Stochastic seeding strategies in networks

Dean Eckles[1] [2]

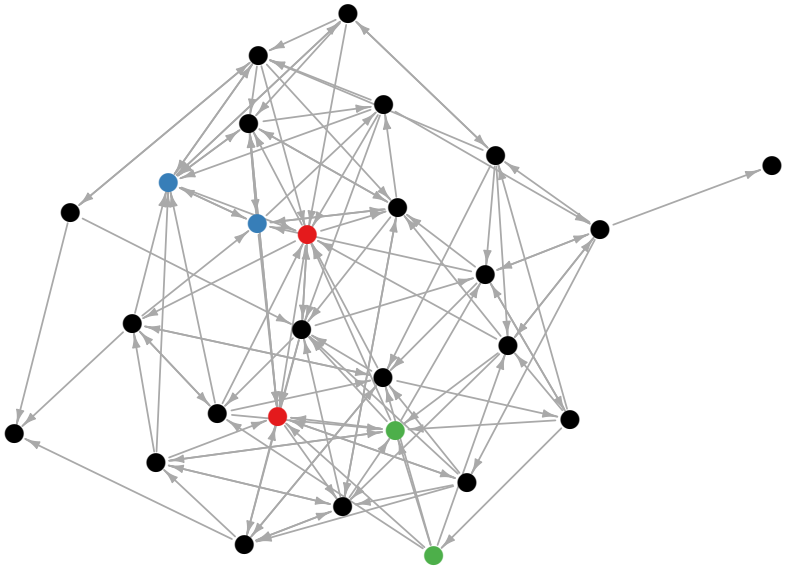
April 24, 2019

[1]Sloan School of Management, Massachusetts Institute of Technology.
[2]Institute for Data, Systems & Society, Massachusetts Institute of Technology.

Evaluating stochastic seeding strategies in networks.
with Alex Chin and Johan Ugander
Working paper. https://arxiv.org/abs/1809.09561

Seeding with partial network information: Hardness and guarantees.
with Hossein Esfandiari, Elchanan Mossel and M. Amin Rahimian
Working paper.

Seed sets and network-level outcomes



$y_i(S = \quad)$
$y_i(S = \quad)$
$y_i(S = \quad)$

Seeding in networks

How can information about a social network be used to target interventions?

- Networks may reveal otherwise latent traits due to homophily etc.
- Social contagion may mean that some vertices are much more influential.
  - Kempe et al. (2003) show that this is NP-hard and provide a greedy approximation assuming submodularity. (Huge subsequent literature.)
- How valuable is measuring the network?
  - Akbarpour et al. (2017) argue that $k +$ random seeds outperform k optimal seeds for small , though this is in an Erdos–Renyi random graph.

Seeding in networks

Can we use the network to seed without measuring the whole network?

- Some “name generators” ask people who is most trusted by them — and others
- Can approximate targeting vertices with max in-degree by sampling from network
- *One-hop targeting* randomly selects vertices, and then takes one step at random. This is repeated to yield k seeds.
- This is designed to exploit a “local” version of the friendship paradox (Feld 1991; Kumar et al. 2018).

One-hop targeting

- Kim et al. (2015) randomly assigned villages to random, one-hop, and maximum in-degree targeting.
- Other applications:
  - vaccination/fragmentation (Cohen et al. 2003; Gallos et al. 2007; Chami et al. 2017)
  - sensing (Leskovec et al. 2007; Christakis and Fowler 2010)

Evaluating one-hop targeting

- Kim et al. (2015) randomly assigned villages to random, one-hop, and maximum in-degree targeting.[3]
- **Village-level difference-in-means estimator:**

$$\hat{\tau}_{DM} = \frac{1}{\sum_{i=1}^{n} z_i} \sum_{i=1}^{n} z_i y_i \quad - \quad \frac{1}{\sum_{i=1}^{n} (1 - z_i)} \sum_{i=1}^{n} (1 - z_i) y_i$$

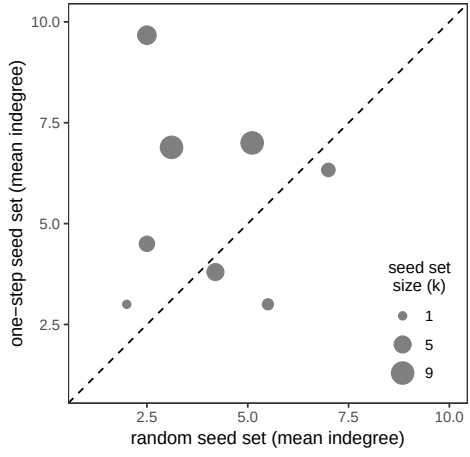
  where $z_i = 1$ iff village i is assigned to one-hop and y_i is a village-level outcome (i.e., number or fraction of adopters).
- $\hat{\tau}_{DM}$ is unbiased for both finite- and infinite-population estimands. But it makes inefficient use of the data.

[3] This uses a blocked design, but ignore that for now and think of a Bernoulli trial at the village-level.

One-hop is importantly stochastic

One village from Kim et al. (2015). Red is random seed set, green is nomination seed set.

One-hop is importantly stochastic



Random targeting

In random targeting, all eligible seed sets are equally likely, so p_i^{rand} is characterized by the uniform probabilities

$$P_i^{rand}(S_i = s) = \binom{m_i}{k}^{-1}, \qquad \text{for all } s \in S_i,\ i = 1, \ldots, n, \tag{1}$$

Notice that this is independent of the graph G_i, the network structure of village i.

One village from Cai et al. (2015)'s study of rural Chinese farmers purchasing insurance.

Consider one-hop with $k = 1$.
Nodes sized proportional to

$$P_i^{hop}(S_i = f v g) = \frac{1}{m_i} \sum_{u2N_{in}(v)} \frac{1}{d_u^{out}}$$

Illustration of one-hop probabilities

Estimand

Goal: estimate the difference in expected outcomes for each of the two targeting strategies, A and B.

- Finite population version:

$$_{fp} = \frac{1}{n} \sum_{i=1}^{n} E_A^i[y_i(S_i)] \quad E_B^i[y_i(S_i)] \ , \tag{2}$$

where E_A^i and E_B^i denote expectation over S_i p_A^i and S_i p_B^i, respectively.

- Superpopulation version:

$$_{sp} = E_A[y(S)] \quad E_B[y(S)]. \tag{3}$$

Notice that $_{sp}$ can be written as

$$_{sp} = E \quad \frac{p_A(S) \quad p_B(S)}{p \ (S)} y(S) \ .$$

Design and target distributions

- Bernoulli design: Z_i Bernoulli( ), where 2 (0, 1) is the treatment assignment probability, yield a mixture:

$$S_i \quad p_i^A + (1 \quad) p_i^B.$$

- This is the design distribution:

$$P_i\ (S_i = s) = \quad P_i^A(S_i = s) + (1 \quad) P_i^B(S_i = s). \qquad (4)$$

- p_i^A, p_i^B, and p_i are all completely known, so we know the exact probabilities corresponding to the observed seed sets. Let s_i be the observed seed set for village i and let

$$a_i = P_i^A(S_i = s_i), \quad b_i = P_i^B(S_i = s_i), \quad {}_i = P_i\ (S_i = s_i)$$

Estimators: Horvitz–Thompson

- De ning the weights

$$w_i^A = a_i = {}_i, \qquad w_i^B = b_i = {}_i.$$

- Horvitz–Thompson estimator:

$$\hat{} = \frac{1}{n} \sum_{i=1}^{n} (w_i^A \quad w_i^B) y_i .$$

- This is the minimum variance unbiased estimator, but often too noisy.

Estimators

- Or de ne the normalized weights

$$\tilde{w}_i^A = \frac{w_i^A}{n^{-1}\sum_j w_j^A}, \qquad \tilde{w}_i^B = \frac{w_i^B}{n^{-1}\sum_j w_j^B},$$

- Hájek estimator:

$$\tilde{\ } = \frac{1}{n}\sum_{i=1}^{n} (\tilde{w}_i^A \quad \tilde{w}_i^B) y_i . \tag{5}$$

- Typically reduces variance compared with Horvitz–Thompson (unnormalized) estimator.

Inference

- Standard Neyman-style variance estimates for the nite population treatment effect $_{fp}$ have problems here. We focus on $_{sp}$.

Proposition

Let S p be a random seed set and let $P_A = p_A(S)$, $P_B = p_B(S)$, and P = p (S) be random variables representing the probabilities corresponding to seed set S. Let $Y = y(S)$, $W_A = P_A/P$, and $W_B = P_B/P$. Then the Horvitz–Thompson estimator ^ has expectation E[^] = $_{sp}$ and Var(^) = $V_{\wedge}/n$, where

$$V_{\wedge} = E\left[\frac{1}{P^2}((P_A \quad P_B)Y \quad {}_{sp}P\ )^2\right] = E[((W_A \quad W_B)Y \quad {}_{sp})^2]. \quad (6)$$

Inference

- We can construct an unbiased estimate of this variance using sample analogs.

$$\hat{\mathbb{V}}_{\hat{}} = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\frac{2}{i}}((a_i \quad b_i)y_i \quad \hat{} \ _i)^2 = \frac{1}{n}\sum_{i=1}^{n} ((w_i^A \quad w_i^B)y_i \quad \hat{})^2. \quad (7)$$

Inference

- Hájek estimator is not unbiased, but is consistent and asymptotically normal:

Proposition

Let S, P_A, P_B, P , and Y be as in Proposition 1. Let $_A = E_A[Y]$ and $_B = E_B[Y]$. Then the Hájek estimator $\sim$ satis es $\sim$! $_{sp}$ as n ! 1 and

$$p\overline{n}(\sim \quad _{sp}) \quad N(0, V_\sim),$$

where

$$V_\sim = E\left[\frac{1}{P^2} \quad _A P_A \quad _B P_B \quad Y(P_A \quad P_B)\right]^2 . \qquad (8)$$

Effective sample size: Designs and estimators

Effective sample size across multiple data sets with $k = 2$ seeds per unit.

Dataset	n	Population effective sample size n_{eff} ~, Bernoulli(0.5)	~, optimal	~, off-policy
Cai et al.	150	631.72	871.16	233.36
Paluck et al.	56	351.60	539.04	128.34
AddHealth	85	319.36	448.80	131.04
Banerjee et al.	75	214.32	274.08	80.24
Chami et al.	17	37.84	47.92	9.32

Estimator provides large gains. Additional gains available from optimal design. Off-policy evaluation (from random seeding only) often beats difference-in-means with Bernoulli design.

Re-analysis of Cai et al. (2015)

- Experimental setup: Farmer's insurance in 150 rural Chinese villages.
  - Treatment: Intensive information session about insurance product to seed set; in rst round
  - Response: Fraction of adopters in village
- All villages were assigned to a uniform random strategy (p^{rand}), but our method actually enables off-policy evaluation, letting us estimate the mean response under both p^{rand} and one-hop targeting p^{hop}.
- This is thus off-policy estimation

Probabilities, weights, and outcomes in Cai et al. (2015) reanalysis.

Results for Cai et al.

Effect of seeding strategy on insurance purchase

estimate (one-hop rand)	-0.0436
SE (analytic)	0.0209
SE (bootstrap)	0.0257
95% CI (analytic)	[-0.0846, -0.0027]
95% CI (bootstrap)	[-0.0909, 0.0088]
p-value (analytic)	0.0367
p-value (Fisherian)	0.0974

Hájek estimate and inference for the difference in insurance takeup rates between one-hop and random seeding for Cai et al. (2015), which provide some evidence that one-hop seeding would have reduced adoption of insurance.

Re-analysis of Paluck et al. (2016)

- Experimental setup: Reducing peer con ict in 56 middle schools in New Jersey
  - Treatment: Invitation to participate in program encouraging taking a public stand against peer con ict
  - Response: Peer con ict events per student, from administrative records
- 28 schools were assigned to have any treated students. These students were selected randomly with blocking on grade and gender.
- We conduct off-policy evaluation for one-hop seeding, conditional on balancing grade and gender.

Probabilities, weights, and outcomes in Paluck et al. (2016) reanalysis.

Results for Paluck et al.

Effect of seeding strategy on rates peer con ict

estimate (one-hop rand)	0.0997
SE (analytic)	0.0231
SE (bootstrap)	0.0432
95% CI (analytic)	[0.0543, 0.1451]
95% CI (bootstrap)	[0.0098, 0.1542]
p-value (analytic)	1.7e-05
p-value (Fisherian)	0.0846

Hájek estimate and inference for the difference in peer con ict events per student between one-hop and random seeding for Paluck et al. (2016), which provide some evidence that one-hop seeding would have increased peer con ict.

Comparing fraction of seeds who are “social referents” (top decile of in-degree) and weights for one-hop vs. random contrast. (On right, the school with very large positive $w_i^{hop} - w_i^{rand}$ (17.8) is not shown; 17% of its seeds were social referents.)

Summary

- One-hop seeding is a practically and theoretically appealing seeding strategy
- We develop designs and estimators for evaluating it and other such stochastic strategies
- Very large precision gains are possible, especially from the proposed estimators
- Preliminary evidence from Cai et al. (2015) and Paluck et al. (2016) actually suggests one-hop may be no better, or even worse, than random seeding

Next steps: More sophisticated stochastic seeding

- One-hop discards much of the network information it collects.
- Can we design better algorithms for seeding when network information is costly?
- New working paper with Hossein Esfandiari, Elchanan Mossel and M. Amin Rahimian

Algorithm 1: PROBE (, T,)

Input: Query access to graph G

Output: $G_{,}^{(1)}, \ldots, G_{,}^{(T)}$

1. SAMPLE: Choose n nodes uniformly at random and call them V .
2. PROBE: For each node v in V :
   - Probe the neighborhood of v, asking each person to reveal the identity of each of their neighbors with probability p. Call any such person who is asked to reveal the identity of their neighbors a probed node.
   - Proceed to probe the newly revealed nodes, only if they are not probed before.
   - Stop probing if there are no more new nodes to probe or if the size of the connected component containing v exceeds .
3. REPEAT: Repeat the PROBE step T times to obtain T independent copies $G_{,}^{(1)}, \ldots, G_{,}^{(T)}$.

Algorithm 2: SEED (^{0})

Input: The T copies, $G^{(1)}_{,}, \ldots, G^{(T)}_{,}$
Output: $^{?}$, (1 1=e ^{0})-approximate solution to k-IM for $^{T}_{,}$

1. Find the connected components of $G^{(1)}_{,}, \ldots, G^{(T)}_{,}$
2. For every connected component in each of the T copies initialize the current value of the component equal to the number of sampled points (belonging to V) in that component.
3. Initialize $^{?} = \varnothing$
4. For i form 1 to k, do:
   - 4.1 Choose a random subset, R V n $^{?}$, $jRj = (n/k) \log(1/ ^{0})$.
   - 4.2 For each v 2 R compute $(vj$ $^{?})$ by adding the current values of the connected components containing v and set $v^{?} = \arg\max_{v2R}$ (vjS)
   - 4.3 Add $^{?}$ $^{?}$ [$fv^{?}g$ and set the current value of the connected components containing $v^{?}$ to zero.

Proposition (Approximation guarantee with bounded edge queries)

For any arbitrary $0 <$ 1, *there exist a polynomial-time algorithm for influence maximization that covers* $(1 \quad 1/e)L \quad n$ *nodes in expectation, using* $O(pn^2 + \sqrt{p}\,\bar{p}n^{1.5})$ *queries, where* L *is the expected number of nodes covered by the optimum seed set.*

Thanks

We thank Jing Cai, Elizabeth Paluck, Hana Shepherd, and Peter Aronow for assistance in working with the data from their field experiments. This work benefited from comments by John Hauser, Maurits Kaptein, Duncan Simester, Juanjuan Zhang, and Yunhao Zhang. This work was supported in part by NSF grant IIS-1657104.

Chin, A., Eckles, D., & Ugander, J. (2018). Evaluating stochastic seeding strategies in networks. Working paper.
https://arxiv.org/abs/1809.09561

References

Akbarpour, M., Malladi, S., and Saberi, A. (2017). Just a few seeds more: Value of targeting for diffusion in networks. Working paper, Stanford Graduate School of Business.

Cai, J., De Janvry, A., and Sadoulet, E. (2015). Social networks and the decision to insure. *American Economic Journal: Applied Economics*, 7(2):81–108.

Chami, G. F., Ahnert, S. E., Kabatereine, N. B., and Tukahebwa, E. M. (2017). Social network fragmentation and community health. *Proceedings of the National Academy of Sciences*, 114(36):E7425–E7431.

Christakis, N. A. and Fowler, J. H. (2010). Social network sensors for early detection of contagious outbreaks. *PloS One*, 5(9):e12948.

Cohen, R., Havlin, S., and Ben-Avraham, D. (2003). Efficient immunization strategies for computer networks and populations. *Physical Review Letters*, 91(24):247901.

Feld, S. L. (1991). Why your friends have more friends than you do. *American Journal of Sociology*, 96(6):1464–1477.

Gallos, L. K., Liljeros, F., Argyrakis, P., Bunde, A., and Havlin, S. (2007). Improving immunization strategies. *Physical Review E*, 75(4):045104.

Kempe, D., Kleinberg, J., and Tardos, É. (2003). Maximizing the spread of influence through a social network. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 137–146. ACM.

Kim, D. A., Hwong, A. R., Stafford, D., Hughes, D. A., O'Malley, A. J., Fowler, J. H., and Christakis, N. A. (2015). Social network targeting to maximise population behaviour change: a cluster randomised controlled trial. *The Lancet*, 386(9989):145–153.

Kumar, V., Krackhardt, D., and Feld, S. (2018). Network interventions based on inversity: Leveraging the friendship paradox in unknown network structures. Working Paper, Yale University.

Leskovec, J., Krause, A., Guestrin, C., Faloutsos, C., VanBriesen, J., and Glance, N. (2007). Cost-effective outbreak detection in networks. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 420–429. ACM.

Paluck, E. L., Shepherd, H., and Aronow, P. M. (2016). Changing climates of conflict: A social network experiment in 56 schools. *Proceedings of the National Academy of Sciences*, 113(3):566–571.