Cross-Unit Spillovers in A/B testing: Empirical Evidence from Ads

Ronak Jain Amazon.com
Stefan Hut Amazon.com
Mahnaz Islam Amazon.com
Yao Pan Amazon.com

Code 2023: Extended Abstract

Randomized Control Trials (RCTs) are widely used across Amazon to causally estimate impacts of proposed feature changes, in order to make data-driven launch decisions. A key element of experimental design is the level of randomization, and the choice often relies on the cross-unit interaction structure. For instance, in the context of advertiser experiments, a treatment may affect the outcome of control advertisers who compete with treated advertisers in ad auctions. Such spillover effects lead to biased treatment effect estimates under simple advertiser randomization due to the violation of the required Stable Unit Treatment Value Assumption (SUTVA). While grouping similar advertisers into clusters and performing cluster randomization potentially mitigates this bias, this comes at the cost of reduced statistical power. Quantifying the magnitude of intra-cluster spillovers is critical to evaluating the trade-offs between simple unit randomization and cluster randomization, and make informed decisions on the level of randomization when spillover effects are a concern.

This paper proposes an empirical approach to estimate spillover effects within advertiser clusters (i.e. a group of advertisers who often interact with each other), using a historical experiment randomized on advertiser-id, joined with advertiser cluster data. The general idea is to investigate whether advertiser outcomes vary by the fraction of advertisers being treated in their clusters, i.e. treatment intensity. Specifically, we compute each advertiser's treatment intensity by overlaying clusters which were generated prior to the experiment. We then assess whether an individual advertiser's outcomes are impacted by their within-cluster peer's treatment status, to quantify the degree of spillovers between advertisers that are grouped together in the same cluster. Although we focus on the use case of advertiser-facing experiments, the approach can more broadly be used to assess the importance of spillovers in other experimental settings where we may expect cross-unit interference (e.g. seller, regional, product-level experiments).

We first define and calculate within-cluster treatment intensity. For advertiser i in cluster j, the treatment intensity is defined as the fraction of advertisers being treated in the cluster, excluding advertiser i:

$$T_{ij}^{intensity} = \frac{\sum_{a \neq i} T_{aj}}{N_j - 1} \ , \tag{1}$$

where T_{ij} is the treatment status of advertiser i in cluster j that takes the value of 1 if treated and 0 otherwise. N_j is the size of cluster j. Empirically, treatment intensity for each advertiser can be calculated by overlaying the clusters which were generated before the experiment on top of advertisers' treatment status for any advertiser-randomized experiment.

We then proceed to investigate whether advertiser outcomes vary by treatment intensity conditional on their own treatment status. We focus on the following simplified model in our empirical exercise.

$$Y_{ij} = \alpha + \beta T_{ij} + f(T_{ij}^{intensity}) + g(N_j) + \theta Y_{ij}^0 + \epsilon_{ij} \ , \tag{2}$$

where Y_{ij} is the outcome of interest (i.e. key launch criterion metric) for advertiser i in cluster j, T_{ij} is a binary variable that is 1 if advertiser i is treated, and $f(\cdot)$ describes the within-cluster spillover effects as a function of treatment intensity, $T_{ij}^{intensity}$. We flexibly control for the size of the cluster, N_j, via $g(\cdot)$ in our estimation to ensure $f(\cdot)$ only captures how treatment effect varies by treatment intensity for clusters of the same size. We include baseline value of the outcome, Y_{ij}^0, to increase the precision of our estimates. We cluster standard errors at the cluster level to allow for within-cluster correlations in the error term.

We rely on the following assumptions in the spillover effect estimation.

1. **There are no inter-cluster spillovers.** We assume local/partial interference and allow the potential outcome of an advertiser in cluster j to depend on treatment assignments of other advertisers $i \in j$ in the same cluster but not on advertisers outside of one's cluster i.e. $i \notin j$. In other words, we assume that SUTVA (Stable Unit Treatment Value Assumption) holds across different clusters of advertisers. While we cannot empirically test for the validity of this assumption in this exercise, we evaluate the robustness of our estimates to different clustering parameters (and therefore number of clusters) used. This is also a typical assumption made in the analysis of spillovers when using two-stage cluster randomized experiments (Baird et al., 2018).
2. **Within-cluster spillovers can be captured by fraction of other advertisers treated in a cluster.** We assume that the spillover effect can be captured by the "exposure to treatment," i.e. the fraction of other advertisers treated in the cluster (which we henceforth refer to as the "treatment intensity"). This is standard practice for evaluating the effect of spillovers (Athey and Imbens, 2017, Saveski et al., 2017, Baird, et al. 2018).
3. **Treatment-intensity is exogenous conditional on cluster size.** As the clusters were constructed independently and prior to treatment assignment, and the randomization was done at the advertiser level for these historical experiments, the fraction of advertisers treated in each cluster should be random as well conditional on cluster size. The treatment intensity is naturally likely to vary if the size of the cluster, N_j, is relatively small; for larger clusters, we may however expect treatment intensity to converge to the treatment allocation share (e.g. 0.5 for 50%/50% T/C allocation) within a cluster given the law of large numbers. We show a scatter plot of treatment intensity with cluster size below, and we test that treatment intensity is not correlated with the pre-period values of the outcomes conditional on cluster size as to shed light on this assumption.
4. **There is sufficient variation in treatment intensity across clusters.** The proposed identification strategy for spillover effect relies on exploiting variation in treatment intensity across clusters. This is a testable assumption, and we check it by plotting the density of treatment intensity.

We apply the approach to an advertising experiment run at Amazon, which introduced budget recommendations for new advertiser campaigns. The new feature was tested in an 5-week (November 2022 to mid-December 2022) advertiser randomized experiment with 50% of advertisers in treatment and the rest in control. We use advertiser clusters generated for the period just preceding the experiment start date (i.e. October to November 2022). Clusters are generated from advertiser/query bipartite graphs where the edges are the sum of cost per click (CPC) between an advertiser/keyword pair (i.e. linking advertisers who often appear together through common keywords). Spectral co-clustering is then used to generate clusters that minimize the number of edges between clusters. For our main results, we use the parameter that groups advertisers into 10k clusters. Figure 1 below shows the scatter plot of treatment intensity with cluster size. We see that the treatment intensity variation is symmetric around 0.5 with most of the variation coming from smaller clusters. As we would expect, given randomization and the law of large numbers, large clusters have treatment intensity close to 0.5. This in turn means that the identification of spillovers will mostly leverage the variation in treatment intensity across smaller clusters, leading to concerns about external validity and the estimated spillovers may not necessarily generalize to larger clusters as we lose identifying variation in treatment intensity.

Table 1 presents the results from the estimation of equation (2) with a linear functional form for $f(\cdot)$ and $g(\cdot)$. The results indicate substantial negative spillovers within clusters as the coefficient of treatment intensity (which measures the effect of increasing treatment intensity to 1, i.e. 100%) is statistically significant for all outcome metrics except one. For instance, the first metric we examine would fall by 99.3 if there was a 10% increase in treatment intensity within a cluster and this effect is in the opposite direction to the direct treatment effect. In the extreme case where an advertiser and all other advertisers in the cluster were treated (i.e. under cluster randomization), we would expect the total treatment to have impacted this metric by 945.99 (=47.26 - 993.25), which is 13% of the control group mean and 20 times the size of the original A/B test estimate.

We conduct several robustness checks to confirm the presence of sizable spillover effects. First, we conduct a placebo test regressing pre-period outcomes on treatment intensity. We find that treatment intensity does not predict any of the pre-period outcomes, supporting our assumption of exogeneity conditional on cluster size. Second, we check for robustness of the estimate to a number of alternative specifications: (a) allowing for spillovers to vary by treatment status, (b) controlling for all pre-period outcomes, (c) interacting treatment and treatment intensity with preperiod value of the outcome, (d) controlling more flexibly for cluster size via quartiles, (e) estimating treatment intensity dropping very large (top 25 percentile) clusters. Third, we conduct a permutation test, randomly shuffling the treatment status of all advertisers while holding their clusters constant, and find further evidence confirming that our observed spillovers are significant and large relative to what is expected if there was no treatment effect (an A/A test setting) (Figure A3).

Lastly, we relax the linear assumption for treatment intensity, i.e. $f(\cdot)$, and test for alternative functional forms. We start by including treatment intensity deciles instead to motivate our functional form choice. We find suggestive evidence for non-linearity in spillovers. As effect of spillovers seems to vary at the extremes of the distribution, we use tercile dummies in regression to approximate this pattern. Again, we detect strong spillovers (Appendix Table A9). The

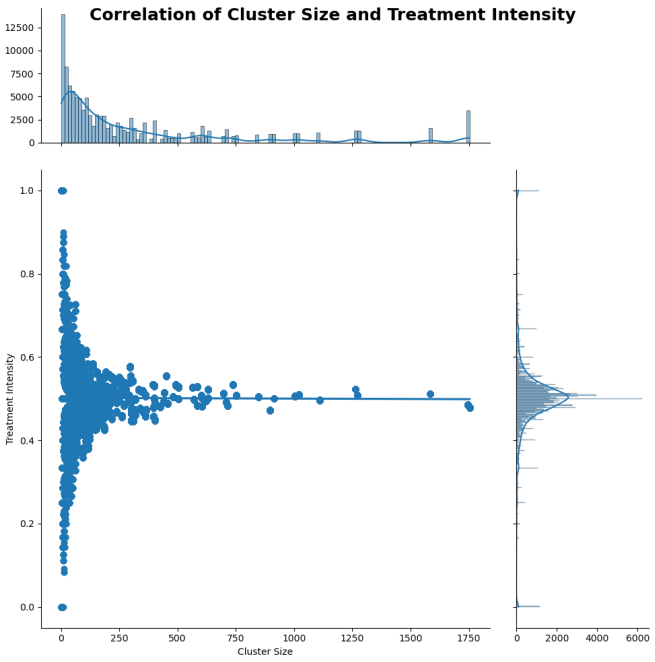

Figure 1: Scatter plot of Treatment Intensity and Cluster Size

Notes: This figure plots the correlation between treatment intensity and cluster size. The chart above plots the histogram of cluster size and the chart on the right shows the density of treatment intensity.

Table 1: Estimation of Spillover Effects

	Metric 1 (1)	Metric 2 (2)	Metric 3 (3)	Metric 4 (4)	Metric 5 (5)	Metric 6 (6)	Metric 7 (7)	Metric 8 (8)	Metric 9 (9)
treat	47.26	-3.65**	-2.08	-0.56	-2.79	-12.46	-2.02	-0.11	-0.1
s.e.	(64.14)	(1.81)	(1.94)	(1.27)	(3.09)	(11.91)	(2.26)	(0.11)	(0.1)
treat_intensity	-993.25***	-24.68***	-26.57***	-16.72***	-42.92***	-47.33	-27.92***	-0.9	-0.81
s.e.	(327.78)	(8.48)	(6.6)	(4.62)	(10.96)	(48.42)	(10.59)	(0.55)	(0.52)
R-squared	0.74	0.78	0.75	0.75	0.76	0.01	0.68	0.54	0.53
R-squared Adj.	0.74	0.78	0.75	0.75	0.76	0.01	0.68	0.54	0.53
N	111,305	111,305	111,305	111,305	111,305	111,305	111,305	111,305	111,305
No. clusters	3,197	3,197	3,197	3,197	3,197	3,197	3,197	3,197	3,197
Dep. Variable Mean	6,937	287	133	98	232	84	373	11	10
Control Group Mean	7,093	292	135	99	235	90	375	11	10
Flywheel Effect (treat + treat_intensity)	-945.99	-28.33	-28.65	-17.28	-45.71	-59.79	-29.94	-1.01	-0.91
P-value: 'treat + treat_intensity=0'	0.01	0	0	0	0	0.23	0.01	0.07	0.08

Notes: This table shows the estimate from equation (1) and controls for the pre-period value of the dependent variable. Standard errors are clustered at the cluster level and reported below coefficients.

magnitude of the effect is smaller than in linear specification. In future iterations, we intend to establish the functional form for spillovers in a a more rigorous and data-driven way.

Taken together, we find evidence of substantial spillovers between treated and control advertisers in the context of an advertiser experiment run at Amazon. These spillovers can be large enough to change an experiment's launch decision. These findings highlight the importance of accounting for spillovers in A/B testing. In particular, the methodology presented in this paper can be used to test for the presence of spillovers in ongoing experiments, to determine whether individual randomized experiments should be re-run as cluster-randomized based on the magnitude of spillover effects.

References

Athey, S. and Imbens, G. W. (2017). The econometrics of randomized experiments. In *Handbook of economic field experiments*, volume 1, pages 73–140. Elsevier.

Baird, S., Bohren, J. A., McIntosh, C., and Özler, B. (2018). Optimal design of experiments in the presence of interference. *Review of Economics and Statistics*, 100(5):844–860.

Nassiri, s. (2022). *Trustworthy Experiments in the Presence of Interference*. Amazon internal paper, submitted to AMLC.

OffersX (2018). *Accounting for Spillovers in ASIN Randomized Experiments: An Econometric and ML Solution*. Amazon internal paper, submitted to AMLC 2018.

OffersX (2022). *Evaluating the Validity of SUTVA in Amazon: The Case of OfferX Experiments.* Amazon internal paper, submitted to AMLC 2022.

OffersX (2023). *Addressing Spillovers in OffersX Experiments Using Parent-Child ASINs*. Amazon internal paper, submitted to AMLC 2023.

Saveski, M., Pouget-Abadie, J., Saint-Jacques, G., Duan, W., Ghosh, S., Xu, Y., and Airoldi, E. M. (2017). Detecting network effects: Randomizing over randomized experiments. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1027–1035.

Weblab (2021). *Flywheel - Clustering Methodology*. Amazon internal paper.

A Appendix: Additional Figures and Tables

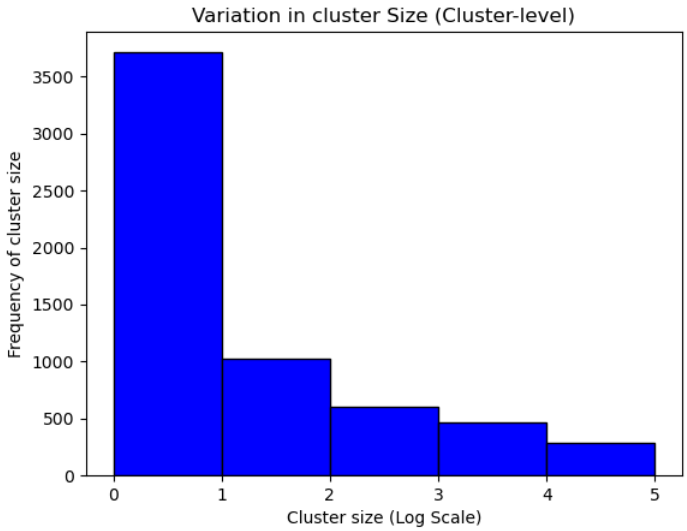

Figure A.1: Variation in Cluster Size

Notes: This figure shows the histogram of cluster size on log scale (i.e. the frequency of clusters between 0-1 on the log scale corresponds to clusters between size 1-10 and those between 1-2 represent clusters with sizes between 10-100).

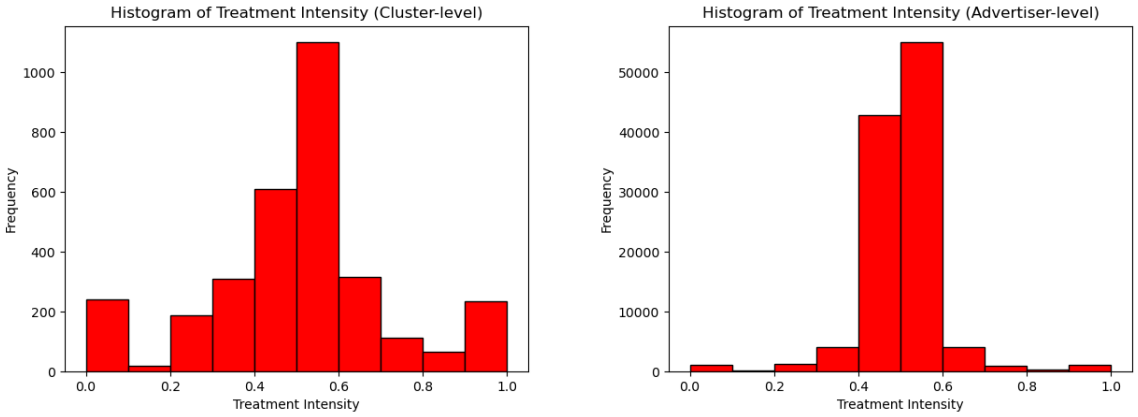

Figure A.2: Histogram of Treatment Intensity

Notes: This figure plots the histogram of treatment intensity after overlaying the advertisers in the experimental data set with clusters. The left panel plots the density at the cluster-level and the right panel plots the histogram at the advertiser-level.

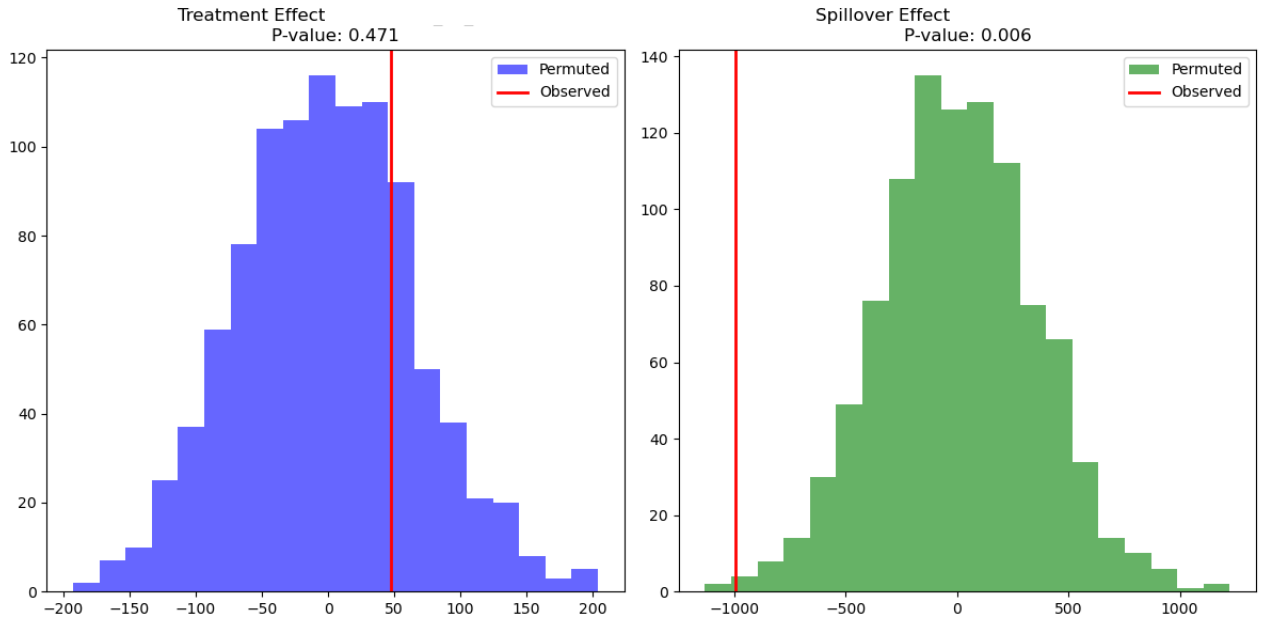

Figure A.3: Permutation Test