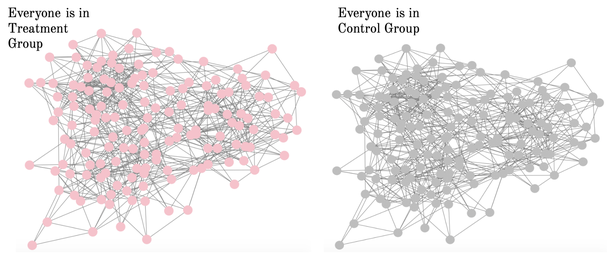
Community Randomized Graph Cluster Randomization

Heather Mathews[1], Alexander Volfovsky[1]

[1]Department of Statistical Science, Duke University

Objective

Our goal is to show how utilizing latent community structure can lead to better estimation of global average treatment effect (GATE) when performing causal inference in a network setting. The GATE is the average difference in response between two versions of a network: one where everyone is treated vs. one where everyone is not treated.



- Usually there is an assumption that the treatment of one individual does not influence the outcome of another (highly unlikely in a network)
- Some methods have tried to address this such as randomized graph cluster randomization (RGCR) (Ugander et al, 2020; Eckles et al, 2017)
- However, it is also common for a network to have latent community structure in which different communities are influenced by treatment in different ways

We propose methods that leverage the true communities to create more meaningful clusters for RGCR in order to allow for **better balance** between the treated and not treated groups and to **gain access to community level treatment effects**. We show how our methods **reduce bias** and how these methods are still incredibly useful when community structure is estimated.

Randomized Graph Cluster Randomization

- The graph is partitioned into clusters, C, using some type of clustering method (epsilon-net, k-means, etc)
- Let $C(i)$ indicate the cluster that node i is assigned to
- Each CLUSTER is then treated with probability p
- Proper cluster randomization can lead to exponentially lower estimator variance when experimentally measuring average treatment effects under interference and reduction in bias

Network Exposure

- Node i is treated network exposed if its response is equal to what it would have been if everyone was treated.
- Some examples of various 'network exposure':
  - i and all of i's neighbors are treated (Full Network Exposure)
  - i and a fraction of i's neighbors are treated
  - i and a certain number of i's neighbors are treated

Our Outcome Model

$$Y(Z) = \alpha 1_N + (\beta I_{N\times N} + [U\Gamma U^T \circ A])Z + N(0,1) \tag{1}$$

- Y: Response of interest
- α: Baseline
- β: Overall direct effect of treatment
- Z: Binary treatment assignment vector
- U: Latent community matrix $n \times K$, K is true number of communities.
- Γ: $K \times K$ matrix describing how strongly a node is influenced by neighbors' treatment
- A: $A_{i,j}|U_i, U_j \sim Bern(a)$ if $U_i = U_j$ or $A_{i,j}|U_i, U_j \sim Bern(b)$ if $U_i \neq U_j$

The GATE:

$$\tau = \frac{1}{N}\sum_i Y_i(Z=1) - Y_i(Z=0)$$

Theorem

Consider expressing the outcome model as (Eckles et al. 2017):

$$E[Y_i(Z)] = a_i + \sum_{j\in V} B_{ij}Z_j \tag{2}$$

where a is an N dimensional baseline vector and B is a $N \times N$ matrix with nonnegative entries. Under this model, the true GATE is:

$$\tau(1,0) = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} B_{ij} \tag{3}$$

Under independent assignment, $\tau^{ind}_{ITR} = \frac{1}{N}\sum_{i=1}^{N} B_{ii}$ and standard graph cluster randomization, $\tau^{gcr}_{ITR} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} B_{ij}1[C(i) = C(j)]$

But if we condition on communities...

Now, if A has true community structure, then we can consider clustering by community,

$$\tau^{gcr}_{ITR}|U = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} B_{ij}1[U(i) = U(j)] \tag{4}$$

If our method for picking clusters only allows nodes within the same community to be assigned to the same cluster, then

$$\tau^{gcr}_{ITR}|U \geq \tau^{gcr}_{ITR} \geq \tau^{ind}_{ITR}$$

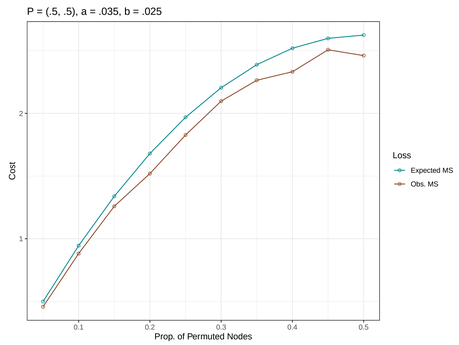
and bias is reduced when communities are known (since block diagonal structure is guaranteed whereas with regular clustering methods, it is not).

Cost of Estimating Community Labels

- While knowing the true community structure is ideal, this is typically not the case
- The community labels must be estimated. Recall, treatment assignment relies on the community labels and part of the GATE depends on the network/communities
- How much bias is added when community labels are not correctly estimated?

Notice that from the model, loss is associated with whether $U_i = U_j \rightarrow \hat{U}_i = \hat{U}_j$ rather than whether i is actually correctly classified. This means we really care about whether a node has been $mispaired$. The expected loss due to mispairing can be calculated:

$$E[Y(Z=1)_i - Y_i|Z_i = 1] = E[\gamma U_i \sum_j A_{ij} \times 1[U_i = U_j] - \gamma U_i \sum_j A_{ij} \times 1[U_i = U_j] \times Z_j)]$$



- The figure above shows the expected vs observed cost of mispairing a node when controlling how many labels in a network are permuted.
- For example, if 30% of nodes are permuted, 300 nodes are randomly selected and their true labels are switched
- This vector of permuted labels acts as an estimated label thus showing what can happen in the worse case scenario when nodes are misclassified

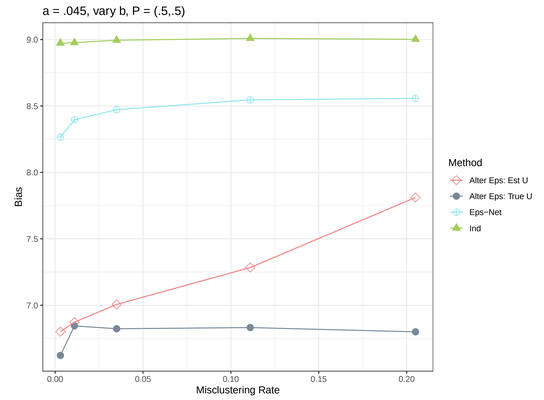
Methods for Graph Clustering

Recall, we are showing how to best allocate treatment assignment to better estimate the GATE. Below, we compare current methods with our methods that incorporate community structure. Current methods:

- **Ind**: Individual randomized assignment
- **Eps-Net**: Creates clusters based off of the epsilon net clustering algorithm but ignores community structure

Our methods:

- **Eps-Net that knows U but does not hold properties (Alter Eps):** Perform an eps-net on each subgraph (by community label)
- **Eps-Net that knows estimated U but does not hold properties (Alter Eps Est U):** Perform an eps-net on each subgraph (by ESTIMATED community label)
  - Community labels are typically not known and must be estimated using a community detection algorithm



$N = 1000, K = 2$ with $p_{comm} = [.5, .5]$

- Above, the network between community probability is varied to increase the difficulty of the community detection problem (and thus increase the misclustering rate)
- The altered eps-net that knows the true communities performs the best
- As the community detection problem becomes harder, the altered eps-net with estimated U begins to increase bias, however it still performs better than the other methods

Conclusions

- Knowledge of latent communities can lead to better treatment design and thus better estimation of average treatment effect!
- Easily can estimate the treatment effect for one community versus others
- Allows for better balance between treated and control groups
- Can be useful when we have limited resources! If we want a representative sample of our population, understanding and utilizing community structure is helpful

References

[1] D. Eckles, B. Karrer, and J. Ugander. Design and analysis of experiments in networks: Reducing bias from interference. *Journal of Causal Inference*, 5(1), 2016.

[2] J. Ugander and H. Yin. Randomized graph cluster randomization. *arXiv preprint arXiv:2009.02297*, 2020.

Contact Info

heather.mathews@duke.edu

Duke UNIVERSITY