Estimating Policy Effects in a Network through Independent Set Sampling

Guest Lecture

Eugene Ang

eugene.ang@u.nus.edu

National University of Singapore

Mar 6th , 2023

Problem Statement

- When policy makers want to understand the effects of their policies, they would conduct RCT or A/B testing on a group of people in the population.

- In most cases, the sampling techniques used such as random sampling, do not account for the underlying network relationship that the population possesses.

- Network effect could affect how policies influence the behaviors of the people in the communities and how their relationships could influence the effect of the policies.
  1. Isolate the “direct” effect of the policy change from any “indirect” effect of the policy change via network influence
  2. Find the “net effect” of a policy change in the presence of homophily and network influence in the population.

Current Approaches

1. Random Selection with Naïve linear regression
    - Regress on observable covariates to explain the policy effects

2. Linear-in-means model
    - Use aggregated values of nodes' neighbors as instrumental variables to explain peer effect

3. Graph clustering selection
    - Sample random clusters in network for external test exposure

4. Co-evolution model
    - Jointly model the evolution of networks and behavioral dynamic as a Markov chain after conditioning on initial state

Why should we care?

1. Presence of network interference within treatment groups and across groups within network

2. Bias could over/under correct the policy effect
   - Estimation of policy influence is generally confounded with homophily
   - Better manage the resources for the policy implementation

3. Current approaches address such problems in stylized, assumption-intensive contexts
   - Econometrics method: Cannot guarantee the strength of IVs and would fail in certain network structure
   - Cluster sampling: Vulnerable to network interference within sampled clusters

Several questions to ask…

1. How do we reduce the bias and estimate the true policy effects?
2. Do different samples of respondents affect the estimation of the policy effects?
3. How do we identify and separate these effects from confounders?

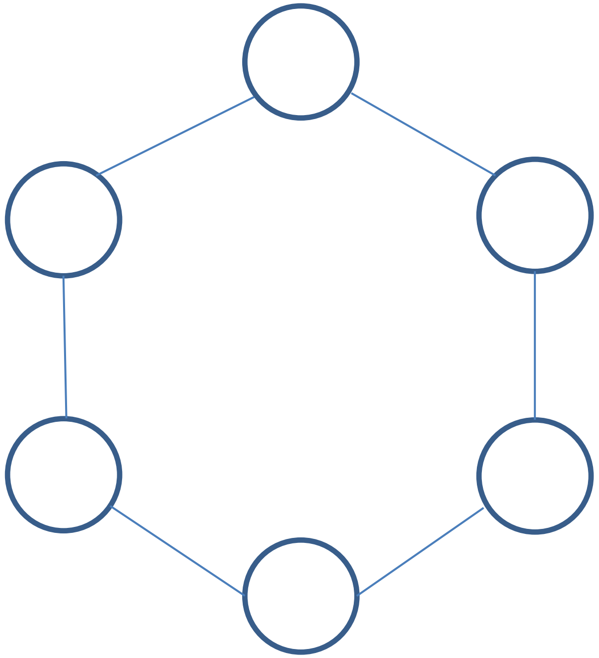
Proposed Methodology

Combines existing work in stochastic actor-oriented models (SAOM) with an independent set sampling

→ So, what is an independent set?

Independent Set

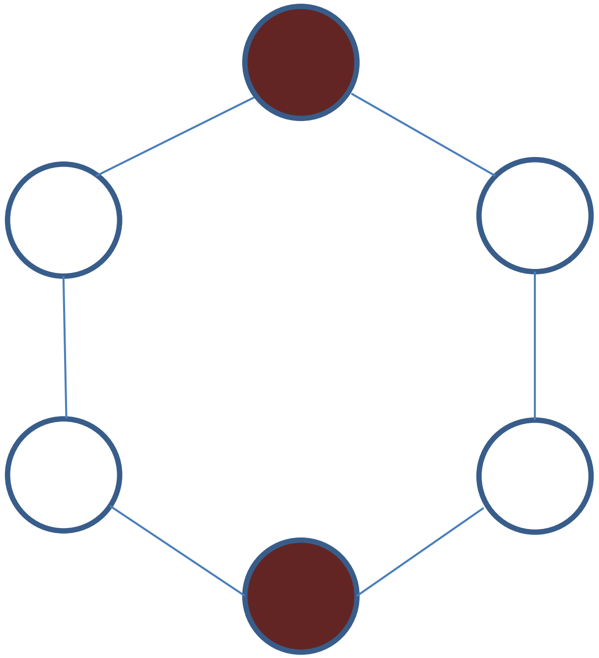
Definition: A set of vertices S is called an independent set if no two vertices in this set S are adjacent to each other



Let’s call this graph G, technically it’s C_6.
Let construct some independent set S

Independent Set

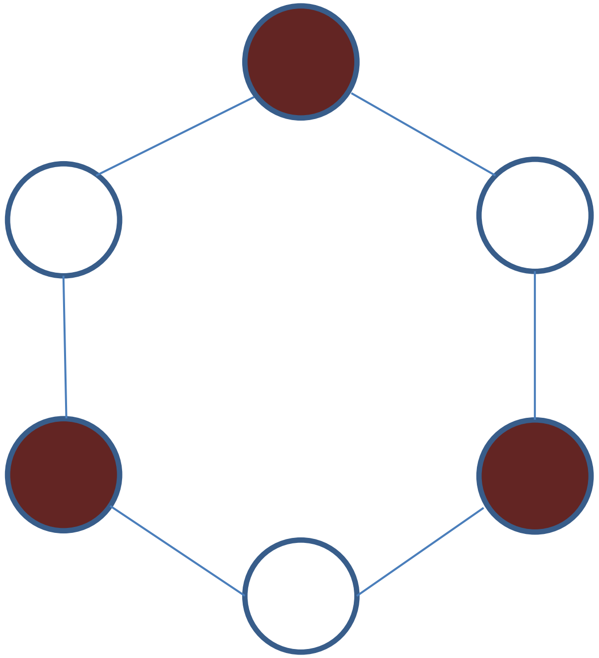
Definition: A set of vertices S is called an independent set if no two vertices in this set S are adjacent to each other



This is one possible S. It is a **maximal** independent set.

Independent Set

Definition: A set of vertices S is called an independent set if no two vertices in this set S are adjacent to each other



This is another possible S. It is a **maximum** independent set.

Why use independent set sampling?

- Social networks are known to be sparse and have bounded degree
  - Independent sets can be used to sample large numbers of nodes relatively efficiently

- Obtain a more representative sample of a network
  - Ensure that the sample is not overly influenced by the presence of dense subgraphs in the original network

- Eliminate interactions within sample groups and reduce the bias
  - Through such construction, it avoids selecting connected groups of nodes, so it eliminates the bias in any of such sample from the network

- Better identify the potential network formation within sample group due to policy
  - Since the sample is isolated by construction, any network formation is purely due to homophily of being exposed to treatment

Some applications

1. Economic/political policies implementation such as taxes, subsidies
   - Study the reception among the masses and the effectiveness of policies

2. Policies in workplace/social/financial network such as collaboration incentive
   - Investigate potential connection between isolated clusters and their tangible benefits

3. Graph-based learning
   - Generate subsets from large graphs for machine learning tasks, which can lead to a more efficient training process and help the generalization of the model

So the plan is…

1. For a given network, we find an independent set/cluster sample/random sample to be exposed to the treatment

2. Simulate the evolution of the network and behavioral dynamic using the co-evolution model

3. Obtain estimates for homophily and influence; compare across the 3 samples

4. On top of it, we use a (pseudo)-contagion model to detect the spread of the policy onto the network; compare across the 3 samples

Use case: Cigarette Price on Smoking

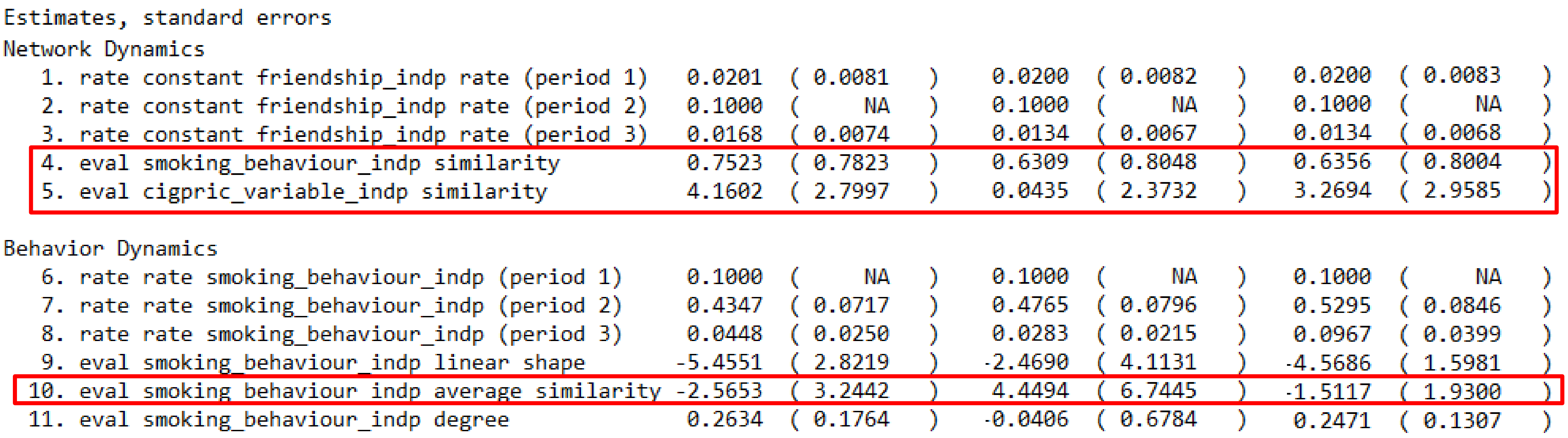
- ➢ Want to investigate the effects of smoking behavior by increasing cigarette prices

- Use a smoking data from R (with random assignment on gender)

- Model a logistic regression based on covariates at individual level

- Create 3 waves (4 stages) for co-evolution model to simulate evolution
  1. Initialize random network
  2. Choose an independent set sample/random sample/cluster sample with small noise
  3. Change behavior according to logistic regression (no change in network)
  4. Parameterize the evolution based on certain probabilities of change

Results from co-evolution model



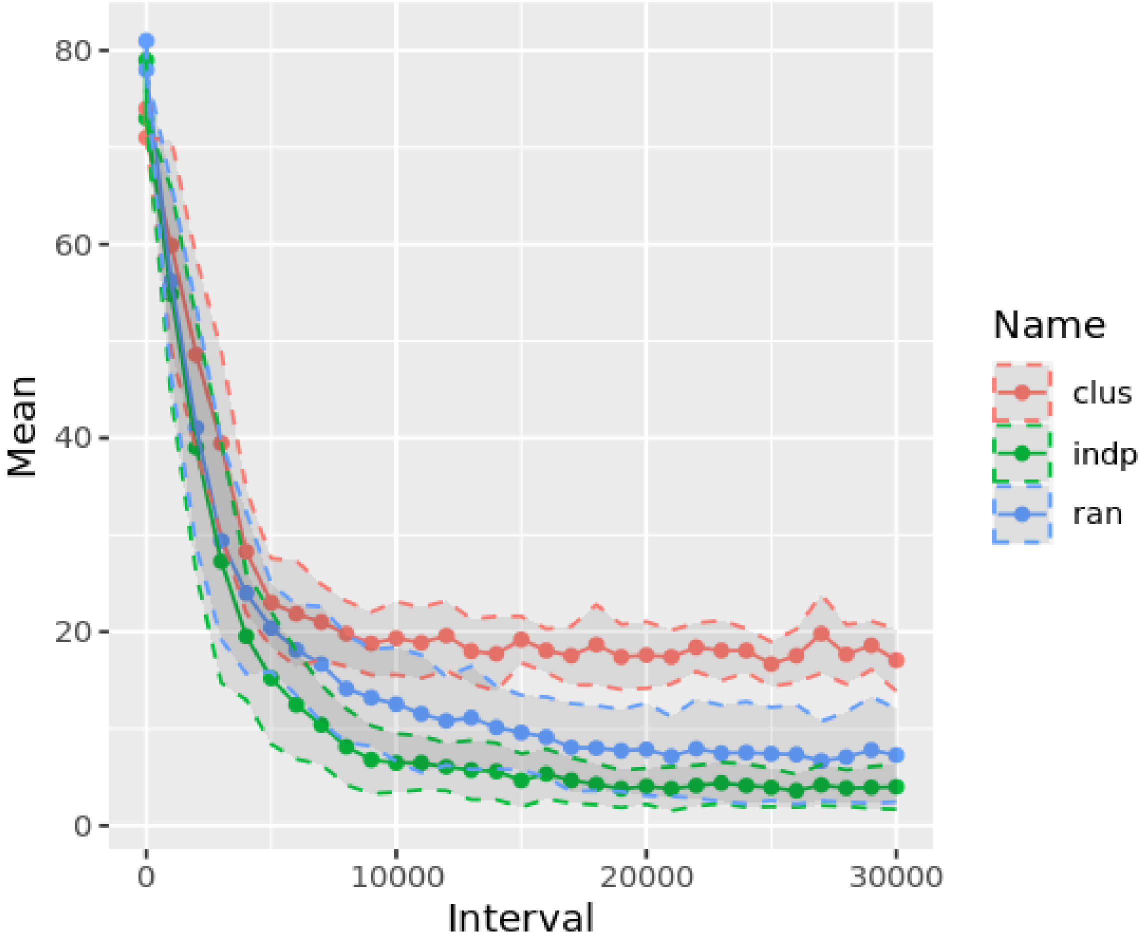
```
Estimates, standard errors
Network Dynamics
   1. rate constant friendship_indp rate (period 1)   0.0201  ( 0.0081   )    0.0200  ( 0.0082   )    0.0200  ( 0.0083   )
   2. rate constant friendship_indp rate (period 2)   0.1000  (      NA   )    0.1000  (      NA   )    0.1000  (      NA   )
   3. rate constant friendship_indp rate (period 3)   0.0168  ( 0.0074   )    0.0134  ( 0.0067   )    0.0134  ( 0.0068   )
   4. eval smoking_behaviour_indp similarity          0.7523  ( 0.7823   )    0.6309  ( 0.8048   )    0.6356  ( 0.8004   )
   5. eval cigpric_variable_indp similarity           4.1602  ( 2.7997   )    0.0435  ( 2.3732   )    3.2694  ( 2.9585   )

Behavior Dynamics
   6. rate rate smoking_behaviour_indp (period 1)     0.1000  (      NA   )    0.1000  (      NA   )    0.1000  (      NA   )
   7. rate rate smoking_behaviour_indp (period 2)     0.4347  ( 0.0717   )    0.4765  ( 0.0796   )    0.5295  ( 0.0846   )
   8. rate rate smoking_behaviour_indp (period 3)     0.0448  ( 0.0250   )    0.0283  ( 0.0215   )    0.0967  ( 0.0399   )
   9. eval smoking_behaviour_indp linear shape       -5.4551  ( 2.8219   )   -2.4690  ( 4.1131   )   -4.5686  ( 1.5981   )
  10. eval smoking behaviour indp average similarity -2.5653  ( 3.2442   )    4.4494  ( 6.7445   )   -1.5117  ( 1.9300   )
  11. eval smoking_behaviour_indp degree              0.2634  ( 0.1764   )   -0.0406  ( 0.6784   )    0.2471  ( 0.1307   )
```

Independent Cluster Random

- A higher cigarette price is a proxy of being included in the treatment set
- Observe a higher homophily based on smoking behaviour and cigarette price for independent set

Results from (pseudo)-contagion model

- Incidence curve to measure the spread of policy effect onto the entire network
- Run a (pseudo)-contagion model to obtain mean of number of smokers remaining in the network after some time
- Observe that independent set sampling bring about lower number of smokers over time



Key findings

1. Through independent set sampling, we eliminate any network interference within the treatment group
   - The bias in the estimated measures are greatly reduced, thus uncovering the true policy effect

2. By implementing policy on independent sample sets, we attain
   - Higher homophily based on targeted behavior and policy exposure
   - Greater and faster coverage of intended policy effects throughout the network

3. Encourages network formation through policy implementation

4. Policy makers can spend less resources by exposing the policy on an independent sample and let the network do the work

FAQ – Future Work

1. What if we obtain a weakly independent set sample due to incomplete data/unobservable links? Any sensitivity/robustness analysis?

2. Do size/certain centrality measures in the independent set affect the speed of influence/coverage?

3. Since the construction of independent set is affected by the graph structure, how would different graph structure affect the effectiveness of such sampling?

4. Which is the “best” independent set to use, in terms of cost of policy implementation or rate of coverage?

Thank you

Please like, share and cite

Eugene Ang

eugene.ang@u.nus.edu

National University of Singapore
Mar 6th , 2023