**Network A/B Testing : From Sampling to Estimation (from Bortik)**

**Summary:**

The authors define the problem of conducting A/B testing in a social network in presence of interference effect as Network A/B testing problem. They propose a heuristic based randomized balanced graph partition strategy followed by random cluster sampling to assign a cluster of nodes as treatment group and control group respectively. Then they measure the ATE under different fraction neighborhood exposure conditions using linear additive models (which are shown to be generalizations of existing models). Empirically it is shown that the proposed linear model based estimators give the best bias and variance wrt any of the baselines.

**Pros:**

- This paper studies the ATE estimation error under network influence effect which is generally a common scenario for real social networks, where the response of one individual is dependent on the response of that person's immediate neighbors irrespective of whether the neighbors are in treatment or control group. Due to this interference effect the SUTVA assumption is invalid and hence simple random sampling of nodes cannot be done to get an unbiased estimate of ATE.

- The authors have shown that interference effect exists in real world networks using their experiments on the LinkedIn data set by running the two sample t-test. They have used a linear additive model to accommodate for the social interference and homophily effects as influencing parameters on the response variable of the users. Experiment results confirm that there is a significant network effect, though the effect of homophily is shown to be very less.

- The authors claim that user responses are dependent on cluster size – in presence of network effect, users in larger treatment clusters would get more social influence. By making the cluster sizes all equal, the authors remove the covariance terms between cluster sizes and user response thereby reducing bias.

- Intuitively, nodes with different degrees should behave differently in terms of being “influenced” by their neighbors. Hub nodes will influence more while low degree nodes will get less influenced which needs to be factored in the model. The authors tried to incorporate this in their LinkedIn feed experiment, but failed to provide any further insights into this, except that there could be more confounding variable influencing the ATE estimation.

- The authors have done a very good empirical evaluation of the bias/variance trade-off problem for the neighborhood exposure model for different values of theta.

- During ATE estimation, the authors have tried to model the users' exposure to network effects using fraction neighborhood exposure (linear model 1 & 2) which is a generalization of the SUTVA and the neighborhood exposure models. Thus they propose new estimator which can not only account for the neighborhood exposure but can also utilize all observations regardless to their network exposure status, thereby solving the tough problem of selecting “theta” that controls the bias/variance trade-off.

- During empirical analysis, the authors successfully show that the network structure has a significant role in the estimator quality for a wide variety of networks.

**Cons:**

- The fraction neighborhood exposure model ignores the user specific traits that can potentially influence the ATE estimate.

- The simulation of synthetic response scores could have been done using different adversarial scenarios like BFS, Quartile based etc.

- No comparison of performance numbers.

- One assumption in the motivating example: The user in the control group visits LinkedIn less frequently with the intent of discovering new friends as compared to other operations that they would perform while visiting LinkedIn. In general, at the time of framing the tests and eventually ATE estimation, we need to have strong conviction about the validity of this kind of latent assumptions being made.

- In the LinkedIn feed experiment, the authors have assumed that all the users show similar periods and types of activity – how realistic is this assumption for the experimental data set ?

- In the LinkedIn feed experiment – why is homophily effect nearly insignificant ? Intuitively this should have been higher. Is part of the homophily effect being absorbed into network effect as well ? (given that one implies the other) It seems like the network effect (topological similarity) is able to hide\overshadow the homophily effect (behavioral/contextual similarity).

- The authors did not mention anything about what is the optimal number of clusters ? In their framework, would they prefer small number of large clusters or large number of small clusters ? [2500 nodes = 500 (nodes) \* 5 (clusters) = 25 (nodes) \* 100 clusters]

- Group level v/s Unit level interference – which one is being targeted by the authors is not very clear. Based on the cluster creation strategy, it appears that the authors try to minimize unit level interference (between node interference) assuming that nodes in different groups do not interact/influence each other. Hence the partitioning strategy tries to create equal-sized clusters with highly inter-connected nodes within the same cluster. However since the partitioning strategy is random, the homophily bias in different clusters would be different based on how much node diversity is present in that cluster and this might increase the overall variance in estimation.

- The goal of the partitioning strategy is to remove information diffusion. It is unclear how well the proposed strategy does to that effect due to lack of any theoretical or empirical insights. A heat diffusion based (or maybe entropy based) clustering strategy would have been a better clustering alternative here.

- The clustering strategy has an underlying assumption that partitioning would be hard partitioning and hence the nodes that are part of different clusters will not have any interference effect on nodes of different clusters. But this ignores the commonplace real world scenarios where nodes are part of overlapping communities and play multiple different roles as part of different communities. Maybe such nodes should be ignored from treatment exposure to minimize the network effect or a novel sampling strategy needs to be developed to accommodate the significance of Role\Community membership of nodes while deciding whether to expose it to treatment group or not.

- The experiments could have been run on networks with different generative models (network structure) and then the robustness of the newly proposed estimators could be analyzed from that.