Community Randomized Graph Cluster Randomization

Heather Mathews¹, Alexander Volfovsky²

¹Department of Statistical Science, Duke University

Objective

Our goal is to show how utilizing latent community structure can lead to better estimation of
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
goal of another (highly unlikely in a network)
* Some methods have tried to address this such as randomized graph cluster randomization
(RGCR) (Ligandet 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
different communities are influenced by treatment in different ways

Randomized Graph Cluster Randomization

* The graph is partitioned into clusters, C, using some type of clustering method (epsillon-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’
  * v and all of v’s neighbors are treated (Full Network Exposure)
  * v and a fraction of v’s neighbors are treated
  * v and a certain number of v’s neighbors are treated

Our Outcome Model

\[ Y(Z) = \alpha_N + \beta U_{i,x} + \epsilon Z X N(0, 1) \]  (1)

* Y: Response of interest
  * \( \alpha \): Baseline
  * \( \beta \): Overall direct effect of treatment
* Z: Binary treatment assignment vector
* \( U_i \): Latent community matrix \( n \times K \), \( K \) is true number of communities.
* \( \epsilon \): Error

The GATE:

\[ \tau = \sum_{i=1}^{N} \frac{1}{N} \left( Y(Z = 1) - Y(Z = 0) \right) \]

Risk depends on number of neighbors treated

Theorem

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

\[ E[Y(Z)] = \alpha_i + \sum_{j \in C(i)} B_{ij} Y_j \]  (2)

where \( \alpha_i \) is an \( N \times D \) dimensional baseline vector and \( B \) is a \( N \times N \) matrix with nonnegative entries.

Under this model, the true GATE is:

\[ \tau = \sum_{i=1}^{N} \sum_{j \in C(i)} B_{ij} \]  (3)

Under independent assignment, \( \tau = \sum_{i=1}^{N} \sum_{j \in C(i)} B_{ij} \) and standard graph cluster randomization,

\[ \tau_{GP} = \sum_{i=1}^{N} \sum_{j \in C(i)} B_{ij} \]  (4)

But if we condition on communities...

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

\[ r_{GP}^A(U) = \sum_{i=1}^{N} \sum_{j \in C(i)} B_{ij} \]  (4)

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

\[ r_{GP}^A(U) = \sum_{i=1}^{N} \sum_{j \in C(i)} B_{ij} \]  (4)

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 \( e_i = u_i \rightarrow e_i = \hat{e_i} \) rather than
whether \( i \) is actually correctly classified. This means we really care about whether a node has
been misclassified. The expected loss due to mislabeling can be calculated

\[ E[Y(Z = 1) - Y(Z = 0)] = E[	au | U_i = \hat{U}_i] + \gamma U \sum_{j \in C(i)} B_{ij} | U_i = \hat{U}_i] \]

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 randomization
* 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

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


Contact Info

heather.mathews@duke.edu