A/B Testing Social Networks and Online Marketplaces: The Problem of Network Effects

Carl Piehl
cpiehl@kth.se

Johan Henning
johennin@kth.se

April 2021

1 Introduction

A/B testing is an important, and widely used, practice in large scale software development. A/B testing is a form of controlled experimentation where the effect of a new feature is estimated by releasing it to a certain part of the user base, and comparing their behaviour to that of the users that have not received the feature. This allows companies to scientifically test the potential value of new features. This is crucial for the continuous development process, because assessing the value of a new feature can be very difficult. In a 2009 Microsoft paper[6], it is reported that only one third of the features tested at the company improve the targeted metrics.

Traditionally, a limiting factor in conducting controlled experiments is obtaining a large enough sample. Now, with the massive user bases of many internet companies, this is no longer an obstacle. Companies such as Microsoft and Google are able to run more than 10,000 experiments annually [5]. This means that decision making is transformed into a scientific, data-driven, process. However, new challenges have arisen as well, and basing decisions on badly designed experiments can have costly consequences.

In this essay we will explore the issue of network effects, a phenomenon that arises in A/B tests conducted on social networks and online marketplaces. We will explore how it presents itself on different platforms, and what methods companies have developed to perform experiments in its presence. The essay is outlined as follows. In section 2 we give a brief overview of A/B testing and introduce the relevant terminology. Section 3 explores the problem of network effects. In section 4 we summarize and look at alternative methods.

2 What is A/B Testing?

A/B testing is a form of *controlled experiment*. In a controlled experiment, all variables are held constant while a single variable, the independent variable, is manipulated in order to measure its effect on another variable, the dependent variable. The purpose of the experiment is to measure the causal relation between the variables. It is crucial that proper experimental control is exercised by only allowing the independent variable to change during the experiment. Otherwise, the measured effect might be caused by something else.

A/B testing is a way of testing new designs, features or products of an application. Instead of using test subjects or an internal testing team, a group of users, called *treatment*

group, is exposed to the experimental application version while another group, the *control group*, is not [3]. Importantly, this group assignment should be performed randomly, to avoid any *selection bias*, which might cause the groups to not be representative of the overall population. By observing both groups and measuring the relevant *metrics* it can be determined if the feature has the desired effect. An example of an A/B test can be seen in figure 1.

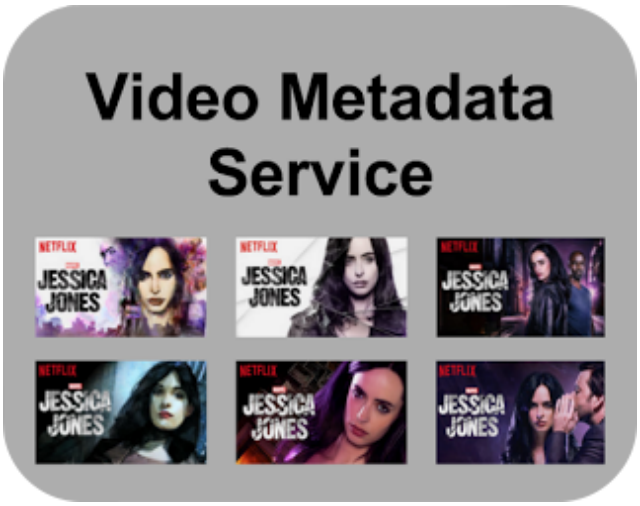

Figure 1: Netflix A/B test for different posters for a series [1].

If a test was successful, for instance if the new design generated more revenue or higher traffic, the build can be deployed into the main pipeline. On the other hand, if the A/B test gave worse results, the test for that specific design is discontinued and documented so history does not repeat itself. A very simplified version of the test process has been visualized in figure 2.

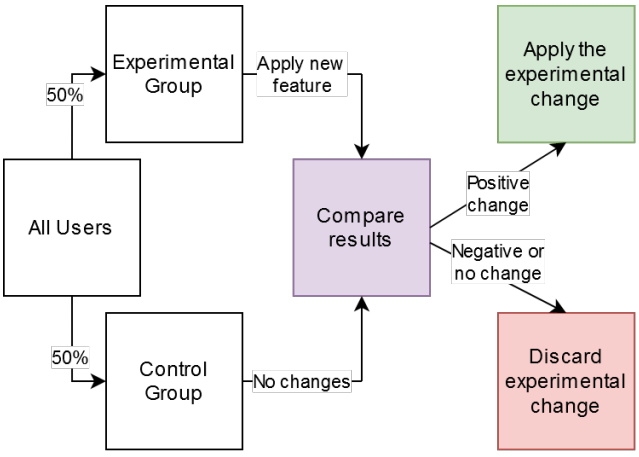

Figure 2: Visualized diagram of the A/B test phases.

3 Network Effects

A fundamental assumption when performing controlled experiments is the *Stable Unit Treatment Value Assumption* (SUTVA). This assumption states that the observed outcome of a unit is unaffected by the assignment of treatment of other units. For instance, in the context of A/B tests, the behaviour of a user in the control group cannot change based on other users receiving treatment. In regular offline experiments, this is usually ensured by limiting interaction between test subjects. Doing this in online experiments is much more difficult, especially on social network and online marketplace platforms, as we will see in this section. On these platforms, users are constantly interacting with each other. This means that users with different treatment assignments might interact with each other, and change their behaviour based on these interactions, thus breaking SUTVA. This phenomenon is called *network effects*. Depending on the tested feature and the type of network, the effects can vary greatly, meaning that there is no catch-all solution for this problem. In the following section we give a few examples from real world platforms of how network effects might appear, and what methods are used to mitigate the negative impact.

3.1 Social Networks

We will look at two examples of how network effects might appear. The first example is of LinkedIn out a new relevance algorithm for users' feeds [9]. We assume that this is an effective algorithm which increases the likelihood of a user engaging with their feed. Thus, the users in the treatment group start engaging more with their feed. However, this also affects the feeds of users in the control group that have connections to treated users, since the content of their feed is based on what their connections are engaging with. This, in turn, causes them to also start engaging with their feed. This is called the *spill-over effect*, where the effect of a treatment spills over to the control group.

The second example is from A/B testing on the Google Cloud Platform (GCP), a collaboration network. In this case, assume that two users are in separate groups, but are working on a project together. If one of the users has a feature the other does not have, this might cause confusion, leading to a worse experience for both users. This is called the *contamination effect* [10].

3.1.1 Cluster-based randomization

A common approach to deal with these problems is to, instead of randomizing at user-level, create clusters of users, and assign the entire cluster to a group. In [9] Saveski et. al present a method for this problem, used at LinkedIn. The idea is to have an "A/B test of A/B tests", where two separate tests are run. In one, the users are randomly assigned to treatment and control groups. In the other, the users are first divided into clusters based on their connections, then the entire cluster is assigned to treatment or control. This way, a user will be assigned to the same group as a significant part of their connections. Then the the difference in metrics between the cluster-based assignment and the random assignment becomes an indication of the level of bias resulting from network effects.

At OkCupid a slightly different approach to clustering is used. The platform is a dating app, where communities are not defined by existing connections, since users are more concerned with finding new connections than interacting with existing ones. This means that community-based clustering is not as effective. Instead, they define a "geo-social network" by investigating between which pairs of locations new connections are made frequently. This

means that two locations between which many new connections are made are likely to be in the same cluster [7].

3.2 Online Marketplaces

Online marketplaces, such as Ebay, are also affected by network effects. In online markets there are two types of users, buyers and sellers. An example of how the seller side might be affected is a feature which helps sellers price their items more realistically. If this is an effective feature, more buyers will engage with the sellers that have this feature. This experiment can give rise to two types of interference. In the first, buyers switch from the control group to the treatment group but do not engage more with the platform in general. This will give the impression that the treatment is effective, but in fact the number of completed auctions stays the same, meaning that revenue does not increase for the marketplace. This is called *cannibalization*. The other type of interference is that sellers from the control groups observe the prices used by the sellers in the treatment group and mirror them, giving the impression that the feature is ineffective. This is another example of the spill-over effect [8].

The way Ebay approaches this problem is similar to what we saw before with social networks. The key idea is to find another unit of randomization, so as to minimize user interactions across groups. In this case, randomization is done based on auction type. Buyers might switch between sellers with similar items, but they are not likely to switch to a completely different item, so the cannibalization is reduced. In the same vein, sellers will, generally, only monitor prices of other sellers with similar items, so the spill-over effect is also reduced [8].

3.2.1 Ridesharing Markets

Ridesharing markets are affected in the same way that marketplaces are affected, and they do not have different types of products to cluster by. Instead randomization is usually performed based on location or time interval [2]. For instance, you might use an entire city as treatment group, and a city with similar properties as control group.

A slightly more sophisticated method, used by Uber, is *synthetic control*. Instead of having a real control city, a model is built for the control city [4]. The objective of this model is to predict the relevant metrics for the chosen city. In order to make these predictions the model uses various features, such as data from similar cities, historic data from the modelled city, weather and event data etc. Once the model is built and verified, the feature is rolled out in the city and the observed metrics are compared to the model's prediction. This is illustrated in figure 3.

Figure 3: Synthetic control, in the pre-treatment period (the yellow portion of the graph) the model is trained and verified. In the post-treatment period (the blue portion) the observed metrics are compared to the model's predictions [4].

4 Conclusions

We have looked at how experimenting on social networks and online marketplaces introduces new challenges caused by network effects, where users from separate groups influence each other's behaviour, causing bias in the measured effects. All of the methods we have looked at deal with the problem by changing the way users are assigned to groups, so that randomization is no longer performed at user-level. This way, users are less likely to interact with users from a different group. It should be noted that this is by no means the only approach. For instance, another popular method is to try to estimate the network effect on the metric and account for it.

We have also seen that different types of platforms require different solutions. For example, Ebay and Uber both have online markets, which might be affected by spill-over effects and cannibalization. But their experiment design looks quite different, because the platforms are quite different. This might be why most larger companies develop their own experimentation platforms, tailored toward their specific use cases.

Finally, there are many other challenges in designing online experiments. Some examples include, choosing the appropriate metrics, estimating long-term effects when only experimenting for a limited amount of time, and avoiding interactions between multiple experiments (*carryover effects*).

References

[1] Netflix Technology Blog. *It's All A/Bout Testing: The Netflix Experimentation Platform*. https://netflixtechblog.com/its-all-a-bout-testing-the-netflix-experimentation-platform-4e1ca458c15?gi=dfc3c8853f8a. 2016.

[2] Nicholas Chamandy. *Experimentation in a Ridesharing Marketplace*. https://eng.lyft.com/experimentation-in-a-ridesharing-marketplace-b39db027a66e. 2016.

[3] Mary Earick Godby. *Control group*. https://www.britannica.com/science/control-group. 2016.

[4] Nick Jones and Sam Barrows. *Uber's Synthetic Control*. Uber, June 2019. URL: https://www.youtube.com/watch?v=j5DoJV5S2Ao.

[5] Ron Kohavi and Stefan Thomke. "The surprising power of online experiments". In: *Harvard Business Review* 95.5 (2017), pp. 74–82.

[6] Ronny Kohavi et al. "Online experimentation at Microsoft". In: *Data Mining Case Studies* 11.2009 (2009), p. 39.

[7] Brenton McMenamin. *The pitfalls of A/B testing in social networks*. https://tech.okcupid.com/the-pitfalls-of-a-b-testing-in-social-networks/. May 2017.

[8] Jason Wang et al. "Designing and analyzing A/B tests in an online marketplace". In: *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE. 2018, pp. 1447–1452.

[9] Ya Xu et al. "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*. 2015, pp. 2227–2236.

[10] Sangho Yoon. *Designing A/B tests in a collaboration network*. 2018. URL: http://www.unofficialgoogledatascience.com/2018/01/designing-ab-tests-in-collaboration.html.