associated Bernoulli trials at each level.

Mixed KPGM:

- Due to the Bernoulli dependencies, KPGM exhibits significantly more variance than we see in real Facebook networks.
- So we outline a mixed-KPGM (mKPGM), which uses independent Bernoulli in the top (global) levels of the hierarchy and then ties the Bernoulli at the lower (local) levels of the hierarchy.
- The amount of variance and clustering depends on the level at which the Bernoulli is tied.

Tied KPGM:

- KPGM uses independent Bernoulli rvs for each edge, which reduces variance in the generated graphs.
- To address this, we develop a tied-KPGM (tKPGM), which ties the Bernoulli trials for each Kronecker multiplication, resulting in increased clustering and variance.
- This can be viewed as a hierarchy of Bernoulli trials, each applied to smaller blocks of the matrix, where the final edge probability is determined from the associated Bernoulli trials at each level.

Experiments

Synthetic experiments to illustrate properties of mKPGM:

- We illustrate the increase in variance across the mKPGM spectrum using 300 generated graphs for \( k=0.2, 0.2, 0.2, 0.2 \) and \( k=10 \), as the level of tying (\( k \)) is varied (solid lines represent the mean, while dash lines are one standard deviation from the mean).
- Observations: Upper left: Variance in the number of edges decreases proportionally with \( L \); upper right: median degree of a node increases proportionally with \( L \); lower left: small degree increases proportionally with \( L \); lower right: large degree increases proportionally with \( L \); mKPGM changes the generated network structure, resulting in graphs with small clustering coefficient and large diameter as \( L \) increases.

Facebook experiments:

- Learn a KPGM model \( \hat{\Theta}_{\text{KPGM}} \) from the network closest to the median of the degree distribution.
- Estimate mKPGM (\( \hat{\Theta}_{\text{mKPGM}} \)) and KPGM (\( \hat{\Theta}_{\text{KPGM}} \)) models using exhaustive search based on the expected number of edges \( \hat{E}(E) = 1975.2 \).
- Generate 200 networks from each model; evaluate the degree, hop plot, and clustering coefficient distributions.
- Observations: KPGM and mKPGM are not able to capture the variance of the population. In contrast, the mKPGM model not only captures the variance of the population, it matches the mean characteristics more accurately as well.

Conclusions

- KPGM lack of variance can be explained due to its independent sampling of the Bernoulli rvs that are used to generate edges.
- We propose a mixed-KPGM, which introduces dependence in the edge generation process and thus increases variance (and clustering).
- Our current work is focused on developing a method to estimate the mKPGM parameters from observed networks.