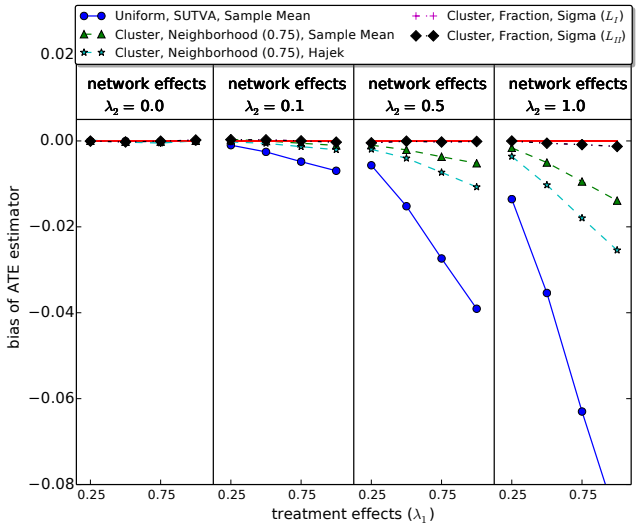
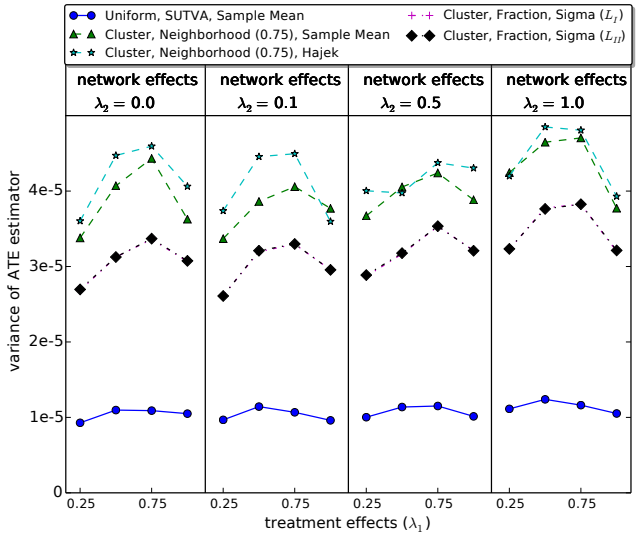


(a) The bias of different estimators



(b) The variance of different estimators

Figure 5: Behavior of different estimators with different parameters, which controls the strength of treatment effects (λ_1) and network effects (λ_2), in the simulation model. The overall percentage of nodes in treatment $\rho = 0.5$.

5.2 Real Online Experiment

In addition to the extensive simulations, we have also conducted a real online experiment at LinkedIn using the network A/B testing framework we have proposed. Specifically, we have done the following:

1. Select a country as the sub-network to experiment on.
2. Apply our randomized balanced graph partition algorithm to assign users from this country into treatment and control groups.
3. Apply different Feed algorithms to the treatment and control groups. Estimate the ATE after running the experiment for two weeks.

Country	$N_{\mathcal{S}}$ ($\times 10^6$)	$R_{\mathcal{S}}$	$\overline{d_{\mathcal{S}}}$
Brazil	19.9	0.932	41.6
United States	119.3	0.910	54.3
Netherlands	**6.1**	**0.868**	**93.0**
Chile	2.8	0.866	38.4
New Zealand	1.3	0.654	29.4

Table 4: Basic statistics about several countries. $N_{\mathcal{S}}$ is the size of the subnetwork, selected by the corresponding country; $R_{\mathcal{S}}$ is the self-containment measure defined in (16); $\overline{d_{\mathcal{S}}}$ is the average degree of the selected subnetwork defined in (17).

We note that unlike simulations, there is no ground truth for this real world experiment. The Feed team has, however, compared these two Feed algorithms globally in a uniformly randomized A/B test, and the treatment Feed algorithm was significantly better than control.

Our goal for the real world experiment is two-fold. First, we would like to compare results from different estimators in a real application setting to complement the observations from simulations. In particular, we want to compare results with and without taking into consideration of the network effects, and further, how our fraction neighborhood exposure model compares with the neighborhood exposure model. Second, as far as we know, we are the first to run a real network A/B test. We would like to establish a process for running network A/B test in practice. As we have seen how the conclusions can differ drastically in real networks compared to simulated networks, we hope this can bridge the gap and encourage more research focusing on real applications in the area of network A/B testing.

5.2.1 Country Selection

We would like to select a country that has a well self-contained LinkedIn social network. Ideally, It should be an isolated sub-network that has as few connections to the outside of the country as possible to prevent network influence to and from users outside. We use the following ratio to quantify such "self-containment" for a set $\mathcal{S}$ of users:

$$R_{\mathcal{S}} = \frac{\sum_{i,j\in\mathcal{S}} A_{i,j}}{\sum_{i\in\mathcal{S}} \sum_{j\in\mathcal{S}\cup\mathcal{S}^c} A_{i,j}}, \tag{16}$$

where $\mathcal{S}^c$ is the complement of $\mathcal{S}$ in the population network. Remark that $R_{\mathcal{S}}$ is the ratio between the edge count within the set $\mathcal{S}$ and all the edges with one end in $\mathcal{S}$.

In addition, we consider the average internal degree such that the selected subnetwork is well connected. Considering selected sub-network $\mathcal{S}$, the average internal degree is defined as

$$\overline{d_{\mathcal{S}}} = \frac{\sum_{i,j\in\mathcal{S}} A_{i,j}}{|\mathcal{S}|}. \tag{17}$$

We calculate $R_{\mathcal{S}}$ and $\overline{d_{\mathcal{S}}}$ for all countries on LinkedIn, and among the ones with top $R_{\mathcal{S}}$ ratios (shown in Table 4), we decide to pick Netherlands as it has the highest average internal degrees and a reasonably large sized network (around 6 million users).

After selecting the Netherlands as the sub-network, we applied the randomized balanced graph partition algorithm

to divide it into 600 shards and randomly picked 300 of them to receive treatment while the rest 300 to receive control. Before performing the A/B test, we have also conducted an A/A test, which is a controlled experiment where treatment is identical to control. This was to confirm that no bias was introduced during the experiment assignment process.

5.2.2 *Online Results*

We let the experiment run for two weeks before we collected data for analysis. The metric we use for evaluation is the average number of social gestures on Feed, such as "like", "comment" or "share".

We compute the ATE based on the various estimators described in Section 5.1.2. The results are shown in Table 5. We have the following observations. (i) The ATE estimators with consideration of network effects are all larger than the estimate under SUTVA. This is yet another good confirmation that there are indeed network effects presented in the A/B experiment. (ii) The choice of θ in the neighborhood exposure model matters. The sample mean estimator almost doubles when θ changes from 0.75 to 0.9. On the other hand, the Hajek estimator gives a smaller estimate when θ changes from 0.75 to 0.9, this is due to small values of π_i's when θ is large. (iv) The fraction neighborhood exposure model gives larger estimates than existing methods.

Method	ATE for social gestures
SUTVA	0.168
Neighbor. Exposure $\theta = 0.75$	0.264
Neighbor. Exposure $\theta = 0.9$	0.520
Hajek. Exposure $\theta = 0.75$	0.625
Hajek. Exposure $\theta = 0.9$	0.133
Fraction Exposure (I)	0.687
Fraction Exposure (II)	0.714

Table 5: ATE estimates from different models for the online Feed experiment.

6. CONCLUSION AND FUTURE WORK

In this paper, we study the problem of network A/B testing in real networks. We start by examining a recent A/B experiment conducted on LinkedIn without considering network structures, which motivates us to set up a framework to study both the sampling and the estimation aspects of the network A/B testing problem. To address the challenge of degree heterogeneity in real social networks, we come up with a new randomization scheme based on balanced graph partitioning, for which an efficient and distributed algorithm is proposed. Based on new sampling scheme, we propose a new method to estimate the average treatment effect (ATE) that is able to take into consideration of the level of network exposure. Extensive simulations are conducted to evaluate these methods and the results show that our new proposals can achieve both a smaller bias and a smaller variance. We have also conducted a real online experiment under the framework we have proposed and the results further validate many observations from simulations.

On the other hand, there are still many open problems in the field of network A/B testing that remain to be addressed, especially with respect to real world applications. First of all, we did not consider the influence strength between pairs of nodes, which may have significant impact on determining users exposure status; Secondly, real social networks are growing all the time, leading to rapid change of network structures, which makes network A/B testing even more challenging considering the effects of newly added edges and nodes. To further complicate the problem, many real experiments on social networks are aiming at increasing network density, making the temporal variability a real, noticeable issue. Thirdly, there are different forms of network interference to be considered. For instance, in discussion *groups*, information propagates from one user to all other users of the same group, so every group acts as a fully connected sub-network. However each user can belong to multiple groups. In this case, the graph clustering randomization can no longer split users into treatment and control under the new information propagation structure. Lastly, our focus here has been on ATE estimation, and we have not touched upon how virality works and how to preserve it in a network A/B testing setting. Given the complex structure of real social networks and the way viral information propagates, the framework proposed here may not be sufficient.

A/B testing in general is widely used and also well studied in the industry as it offers the best scientific approach to understand the causal impact of product changes on end user behavior. However, the problem of A/B testing in a social network setting is no where near solved. A lot of work still remains to be done to make it a well-understood problem in real world applications. We hope our work here can bridge some of the gaps and encourage more research in this area.

Acknowledgement

Firstly, we wish to thank our colleagues and friends who have held many enlightening discussions with us: Deepak Agarwal, Mathieu Bastian, Evion Kim, Wayne Tai Lee, Haishan Liu, Mitul Tiwari, Xiaolong Wang, and many members of the A/B testing team at LinkedIn. Secondly, we wish to thank Bee-Chung Chen, Liang Zhang, Pannagadatta Shivaswamy, Cory Hicks and Caroline Gaffney for working with us to run the network A/B test on LinkedIn Feeds. Lastly, we wish to thank the anonymous reviewers for their valuable comments and suggestions to improve the quality of the paper.

References

[1] K. Andreev and H. Racke. Balanced graph partitioning. *Theory of Computing Systems*, 39(6):929–939, 2006.

[2] P. M. Aronow and C. Samii. Estimating average causal effects under general interference. Citeseer, 2012.

[3] L. Backstrom and J. Kleinberg. Network bucket testing. In *Proceedings of the 20th international conference on World wide web*, pages 615–624. ACM, 2011.

[4] D. A. Bader, H. Meyerhenke, P. Sanders, and D. Wagner. *Graph partitioning and graph clustering*, volume 588. American Mathematical Soc., 2013.

[5] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

[6] J. Cheng, L. Adamic, P. A. Dow, J. M. Kleinberg, and J. Leskovec. Can cascades be predicted? In *Proceedings of the 23rd international conference on World wide web*, pages 925–936. International World Wide Web Conferences Steering Committee, 2014.

[7] N. Cressie and H. Whitford. How to use the two sample t-test. *Biometrical Journal*, 28(2):131–148, 1986.

[8] D. Eckles, B. Karrer, and J. Ugander. Design and analysis of experiments in networks: Reducing bias from interference. *arXiv preprint arXiv:1404.7530*, 2014.

[9] H. Gui, Y. Sun, J. Han, and G. Brova. Modeling topic diffusion in multi-relational bibliographic information networks. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 649–658. ACM, 2014.

[10] M. G. Hudgens and M. E. Halloran. Toward causal inference with interference. *Journal of the American Statistical Association*, 103(482), 2008.

[11] K. L. Judd. The law of large numbers with a continuum of iid random variables. *Journal of Economic theory*, 35(1):19–25, 1985.

[12] L. Katzir, E. Liberty, and O. Somekh. Framework and algorithms for network bucket testing. In *Proceedings of the 21st international conference on World Wide Web*, pages 1029–1036. ACM, 2012.

[13] R. Kohavi, A. Deng, B. Frasca, T. Walker, Y. Xu, and N. Pohlmann. Online controlled experiments at large scale. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1168–1176. ACM, 2013.

[14] R. Kohavi, A. Deng, R. Longbotham, and Y. Xu. Seven rules of thumb for web site experimenters. *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2014.

[15] C. F. Manski. Identification of endogenous social effects: The reflection problem. *The review of economic studies*, 60(3):531–542, 1993.

[16] C. F. Manski. Identification of treatment response with social interactions. *The Econometrics Journal*, 16(1):S1–S23, 2013.

[17] M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, pages 415–444, 2001.

[18] J. A. Middleton and P. M. Aronow. Unbiased estimation of the average treatment effect in cluster-randomized experiments. *Available at SSRN 1803849*, 2011.

[19] M. E. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, 2006.

[20] S. W. Raudenbush. Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods*, 2(2):173, 1997.

[21] M. G. Rodriguez, J. Leskovec, D. Balduzzi, and B. Schölkopf. Uncovering the structure and temporal dynamics of information propagation. *Network Science*, 2(01):26–65, 2014.

[22] P. R. Rosenbaum. Interference between units in randomized experiments. *Journal of the American Statistical Association*, 102(477), 2007.

[23] D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

[24] M. E. Sobel. What do randomized studies of housing mobility demonstrate? causal inference in the face of interference. *Journal of the American Statistical Association*, 101(476):1398–1407, 2006.

[25] L. Tang, R. Rosales, A. Singh, and D. Agarwal. Automatic ad format selection via contextual bandits. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 1587–1594. ACM, 2013.

[26] E. J. T. Tchetgen and T. J. VanderWeele. On causal inference in the presence of interference. *Statistical Methods in Medical Research*, 21(1):55–75, 2012.

[27] P. Toulis and E. Kao. Estimation of causal peer influence effects. In *Proceedings of The 30th International Conference on Machine Learning*, pages 1489–1497, 2013.

[28] J. Ugander, B. Karrer, L. Backstrom, and J. Kleinberg. Graph cluster randomization: network exposure to multiple universes. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 329–337. ACM, 2013.

[29] F. Yates. Sir ronald fisher and the design of experiments. *Biometrics*, 20(2):307–321, 1964.