Clustering the vertices of a random dot product graph works-ish

Lyzinski *et al*., 2015

an ocarf by

Max Collard & Jerry Wang

Opportunity

- Biology needs generative models.
- SBMs are nice such models for random networks.
- Clustering vertices seems like a reasonable thing in this model …
- … and yet we don’t know if it actually works.

Preliminaries

- Model details
  - Random dot product graph (RDPG)
  - Stochastic blockmodel (SBM)
  - Degree-corrected SBM (dcSBM)
- Procedure details
  - Adjacency spectral embedding (ASE)
  - Mean square error clustering, *i.e.* what *k*-means solves

Challenge

The Frobenius norm sucks.

Advances

New norm. Cleverness.

Result

k-means on the adjacency spectral estimates “works”.

Result

- SBM
  - MSE clustering on ASE estimates of latent variables, **probability of making an error goes to 0** as *n* goes to infinity
- dcSBM
  - **Same**, but using projection of latent variables onto the (*d*-1)-sphere

Result

- RDPG
  - **Asymptotically**, clustering on some transformation of the ASE latent variables **works just as well** as clustering on the latent variables themselves

Future: *The "-ish"*

On the negative side:

- Assumptions everywhere
  - We know *k*, number of clusters
  - We know *d*, dimensionality of latent variables
  - ...
- Bounds on probabilities are of near-zero utility without knowledge of hidden variables

Future

- “Perfection” only comes in the large *n* limit. How large is large?
- What if real graphs aren’t RDPGs?

On the positive side:

- Techniques here generalize to consistency proofs for other clustering techniques (*e.g. k*-NN) and other generative models