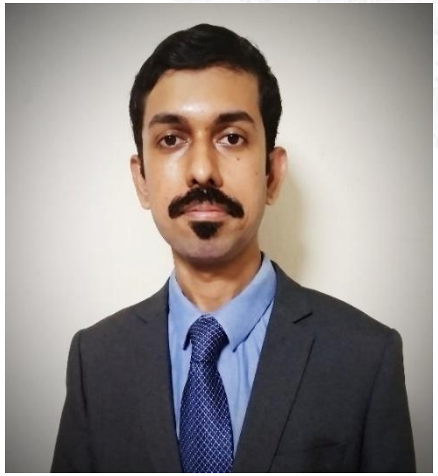
Estimating Policy Effects in a Social Network with Independent Set Sampling

Eugene Ang[1]

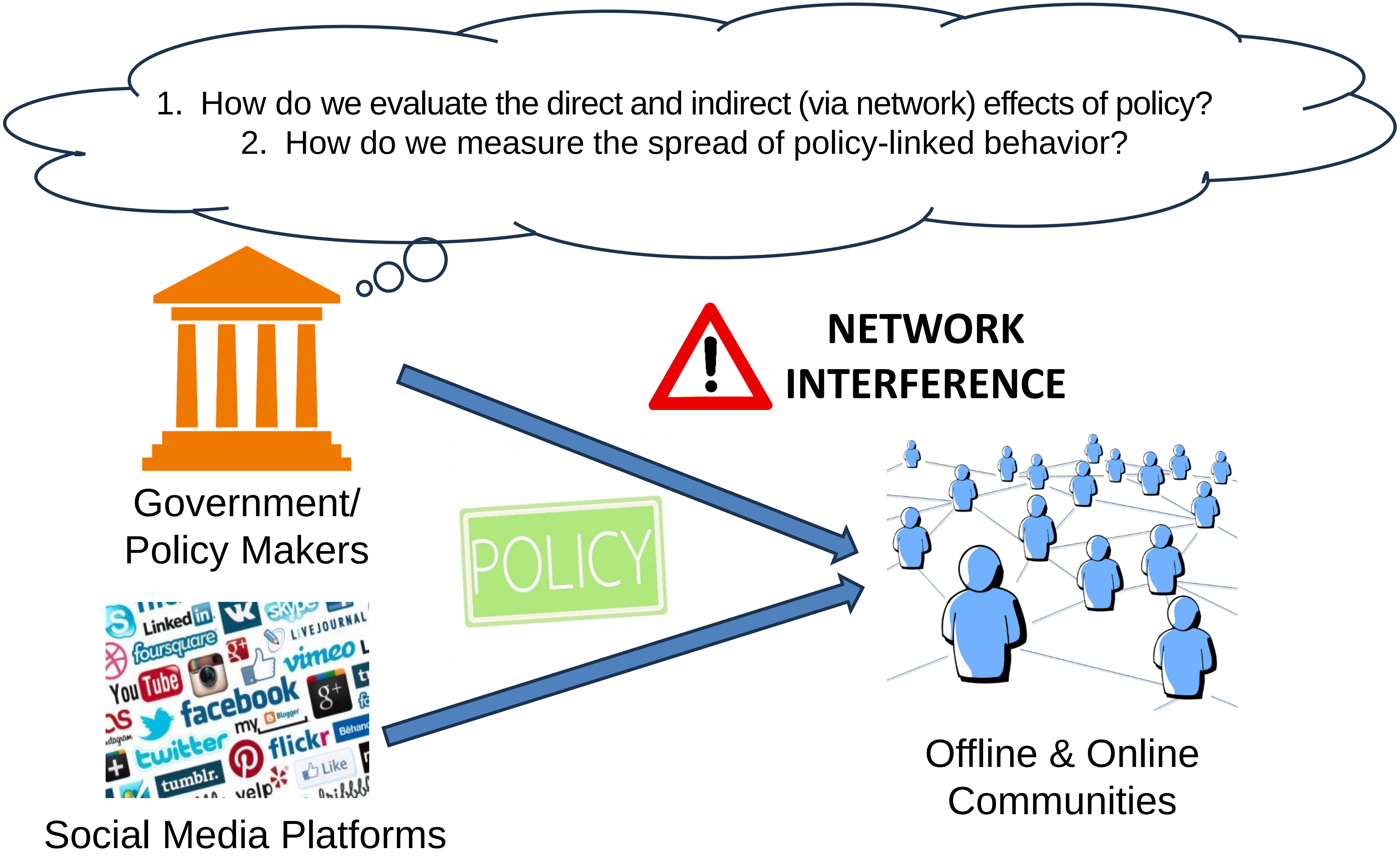


Prasanta Bhattacharya[2]

Andrew Lim[1,2]

[1] Institute of Operations Research and Analytics (IORA), National University of Singapore (NUS)
[2] Department of Analytics and Operations (DAO), NUS Business School

Sunbelt 2023
June 30th 2023


1. How do we evaluate the direct and indirect (via network) effects of policy?
2. How do we measure the spread of policy-linked behavior?
NETWORK INTERFERENCE
Government/ Policy Makers
POLICY
Social Media Platforms
Offline & Online Communities

Policy Evaluation Approaches

- Use (quasi) – experimental approaches or conduct RCT on a group of people in the population *(Coly 2017, White 2017)*

    - Affected by network interference and other network-related confounders, such as homophily

- Require appropriate and efficient sampling methods *(Olsen 2013)*

    - Risk of network interference “within” the treatment and non – treatment groups

Network Interference

- Network interference 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” treatment 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 *(e.g., Porter 1981)*
    - Regress on observable covariates to explain the policy effects

2. Linear-in-means model *(e.g., Manski 1993, Kline 2012)*
    - Use aggregated values of nodes' neighbors as instrumental variables to explain peer effect

3. Graph clustering selection *(e.g., Ugander 2013)*
    - Sample random clusters in network for external test exposure

4. ERGM / Co – evolution model *(e.g., Wasserman 1996, Snijders 2007)*
    - Model social network structures through specified statistics and properties

5. Contagion model *(e.g., Carrington 2005, Rogers 2019)*
    - Treat the exposed respondents 'infectious', and model the spread of the policy-linked behavior to subsequent contacts of these initial respondents

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 have certain limitations
    - Econometric approaches: Cannot guarantee the strength of IVs, and might fail for certain network structures
    - Cluster sampling: Vulnerable to network interference “within” sampled clusters

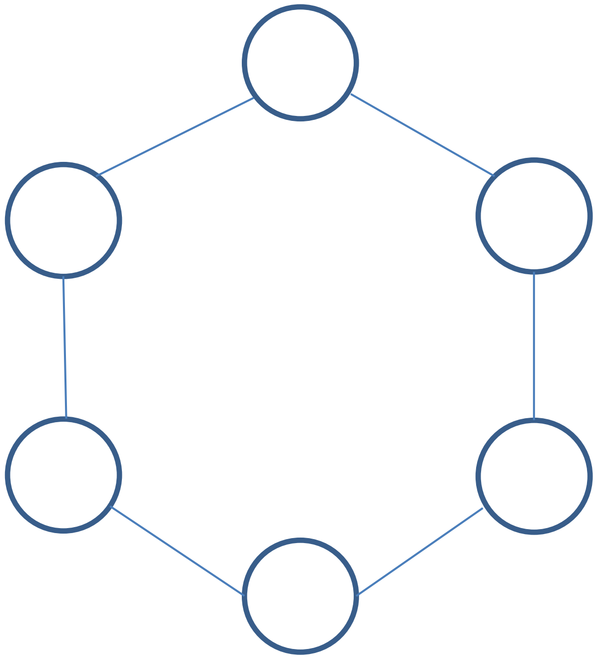
Proposed Methodology

Combines existing work in stochastic actor-oriented models (SAOM) & diffusion contagion models with *independent set sampling technique*

→ So, what is an independent set?

Independent Set

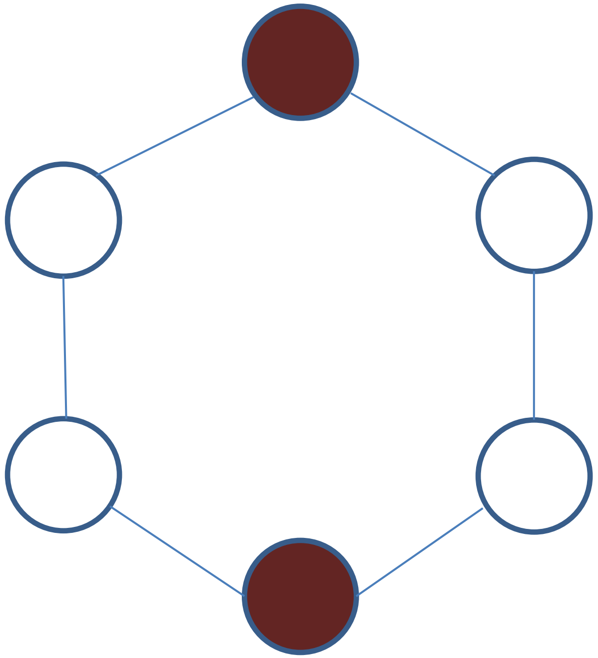
Definition: A set of vertices S is called an independent set if no two vertices in this set S are adjacent to each other



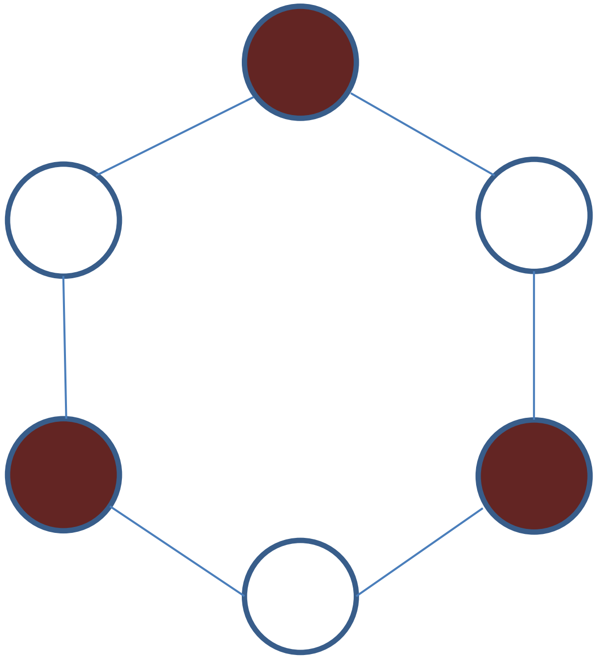
Let's call this graph G, technically it's C_6.
Let construct some independent set S

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.



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

Bounds

Let $G = (V, E)$, $|V| = n, |E| = e, d_v$ be degree of vertex v, Δ is the maximum degree, $\alpha(G)$ is the size of the maximum independent size

Theorem (Kwok): $\boldsymbol{\alpha(G) \leq n - \frac{e}{\Delta}}$

Theorem (Caro – Wei): $\boldsymbol{\alpha(G) \geq \sum_{v \in V} \frac{1}{1+d_v}}$

Guarantee on the size of the largest independent set sample

- Flexible to choose a suitable sample size given their resource constraints and objective

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 isolate policy effects
  - Through such construction, it avoids selecting connected groups of nodes, so it reduces the chances of treatment spill-over in any such sample from the network

- Better identification of network formation within sample (due to policy change)
  - Since the sample is isolated by construction, any network formation can be attributed to homophily of being exposed to treatment (or due to chance)

So, the plan is…

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

2. Simulate the joint evolution of the network and behavior using a stochastic co-evolution model

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

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

Simulation Study

- ➢ Want to investigate the effects on focal behavior due to the policy

- Use a popular dataset accessed from the *wooldridge* package in R *(with random assignment on gender)*

- Model changes in focal behavior using a logistic regression based on individual covariates (policy – increase price level of goods)

- Create 3 waves (4 stages) to simulate evolution
  1. Initialize random scale free 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

SAOM

- Assume loss in utility is equal to respondent's earlier gain. No endowment function

- Model rate and objective functions with network statistics such as
  1. Degree
  2. Transitivity
  3. Homophily based on the respondent's covariates (focal behavior and price)

  and behavioral statistics such as
  1. Similarity measure
  2. Behavioral tendency effect
  3. Peer influence effect

Note: these are not exhaustive, and exact selection will depend on the specific problem context

Stronger homophilous effect

Estimates, standard errors				
Network Dynamics		Independent	Cluster	Random
1	Friendship rate (Period 1)	0.0100** (0.0057)	0.0100** (0.0057)	0.0101** (0.0057)
2	Friendship rate (Period 2)	0.1000 (NA)	0.1000 (NA)	0.1000 (NA)
3	Friendship rate (Period 3)	0.0201*** (0.0081)	0.0134*** (0.0066)	0.0100** (0.0058)
4	Transitivity	-4.9743 (9.8651)	0.0269 (1.0338)	-3.9669 (7.9918)
5	**Behavior homophily**	**1.9260** (1.1031)**	**1.7064 (1.1838)**	**1.4226 (1.1266)**
6	**Policy exposure homophily**	**6.6035** (3.5597)**	**5.0359 (3.6011)**	**2.7138 (3.2889)**
Behavior Dynamics				
7	Behavior rate (Period 1)	0.1000 (NA)	0.1000 (NA)	0.1000 (NA)
8	Behavior rate (Period 2)	0.5439*** (0.0880)	0.4409*** (0.0770)	0.5587*** (0.0854)
9	Behavior rate (Period 3)	0.0158 (0.0156)	0.0685*** (0.0283)	0.1000 (NA)
10	Behavior Tendency (Linear Shape)	-11.1635 (14.7546)	-2.5606*** (1.1993)	0.5258 (3.2434)
11	**Average Peer Influence**	**-0.0035 (6.7040)**	**0.0654 (1.6598)**	**0.4630 (1.4848)**
12	Outdegree	0.9312 (1.4059)	0.0323 (0.1417)	-0.8170 (1.3339)

- A higher price level serves as a proxy for being included in the treatment set
- Observe a higher homophily based on focal behaviour and price level for independent set

Estimating treatment effect

- Compute differences in the proportion of individuals having the focal behavior in both the treatment and non-treatment groups

- Track the difference estimates over 4 time periods

  (A) Before the policy implementation

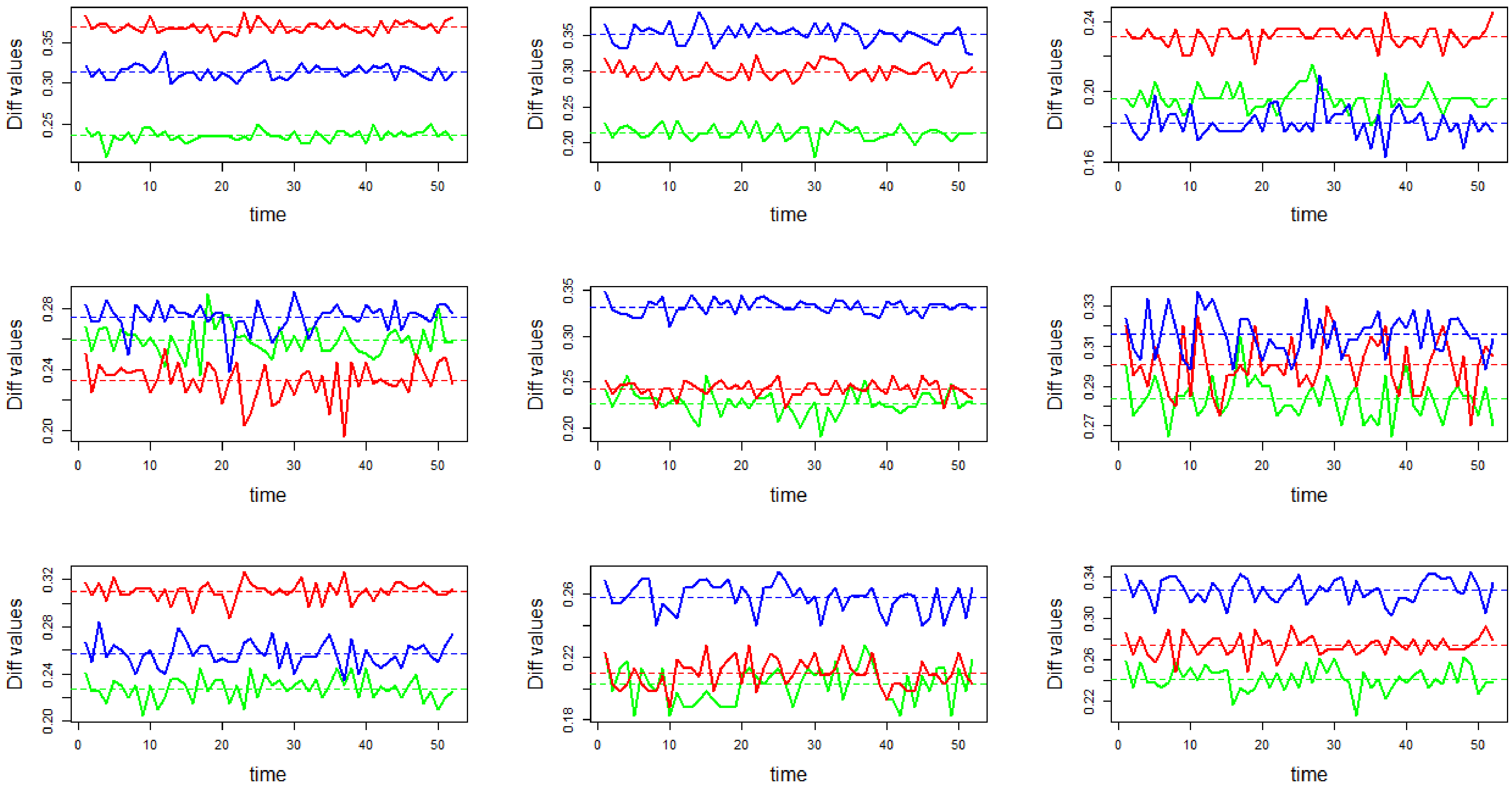
  (B) Right after the policy implementation

  (C) After one wave of simulated evolution

  (D) Future epochs of the predicted networks based on SAOM

- Direct effect: B-A, Short-term net treatment effect: C-A, Long-term net treatment effect: D-A

Distinct *direct* and *net* treatment effect of policy



Difference values for different sampling strategies

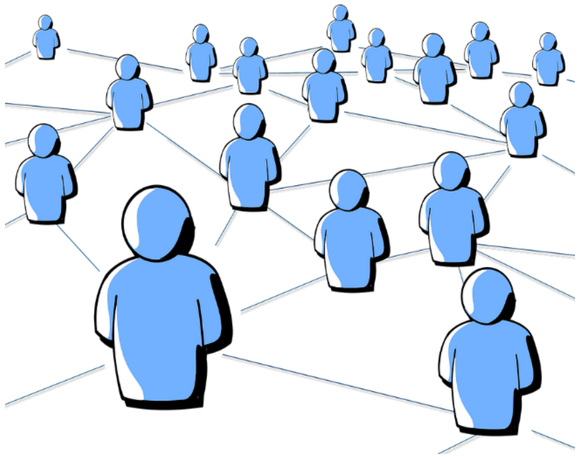
indp ran clus

So, the plan is…

1. How do we evaluate the direct and indirect (via network) effects of policy?
2. How do we measure the spread of policy-linked behavior?

Government/
Policy Makers

Social Media Platforms



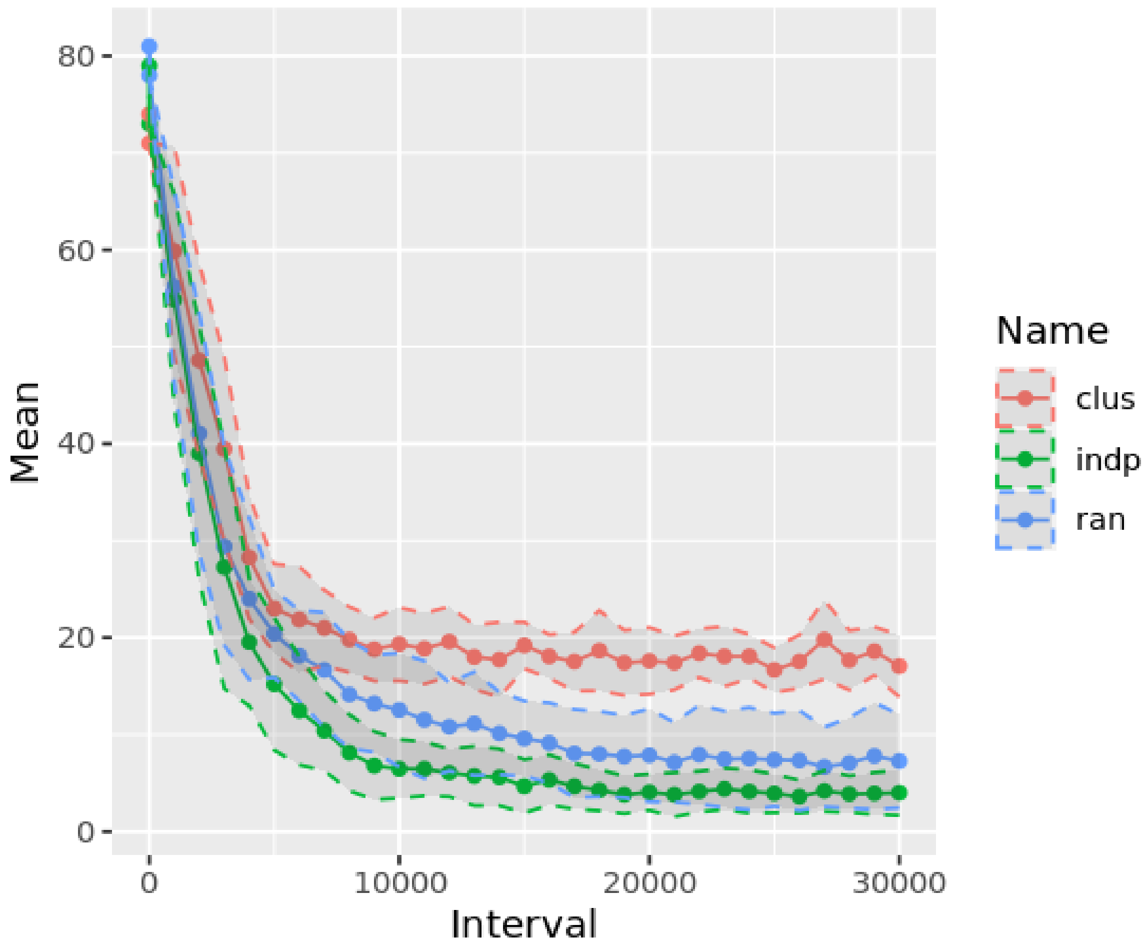
Offline & Online
Communities

Diffusion Contagion Model

- Model the spread of policy – linked behavior through a dynamic network as a continuous time Markov process
- Equip each respondent with a behavioral threshold generated from the standard uniform distribution
- Respondent acquires focal behavior if the proportion of alters, who have the same behavior, exceeds the threshold
- Increase the weight of the alters in the treatment group by the difference in probability computed in the earlier logistic regression
- Get an opportunity to change an edge or behavior at each micro-step with equal probability

Larger and faster policy coverage

- Incidence curve to measure the spread of policy effect onto the entire network
- Run a dynamic diffusion contagion model to obtain mean of number of individuals with focal behavior over time
- Observe that independent set sampling brings about lower number of respondents with focal behavior over time



Key findings

1. Through independent set sampling, we eliminate any network interference within the treatment group
   - Leads to better estimation of the treatment effect

2. By implementing policy on independent sample sets, we attain
   - Higher homophily based on focal 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

Future Work

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

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. Any questions?

Eugene Ang

eugene.ang@u.nus.edu

https://arxiv.org/abs/2306.14142