SPILLOVER EFFECTS IN ONLINE FIELD EXPERIMENTS: OPPORTUNITIES AND CHALLENGES

TECHNICAL REPORT

Organizers and speakers for WINE'2021 Experiment Design Tutorial

ABSTRACT

This tutorial consists of six talks by researchers from a wide range of disciplines such as computer science, economics, statistics, and operation research. Here we provide an overview of one main issue in online field experiments: spillover effects. We compile a reading list for researchers in both academia and industry who are interested in this topic.

In this overview, we first introduce the spillover effects and discuss its impacts on field experiments. We next discuss how to addresses this issue by improving the experimental design and post-experiment analysis.

1 Overview

1.1 An introduction to the spillover effects in field experiments

Field experiments typically aim to quantify how an intervention (e.g., a new policy [1]) affects certain outcomes of all the population. With the popularity of online communities and market places, there has been an outbreak of online field experiments [2]. However, in many online and offline experiments, interference (or spillover effects) exists [3, 4]: a user's outcome may be affected by the treatment assignments of other subjects; for within-subject experiments, the previous treatment assignment the subject received may affect the outcome at a later stage.

Imagine that a "powerful" policymaker wants to evaluate the impact of vaccines on reducing COVID-19 cases and they can "impose" people to get vaccinated. The experiment is to randomly assign some people to get vaccinated while others do not. There exists a strong spillover effect: people who get vaccinated also reduce the risk of their family and friends, while people who do not get vaccinated increase others' risk level. More generally, if a large proportion of the population is assigned to receive vaccines, the observations in the control group will have a lower risk; similarly, if a large proportion of the population is assigned not to receive vaccines, the observations in the treatment group will have a higher risk. In this case, a comparison between the treatment and control group would underestimate the impact of the policy whereby everyone should receive vaccines. In the online setting, many controlled experiments face similar issues. For example, when a user is promoted with a discount coupon, the user may share this coupon with their friends who are assigned to other groups, which may decrease the purchase probability of their friends in the short term. This spillover effect may lead to an overestimate of the average treatment effect when we compare the difference in the purchase probability between the treatment and control groups. Generally speaking, without addressing spillover effects, we may over- or under-estimate the "global treatment effect" which describes the difference between the scenarios where everyone is treated versus non-treated.

With the presence of the interference or spillover effect, the conventional way of randomizing samples may be problematic [5, 6]. Intuitively, the spillover effect means the outcome of an observation is not only affected by their own treatment assignment, but also by the treatment assignments of other observations. The existence of spillover effects violates the stable unit treatment value assumption (SUTVA). Formally, SUTVA is defined as [7]

1. The potential outcomes for any unit do not vary with treatments assigned to other units;
2. For each unit, there are no different forms or versions of each unit level, which lead to different potential outcomes.

The presence of spillover effects violates the first condition – the potential outcome is affected by treatments assigned to other units [5, 8].

Here are more concrete examples of spillover effects:

- *Social contagion*. Similar to contagion, social contagion means that a person's behavior may influence others to do likewise [9]. Therefore, if one person is assigned a treatment, their family, friends, or acquaintances may also indirectly receive this treatment [10].
- *Displacement*. For example, increasing the exposure of an ad on a website may displace other ads with similar topics. Suppressing the crime in location A may increase the crime in location B [11].
- *Carryover effects*. In within-subject analysis, the treatment received in the last stage may affect the potential outcome in the current stage [12]. For example, if a person receives a promotion before, receiving it once again would probably not have the same effect as the first time of receipt.

This tutorial focuses on how to address such spillover effects when we implement experiments or analyze the experimental data.[1]

1.2 Mitigating spillover effects by experimental designs

There are two main approaches to address spillover effects – improving the experimental design or improving analysis. Here we first discuss several types of experimental designs (i.e., the way of randomized treatment assignments).

Cluster randomization

Often spillover effects exist within a larger unit. For example, considering all elementary school students as the population, the treatment of deworming a student may only spill over to their classmates or schoolmates. Therefore, we can randomly assign treatments on the cluster level. In a school-level cluster randomization, all students in the same school will receive the same treatment assignments (deworm or not). By comparing students who are treated or not, we can measure the effect size with consideration of within-school spillover effects [13]. This experimental design is referred to as cluster randomization, which was first proposed in medical science but later rapidly adopted in many other fields [14, 15].

In the online setting, many platforms consider a city (county) as a "cluster." For example, Uber conducted an experiment to test the effect of the introduction of the tipping mechanism. They chose 110 cities in the US and Canada as clusters and performed cluster level randomization. To improve the precision of estimation, they first match group cities in pairs according to their similarity and then assign one in a pair to treatment and the other to control [16].

Graph cluster randomization

This is a special case of cluster randomization in social networks. As mentioned previously, social contagion is widespread in various behaviors, indicating that the treatment of an individual likely affects the outcome of their network neighbors. In social networks, it is challenging to split a whole network into completely separate clusters. A way to address this issue is by implementing community detection algorithms (in graph theory, graph clustering) [17]. That is, a social network is split into relatively separate clusters. With this approach, most edges belong to the same cluster. We thus assign treatments such that all units in a cluster receive the same treatment. In this way, the spillover effects within each graph cluster (a.k.a. community) are taken into account. This method was first proposed by [18], and many improved versions have appeared [6, 19, 20, 21].

For example, Meta (formerly Facebook) reports that they actively employ graph cluster randomization on social networks [21]. A key challenge is the choice of graph cluster algorithms[22]. For example, [23] find that imbalanced graph clusters are typically superior in terms of the bias-variance tradeoff for graph cluster randomization.

Meanwhile, Linkedin Research pointed out several challenges of graph cluster randomization [24]. They find that if the network is so dense that a community detection algorithm cannot give satisfactory partitioning, graph cluster level randomization may fail to consider too many spillover effects across communities.

Bipartite experiments

Many platforms have two sided markets between consumers and suppliers. Examples include customers and deliverymen on food delivery apps or hosts and guests on lodging apps. The platforms may face a challenge of deciding whether a randomization should be implemented on the consumer or supplier level. Moreover, a treatment can be assigned at the consumer-supplier pair level, which is referred to as bipartite experiments [25, 26, 27].

[1] There are studies in parallel aimed to empirically quantify spillover effects, which is beyond the scope of our discussion.

There has been growing literature on designing and analyzing bipartite experiments, especially for two-sided markets. For example, [28] provides an approach to incorporate cluster randomization in bipartite experiments. Specifically, they use observational data on Airbnb to create clusters of similar listing, and perform randomization on the cluster level. [29] discusses when the platform should perform randomization on the customer end or the supplier end.

Switchback randomization

Sometimes the whole population cannot be satisfactorily split into many separate clusters. One solution is to implement within-subject experiments and randomize on the time level. For example, on a mobile app, the platform can randomly switch between two versions of the UI every hour or every day. Then the platform can compare user behavior between the hours (or days) with different versions displayed [30, 12].

For instance, DoorDash [31] uses this switchback experiment to test the effect of a pricing algorithm. However, large spillover effects (network interference), along with customer experience issues may occur. If two users who are close friends are assigned to different experimental groups, they may turn out to use the one with lower prices. Moreover, price discrimination produced by user-level randomization may hurt user experiences. These are the reasons why we conjecture DoorDash employs switchback experiments to detect the effect of a pricing algorithm. Specifically, the experiment setting switches back and forth every 30 minutes.

In addition, Kuaishou, a popular video-sharing mobile app, has its two-sided market between users and hosts. However, as reported by [32], their platform has strong spillover effects between hosts such that cluster randomization is not feasible. Therefore, they also use switchback experiments.

However, one challenge is the existence of the "carryover" effect. Previous experience of another version may affect the outcome in the current time period. One solution to this issue is by first implementing another experiment to detect the decay rate of carryover effects [12]. After this experiment, the researcher determines the time window for the switchback accordingly and implemented the switchback experiment.

1.3 Mitigating spillover effects by post-experiment analysis

After experiments are conducted, we can further mitigate the spillover effects by post-experiment analysis.

Bias-variance trade-off

In general, all solutions may involve a trade-off between bias and variance. Bias comes from the disregard of a great proportion of spillover effects. In the student deworming example, if there are many spillover effects across schools (e.g., many cross-school students play together off-school), we fail to consider a large proportion of spillover effects, and thus our estimation may be biased.

One solution is to have larger clusters (e.g., all schools in the same city receive the same treatment). In this way, the bias would be reduced as more spillover effects are taken into account. However, the tradeoff is the variance. As the cluster is larger, the number of clusters decreases, and our statistical power by comparing the treatment and control also decreases. Therefore, choosing a proper size for clusters is crucial for the trade-off between bias and variance [18].

Switchback experiments face the same trade-off. As the time window for each experimental period is wider, the bias is small as wider windows with potential carryover effects are taken into account. However, the variance increases with the number of available experimental periods [12].

Common assumptions used for considering spillover effects

With spillover effects, SUTVA does not hold anymore. Therefore, we need weaker assumptions when we apply analysis:

- *Direct and indirect effects*. One assumption is that the treatment effect is a summation of the direct effect and the indirect effect. For example, the potential outcome of user i can be written as $y_i(z_i, w_i)$ where z_i denotes the treatment assignment of the user i and w_i denotes if there is any other user who may "spill over" their treatment to user i. Then we can estimate the direct effect and indirect effect separately. However, the drawback of this assumption is that we assume all indirect effects are homogeneous - it is considered an indirect effect regardless of the number of friends of the user i who get treated [8].

- *Stable unit neighborhood treatment value assumption (SUNTVA)*. In the setting of social networks, we should define the network neighborhood that would spill over. Ideally, with the presence of social influence, everyone in the world would have indirectly affected the user's outcome. However, this would indicate that we have to treat the whole social network as a giant cluster and we would not be able to separate them into relatively small clusters where we can assign independent cluster-level treatments. Therefore, we need to define up to

how many hops there is a spillover effect (typically it is just one hop). This is sometimes called Stable Unit Neighborhood Treatment Value Assignment [18, 33].

- *Exposure conditions*. A further extension of the direct/indirect effects is the definition of exposure conditions. Practitioners should define a mapping where the input is the treatment assignments of a user and others who may spill over to the user and the output is a discrete number that labels the user's exposure condition. An example is that we first define $2 \times 2 = 4$ exposure conditions, where the exposure condition is dependent on whether the user is treated and whether the number of treated neighbors in the social network is greater than k. There are also more complex definitions of exposure conditions [18, 8]. In social network experiments, people typically need to first define SUNTVA (i.e., only k-hop network neighbors would matter), and then define exposure conditions.
- *Limited period of carryover effects*. In switchback experiments where we randomize treatments using a period of time as the unit, we need to quantify the strength of carryover effects (i.e., how the treatment in the last time step affects the outcome time step). Ideally, after a period of time, the carryover effects should decay to negligibly small. There are studies that aim to identify the strength and decay rate of carryover effects and then assume the carryover effects only appear after a certain amount of time [12].

Using machine learning for complex spillover effects (interference)

Two main challenges remain for using assumptions for analyzing experimental data. First, some assumptions may be simplistic. For example, simply aggregating all spillover effects as "indirect effects" may lose the granularity of heterogeneous indirect effects (e.g., how many others are treated and who they are may influence the strength of spillovers). Second, strict assumptions may increase the chance of functional form misspecification. That is, the estimation of the treatment effect may be heavily biased if the assumption does not fit this specific experimental setting.

There is growing literature in utilizing machine learning for addressing network interference, which proposes flexible model specification. Here are two examples. First, [34] proposes regression adjustment methods–they convert functions of the treatment assignment vector (the treatment of all subjects) into features in machine learning, and thus addressing spillover effects turns into standard feature engineering in machine learning. [35] proposes another approach for addressing spillover effects in networks—they first use network motifs to generate treatment assignment vectors, and then propose the tree-based algorithm that automatically generates exposure conditions.

1.4 Future directions

In the future, there are a lot of things to explore for addressing spillover effects (interference) in online field experiments. First, while many studies discuss specific experimental design, fewer have investigated how to choose different designs – for example, how to decide on cluster randomization versus switchback randomization. Second, as mentioned in the machine learning section, contrary to conventional approaches where experimenters limit the relationships between treatment assignment vector for all subjects and the outcome by certain model assumptions, it is interesting to investigate how to design algorithms to automatically detect such relationships. Finally, with an increasing number of experiments performed by platforms, it is interesting to study how to collectively examine multiple experiments and obtain insights into experimental design and analysis.

2 Selected reading list for tutorial

[Introduction to online field experiments]

Chen, Yan, and Joseph Konstan. "Online field experiments: A selective survey of methods." Journal of the Economic Science Association 1.1 (2015): 29-42.

This paper presents an overview of the design and analysis for online field experiments. It covers representative studies from both economics and computer science.

[Empirical evidence for interference]

Muchnik, Lev, Sinan Aral, and Sean J. Taylor. "Social influence bias: A randomized experiment." Science 341.6146 (2013): 647-651.

This paper represents the literature on identifying empirical evidence of interference (social influence) between units.

[Cluster randomization]

Bharat K Chandar, Ali Hortaçsu, John A List, Ian Muir, and Jeffrey M Wooldridge. Design and analysis of cluster-randomized field experiments in panel data settings. National Bureau of Economic Research, 2019.

This paper examines best practices for estimating unit-level treatment effects in cluster-randomized field experiments and uses insights from their analysis to evaluate the effect of a nationwide tipping field experiment on Uber.

[Graph cluster randomization]

Ugander, Johan, et al. "Graph cluster randomization: Network exposure to multiple universes." Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2013.

This paper proposed a novel approach to address spillover effects on large-scale social networks–graph cluster randomization. Here clusters are the graph clusters (communities) on networks

Pouget-Abadie, Jean, et al. "Testing for arbitrary interference on experimentation platforms." Biometrika 106.4 (2019): 929-940.

This paper proposes an approach to compare graph cluster randomization and individual level assignment. Graph cluster randomization is shown to largely reduce bias in estimating treatment effects.

[Bipartite experiments]

Holtz, David, et al. "Reducing interference bias in online marketplace pricing experiments." Available at SSRN 3583836 (2020).

This paper aims to address the interference on two-sided markets, which is a good example for bipartite experiments.

[Switchback randomization]

Bojinov, Iavor, David Simchi-Levi, and Jinglong Zhao. "Design and analysis of switchback experiments." Available at SSRN 3684168 (2020).

This paper derives the optimal design of switchback experiments under the assumption of limited period of carryover effects.

[Relaxing SUTVA]

Peter M Aronow and Cyrus Samii. "Estimating average causal effects under general interference, with application to a social network experiment." Ann Appl Stat (2017).

This paper presents a randomization-based framework that relaxes SUTVA under general interference between units.

[Incorporating machine learning]

Alex Chin. Regression adjustments for estimating the global treatment effect in experiments with interference. Journal of Causal Inference, 7(2), 2019.

This paper proposes a novel approach that allows researchers to convert the problem of detecting interference into a feature engineering problem.

[Applications]

Ya Xu, Nanyu Chen, Addrian Fernandez, Omar Sinno, and Anmol Bhasin. "From infrastructure to culture: A/B testing challenges in large scale social networks." In KDD(2015). 2227–2236.

From an industry perspective, this paper discusses challenges, best practices and pitfalls in evaluating results of online controlled experiments.

References

[1] Cassandra Handan-Nader, Daniel E Ho, and Becky Elias. Feasible policy evaluation by design: A randomized synthetic stepped-wedge trial of mandated disclosure in king county. *Evaluation Review*, 44(1):3–50, 2020.

[2] Yan Chen and Joseph Konstan. Online field experiments: A selective survey of methods. *Journal of the Economic Science Association*, 1(1):29–42, 2015.

[3] David Roxbee Cox. *Planning of experiments.* Wiley, 1958.

[4] Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

[5] Ya Xu, Nanyu Chen, Addrian Fernandez, Omar Sinno, and Anmol Bhasin. From infrastructure to culture: A/B testing challenges in large scale social networks. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2227–2236, 2015.

[6] Jean Pouget-Abadie, Guillaume Saint-Jacques, Martin Saveski, Weitao Duan, S Ghosh, Y Xu, and Edoardo M Airoldi. Testing for arbitrary interference on experimentation platforms. *Biometrika*, 106(4):929–940, 2019.

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

[8] Peter M Aronow and Cyrus Samii. Estimating average causal effects under general interference, with application to a social network experiment. *The Annals of Applied Statistics*, 11(4):1912–1947, 2017.

[9] Sinan Aral and Dylan Walker. Identifying influential and susceptible members of social networks. *Science*, 337(6092):337–341, 2012.

[10] Adam DI Kramer, Jamie E Guillory, and Jeffrey T Hancock. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111(24):8788–8790, 2014.

[11] David Weisburd and Cody W Telep. Hot spots policing: What we know and what we need to know. *Journal of Contemporary Criminal Justice*, 30(2):200–220, 2014.

[12] Iavor Bojinov, David Simchi-Levi, and Jinglong Zhao. Design and analysis of switchback experiments. *Available at SSRN 3684168*, 2020.

[13] Guillaume Basse and Avi Feller. Analyzing two-stage experiments in the presence of interference. *Journal of the American Statistical Association*, 113(521):41–55, 2018.

[14] J Martin Bland. Cluster randomised trials in the medical literature: Two bibliometric surveys. *BMC Medical Research Methodology*, 4(1):1–6, 2004.

[15] David M Murray, Sherri P Varnell, and Jonathan L Blitstein. Design and analysis of group-randomized trials: A review of recent methodological developments. *American Journal of Public Health*, 94(3):423–432, 2004.

[16] Bharat K Chandar, Ali Hortaçsu, John A List, Ian Muir, and Jeffrey M Wooldridge. Design and analysis of cluster-randomized field experiments in panel data settings. *National Bureau of Economic Research*, 2019.

[17] Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75–174, 2010.

[18] Johan Ugander, Brian Karrer, Lars Backstrom, and Jon 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, 2013.

[19] Dean Eckles, Brian Karrer, and Johan Ugander. Design and analysis of experiments in networks: Reducing bias from interference. *Journal of Causal Inference*, 5(1), 2017.

[20] Johan Ugander and Hao Yin. Randomized graph cluster randomization. *arXiv preprint arXiv:2009.02297*, 2020.

[21] Brian Karrer, Liang Shi, Monica Bhole, Matt Goldman, Tyrone Palmer, Charlie Gelman, Mikael Konutgan, and Feng Sun. Network experimentation at scale. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 3106–3116, 2021.

[22] Joel Nishimura and Johan Ugander. Restreaming graph partitioning: Simple versatile algorithms for advanced balancing. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1106–1114, 2013.

[23] Satu Elisa Schaeffer. Graph clustering. *Computer Science Review*, 1(1):27–64, 2007.

[24] Preetam Nandy, Kinjal Basu, Shaunak Chatterjee, and Ye Tu. A/B testing in dense large-scale networks: Design and inference. *Advances in Neural Information Processing Systems*, 33, 2020.

[25] Nick Doudchenko, Minzhengxiong Zhang, Evgeni Drynkin, Edoardo M Airoldi, Vahab Mirrokni, and Jean Pouget-Abadie. Causal inference with bipartite designs. *Available at SSRN 3757188*, 2020.

[26] Jean Pouget-Abadie, Kevin Aydin, Warren Schudy, Kay Brodersen, and Vahab Mirrokni. Variance reduction in bipartite experiments through correlation clustering. *Advances in Neural Information Processing Systems*, 32, 2019.

[27] Corwin M Zigler and Georgia Papadogeorgou. Bipartite causal inference with interference. *Statistical Science: A Review Journal of the Institute of Mathematical Statistics*, 36(1):109, 2021.

[28] David Holtz, Ruben Lobel, Inessa Liskovich, and Sinan Aral. Reducing interference bias in online marketplace pricing experiments. *Available at SSRN 3583836*, 2020.

[29] Ramesh Johari, Hannah Li, Inessa Liskovich, and Gabriel Weintraub. Experimental design in two-sided platforms: An analysis of bias. *arXiv preprint arXiv:2002.05670*, 2020.

[30] Iavor Bojinov and Neil Shephard. Time series experiments and causal estimands: Exact randomization tests and trading. *Journal of the American Statistical Association*, 114(528):1665–1682, 2019.

[31] David Kastelman and Raghav Ramesh. Switchback tests and randomized experimentation under network effects at doordash. *https://medium.com/@DoorDash/switchback-tests-and-randomized-experimentation-under-network-effects-at-doordash-f1d938ab7c2a*, 2018.

[32] Yaran Jin. Kuaishou causal inference and experimental design. *https://mp.weixin.qq.com/s/svVl1eiVUH6rOYG3p2YiGg*, 2021.

[33] Michael P Leung. Treatment and spillover effects under network interference. *Review of Economics and Statistics*, 102(2):368–380, 2020.

[34] Alex Chin. Regression adjustments for estimating the global treatment effect in experiments with interference. *Journal of Causal Inference*, 7(2), 2019.

[35] Yuan Yuan, Kristen Altenburger, and Farshad Kooti. Causal network motifs: Identifying heterogeneous spillover effects in a/b tests. In *Proceedings of the Web Conference 2021*, pages 3359–3370, 2021.