



# Dealing with Interference on Experimentation Platforms

## Link

<http://nrs.harvard.edu/urn-3:HUL.InstRepos:39947197>

## Terms of use

This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material (LAA), as set forth at

<https://harvardwiki.atlassian.net/wiki/external/NGY5NDE4ZjgzNTc5NDQzMGIzZWZhMGFIOWI2M2EwYTg>

## Accessibility

<https://accessibility.huit.harvard.edu/digital-accessibility-policy>

## Share Your Story

The Harvard community has made this article openly available.  
Please share how this access benefits you. [Submit a story](#)

# Dealing with Interference on Experimentation Platforms

A dissertation presented

by

Jean Pouget-Abadie

to

The School of Engineering and Applied Sciences

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Computer Science

Harvard University

Cambridge, Massachusetts

September 2018

© 2018 Jean Pouget-Abadie

All rights reserved.

*Dissertation Advisor:*  
**Professor Edoardo M. Airoidi**

*Author:*  
**Jean Pouget-Abadie**

## **Dealing with Interference on Experimentation Platforms**

### **Abstract**

The theory of causal inference, as formalized by the potential outcomes framework, relies on an assumption that the experimental units are independent. When independence is not tenable, we say there is interference, and the core results of causal inference can no longer be guaranteed. Recent research efforts have focused on extending the theory to a setting where interference is present. The many advantages of experimentation platforms over more traditional settings of causal inference—no issue of non-compliance, large number of experimental units, ease of collecting outcomes over the course of an experiment—make them an ideal setting for studying causality with interference. With this setting in mind, we explore how multi-level designs, Experiment-of-Experiments, can allow us to detect and mitigate the effects of interference on experimentation platforms. In particular, we develop a design-based statistical test for the no-interference assumption. We further design an empirical procedure for comparing the effectiveness of cluster-based randomized designs. Finally, we show that randomized saturation designs can be optimized to improve the bias and variance of standard estimators, and extend these results to a new category of randomized designs: optimized saturation designs.

# Contents

Abstract . . . . .	iii
Acknowledgments . . . . .	ix
<b>Introduction</b>	<b>1</b>
<b>1 Testing for interference</b>	<b>6</b>
1.1 An Experiment-of-Experiments design . . . . .	8
1.2 Theoretical results . . . . .	12
1.2.1 The Type I error rate . . . . .	13
1.2.2 The type II error rate . . . . .	15
1.3 Variations of the Experiment-of-Experiments design . . . . .	16
1.3.1 Bernoulli randomization as an alternative assignment strategy . . . . .	16
1.3.2 Stratification and subsampling considerations . . . . .	17
1.4 Illustration on the <i>LinkedIn</i> experimentation platform . . . . .	18
1.4.1 Experimental set-up . . . . .	19
1.4.2 Clustering the <i>LinkedIn</i> graph . . . . .	19
1.4.3 Experimental results . . . . .	24
1.5 Conclusion . . . . .	26
<b>2 Optimizing cluster-based randomized experiments</b>	<b>27</b>
2.1 An Experiment-of-Experiments design . . . . .	29
2.1.1 A monotonicity assumption . . . . .	29
2.1.2 Design and analysis . . . . .	32
2.2 Application to reserve price experiments . . . . .	36
2.2.1 Single-item second price auctions . . . . .	37
2.2.2 Positional ad auctions . . . . .	38
2.3 Illustration on the <i>Yahoo!</i> bid dataset . . . . .	41
2.3.1 Simulating a reserve price experiment . . . . .	43
2.3.2 Experimental results . . . . .	46
2.4 Conclusion . . . . .	47

<b>3</b>	<b>Randomized and Optimized Saturation designs</b>	<b>49</b>
3.1	Randomized saturation designs . . . . .	51
3.1.1	Bias and variance under no interference . . . . .	53
3.1.2	Bias under a linear interference model . . . . .	55
3.1.3	Extension to a random graph model . . . . .	60
3.1.4	Extension to a stratified estimator . . . . .	61
3.2	Optimized saturation designs . . . . .	63
3.3	Simulation study . . . . .	66
3.3.1	The bias-variance tradeoff of randomized saturation designs . . . . .	67
3.3.2	The benefits of optimized saturation designs . . . . .	72
	<b>Conclusion</b>	<b>73</b>
	<b>References</b>	<b>77</b>
	<b>Appendix A Appendix to Chapter 1</b>	<b>82</b>
A.1	Review of the completely and cluster-based randomized assignments . . . . .	82
A.2	Proof of Lemma 1 . . . . .	85
A.3	Proof of Proposition 2 . . . . .	86
A.4	Proof of Theorem 1 . . . . .	86
A.5	Proof of Theorem 2 . . . . .	90
A.6	Proof of Theorem 3 . . . . .	91
A.7	Proof of Theorem 4 . . . . .	92
A.8	Proof of Theorem 5 . . . . .	94
	<b>Appendix B Appendix to Chapter 2</b>	<b>97</b>
B.1	Proof of Proposition 5 and 6 . . . . .	97
B.2	Proof of Proposition 7 . . . . .	98
B.3	Proof of Proposition 8 . . . . .	98
B.4	Proof of Proposition 9 . . . . .	99
	<b>Appendix C Appendix to Chapter 3</b>	<b>101</b>
C.1	Proof of Lemma 2 . . . . .	101
C.2	Proof of Proposition 10 . . . . .	102
C.3	Proof of Proposition 11 . . . . .	102
C.4	Proof of Corollary 1 . . . . .	106
C.5	Proof of Proposition 13 . . . . .	107
C.6	Proof of Proposition 14 . . . . .	107
C.7	Proof of Theorem 8 . . . . .	108
C.8	Proof of Proposition 15 . . . . .	110

C.9 Proof of Corollary 2 . . . . .	111
C.10 Proof for Corollary 3 . . . . .	112
C.11 Proof of Theorem 9 . . . . .	112
C.12 Proof of Example 1 . . . . .	113
C.13 Proof of Example 2 . . . . .	113

## List of Tables

1.1	Evaluation of clustering algorithms on the Netherlands LinkedIn graph. . . .	21
1.2	Evaluation of different cluster-sizes for the ReLDG algorithm. . . . .	22
1.3	Results of the LinkedIn test for interference . . . . .	25
2.1	Summary statistics of the <i>Yahoo!</i> dataset . . . . .	41
3.1	Block-model matrices used in simulation . . . . .	73
3.2	Evaluation of optimized saturation designs under interference . . . . .	74

## List of Figures

1.1	Illustration of the suggested experiment-of-experiments design . . . . .	11
1.2	Impact of clustering quality on variance upper-bound estimator . . . . .	23
1.3	Impact of interference and clustering quality on experimental power . . . . .	24
2.1	Illustration of the suggested experiment-of-experiments design . . . . .	33
2.2	Click-through-rate per rank in <i>Yahoo!</i> dataset . . . . .	39
2.3	Graph-cut values for the reLDG algorithm . . . . .	43
2.4	Experimental results of the reserve price simulation . . . . .	44
3.1	Various forms of the beta distribution . . . . .	68
3.2	Standard deviation of $\hat{\tau}$ under SUTVA . . . . .	69
3.3	Mean-squared error of $\hat{\tau}$ under interference . . . . .	71

## Acknowledgments

This thesis would not have been possible without the guidance of my incredibly patient advisor and wonderful human being, Edoardo M. Airoidi, as well as the advice and supervision of my committee members, David C. Parkes, Donald B. Rubin, and Salil Vadhan.

I am equally grateful for my Harvard colleagues and friends for a very sweet and instructive four years. I want to thank my friends in the EconCS group: Eric Balkanski, Hongyao Ma, Debmalya Mandal, Paul Tylkin, Andrew Mao, Greg Stoddard, Bo Waggoner, Rediet Abebe, Chara Podimata, Zhe Feng, and Dimitris Kalimeris; my officemates and friends in the Theory Group: Jack Murtagh and Jarek Blasiok; and my friends in the Statistics department: Guillaume Basse, Niloy Biswas, Nicole Pashley, Kathryn McKeough, Albert Wu, and Stéphane Shao. A special thank you to Thibaut Horel, who gave me more of his time than I could have possibly deserved.

I want to thank all my collaborators at LinkedIn, Spotify, Facebook, and Google, namely Vidhya Murali, Romain Yon, Rohan Agrawal, Isabel Kloumann, Jonathan Tannen, Kevin Aydin, Souvik Ghosh, Weitao Duan, Guillaume Saint-Jacques, and Ya Xu. I particularly want to thank Udi Weinsberg and Vahab Mirrokni for their mentorship and continued encouragement.

I would like to thank my friends for their love and support and the Grants and Zanetises for treating me like family. Finally, I could not be more proud of and thankful for my own wonderful family, Bénédicte, Marc, Théophile, Thomas, and Blandine Pouget-Abadie.

This work was supported by National Science Foundation grants IIS-1149662 and IIS-1409177, Office of Naval Research Grant N00014-17-1-2131, a Siebel fellowship, and a Google Research Award.

To Tripp Zanetis. May I continue to do you proud.

# Introduction

Causal inference is the study of the relationship between cause and effect of an intervention on a system. The potential outcomes approach to causal inference finds its roots in Jerzey Neyman's (Splawa-Neyman *et al.*, 1923, 1990) and Ronald A. Fisher's work (Fisher, 1919, 1925, 1935), and was later continued by Donald Rubin in (Rubin, 1974, 1978). The framework postulates that for every possible intervention on a collection of units, there are two possible outcomes per unit: one "potential outcome" if the intervention occurs and one "potential outcome" if the intervention does not occur.

For example, consider the cholesterol level of an individual if they take a specific medication as well as their cholesterol level if they do not. The outcome metric of interest is the individual's cholesterol, and the intervention is whether or not they took the medication. Together, these two possible values of cholesterol make up the potential outcomes. We can define the causal effect of the medication as the difference between the two potential cholesterol levels. This framework can be easily extended to multi-level treatments.

The potential outcomes notation frames the question of causality as a counter-factual problem: we can only ever observe one potential outcome. Fisher's intuition was that this could be resolved with randomization. By randomizing over which units receive the intervention and which units do not receive the intervention, it is possible to estimate the causal effect unbiasedly under a specific set of assumptions.

There are many possible kinds of randomized assignment mechanisms, and not every one is appropriate for drawing causal inferences. We will restrict ourselves to assignment mechanisms that are *individualistic, probabilistic, and unconfounded* (Imbens and Rubin, 2015),

for which the common difference-in-means estimator is unbiased for the treatment effect of an intervention. One additional assumption must hold true: the Standard Unit Treatment Value Assumption (SUTVA) (Rubin, 1980), also known as the Individualistic Treatment Response (ITR) (Manski, 2013) assumption. The stable unit treatment value assumption states that

*“The potential outcomes for any unit do not vary with the treatments assigned to other units, and, for each unit, there are no different forms or version of each treatment level, which lead to different potential outcomes.”*

— (Imbens and Rubin, 2015)

In other words, the assumption holds if there is only a single version of each treatment level (every prescribed cholesterol tablet of identical dose is considered equally efficacious) *and* the treatment assignment of one unit does not interfere with another unit’s potential outcomes. The no-interference component of this assumption—the treatment of one unit does not affect the outcome of another unit—is not always tenable.

Mitigating interference was a concern from the start of randomized experiments. A famous historical example was the use of “guard rows” to separate plots of land and prevent the run-off from fertilized acres from contaminating unfertilized acres (Kempton, 1997). Immunization campaigns are another classic example of randomized experiments where measures were taken to mitigate interference. Indeed, the concept of “herd immunity” alerted researchers to the fact that vaccinating an individual might impact the outcome of other individuals in his or her social circle (Struchiner *et al.*, 1990).

As the scope of randomized experiments grew over time, so did our awareness of interference. Precautions taken to mitigate the effects of interference can be found in recent causal analyses, from education policy (Hong and Raudenbush, 2012), viral marketing campaigns (Aral and Walker, 2011; Eckles *et al.*, 2016), and healthcare (Shakya *et al.*, 2017). Large technology companies, with their reliance on ubiquitous A/B tests, have become a new source of randomized experiments with interference (Eckles *et al.*, 2016; Gui *et al.*, 2015). The experiments conducted by technology companies are an ideal testing ground

for quantifying and mitigating interference. In contrast with clinical trials or public policy experiments (Sobel, 2006), technology companies have millions of units at their disposal and rarely have to worry about notions of compliance. Furthermore, their mature experimentation platforms, (e.g. Google (Tang *et al.*, 2010), Microsoft, (Kohavi *et al.*, 2013), Facebook (Bakshy *et al.*, 2014), LinkedIn (Xu *et al.*, 2015)), simplify the collection of outcome metrics and covariates.

With the advent of online social networks in so many technological products today, examples of A/B tests that suffer from interference have multiplied. Consider, for example, an intervention on a user of a messaging platform designed to modify her behavior, causing a faster response time and encouraging her to increase the number of messages she initiates. The resulting behavioral change will invariably affect the friends on the platform she chooses to communicate with; their own response time might decrease, which might in turn reduce the response time of *their* friends. If one were simply to compare the response time of units assigned to the intervention with the response time of units not assigned to the intervention, both having decreased due to the intervention and the social mechanisms at play, the true impact of the change to the messaging platform would be underestimated. In the extreme case that the response time of every user decreases equally due to interference, the unwary statistician might even conclude that the intervention has had no effect on the response time of its users!

When interference is present, and the stable unit treatment value assumption is violated, many fundamental results of the causal inference literature no longer hold. As illustrated in the previous example, the difference-in-means estimator under a completely randomized assignment is no longer guaranteed to be an unbiased estimator of the average treatment effect (Imbens and Rubin, 2015). Similarly, controlling the variance of common estimators becomes infeasible without a parametric model for outcomes (Basse and Airoidi, 2017).

The research community has made significant efforts to extend the theory of causal inference to scenarios where the stable unit treatment value assumption does not hold. A popular approach to minimizing the effects of interference, *cluster-based randomized designs*,

have been extensively studied, spanning from the early work of (Cornfield, 1978; COMMIT, 1991; Donner and Klar, 2004; Murray *et al.*, 2004) to more recent contributions by (Ugander *et al.*, 2013; Walker and Muchnik, 2014; Eckles *et al.*, 2017). Cluster-based randomized designs assign units to treatment or control in groups to limit the amount of interaction between different treatment buckets.

Designs where treatment is applied with different proportions across the population, known as *randomized saturation designs* (Hudgens and Halloran, 2008; Tchetgen and VanderWeele, 2012; Baird *et al.*, 2016), are another important tenet of the literature on improving causal estimates under interference, having been applied to vaccination trials (Datta *et al.*, 1999) and voter-mobilization campaigns (Sinclair *et al.*, 2012). More recently, a literature has developed around various assignment strategies and estimators, beyond cluster-based randomized design or randomized saturation designs with specific guarantees under specific models of interference (Backstrom and Kleinberg, 2011; Katzir *et al.*, 2012; Toulis and Kao, 2013; Manski, 2013; Choi, 2014; Basse and Airoidi, 2015; Gui *et al.*, 2015).

In the following chapters, we will provide answers to the following questions on causal inference with network interference:

1. Can we design a randomized experiment that tests, with minimal assumptions, whether interference is present? [**Chapter 1**]
2. Can we determine which of two cluster-based randomized designs is most appropriate without explicit knowledge of the interference structure of the experiment? [**Chapter 2**]
3. When do we stand to gain from running randomization saturation designs in the presence of interference? [**Chapter 3**]

In tackling these questions throughout the following chapters, we will often look to a particular type of randomized experiment, which we name *Experiment-of-Experiments designs*, to provide answers. Experiment-of-Experiments designs implement the idea of testing multiple randomized design strategies within the same experiment. In Chapter 1, we develop an Experiment-of-Experiments design to test whether the no-interference assumption is

violated in an experiment. In Chapter 2, we introduce an Experiment-of-Experiments design to allow an empirical comparison of cluster-based randomized designs and their effectiveness. In Chapter 3, we show that randomized saturation designs—also a type Experiment-of-Experiments design—can be optimized to reduce the bias and variance of common estimators. In fact, we show that these results can be further improved with a new category of randomized designs: optimized saturation designs. In developing solutions from this common thread, we hope to show the power and versatility of Experiment-of-Experiments designs in answering questions of causal inference with interference.

# Chapter 1

## Testing for interference

In this chapter, we tackle the first research question listed in the introductory chapter, which we restate here:

*Can we design a randomized experiment that tests, with minimal assumptions, whether interference is present?*

Statisticians at large technology companies have millions of experimental units at their disposal, and rarely deal with issues of compliance<sup>1</sup>. One remaining primary concern that experimenters face is the presence of bias in their experiments, primarily due to interference. Without ground truth knowledge, it is impossible to know whether the estimators of an experiment are biased or not. Testing whether SUTVA holds serves as a litmus test for whether the standard estimators are unbiased or whether more sophisticated causal inference designs or imputation mechanisms need to be deployed.

The null hypothesis that SUTVA holds can be formulated as such:

$$\forall \mathbf{Z}, \mathbf{Z}' \in \{0, 1\}^N, \forall i, Z_i = Z'_i \implies Y_i(\mathbf{Z}) = Y_i(\mathbf{Z}') \quad (1.1)$$

It is not a sharp null and the standard Fisher randomization test does not apply here.

---

<sup>1</sup>There are exceptions; for example, when analysing the launch of a new feature/interface on a website or mobile application that is accessible to a user only if that user accepts the change, by e.g. upgrading their mobile application.

(Rosenbaum, 2007) was the first to formulate two sharp nulls, which are necessary but not sufficient conditions for SUTVA to hold. More recent work (Aronow, 2012; Athey *et al.*, 2015; Basse *et al.*, 2017) explicitly tackles testing for the non-sharp null that SUTVA holds by considering the distribution of an interference effect parameter of the experimenter’s choosing for a subpopulation of the graph under SUTVA. The clear advantage of these analysis-centric approaches is that they does not interfere with the design of the experiment, hence giving the experimenter the freedom to choose the method of analysis. However, the result is ultimately dependent on the chosen interference parameter.

In this paper, we propose an Experiment-of-Experiments design that allows the experimenter to test whether SUTVA holds. Our design-centric approach is in the spirit of the Durbin-Wu-Hausman test for endogeneity in econometrics (Durbin, 1954; Wu, 1973; Hausman, 1978), in which multiple estimators return the same estimate if and only if the null hypothesis holds. The design that we introduce makes no assumptions on the interference model between units; it comes with a sharp bound on the variance (under some conditions) and an implied analytical bound on the Type I error rate. Most importantly, the proposed design is non-intrusive in that it allows the experimenter to analyse the experiment in a classical way with a smaller sample size.

In Section 1.1, we introduce and provide intuition for the suggested Experiment-of-Experiment design and resulting statistical test for interference. We provide guarantees on the Type I and II error of the suggested test in Section 1.2. In Section 1.3, we discuss possible modifications to the Experiment-of-Experiments design that can be implemented to meet practical constraints. Finally, in Section 1.4, we present some of the results obtained for an experiment launched in August 2016 on LinkedIn’s experimentation platform using our suggested framework.

**Acknowledgements:** The results of this chapter are based on two research papers (Saveski *et al.*, 2017; Pouget-Abadie *et al.*, 2017), which are the results of a collaboration with Edoardo M. Airoidi, Weitao Duan, Souvik Ghosh, Guillaume Saint-Jacques, Martin Saveski, and Ya Xu.

## 1.1 An Experiment-of-Experiments design

Consider  $N$  experimental units which we can assign to either treatment or control. Let  $Z_i$  indicate the intervention assigned to unit  $i$ :  $Z_i = 1$  if unit  $i$  is assigned to treatment and  $Z_i = 0$  if unit  $i$  is assigned to control. Per the potential outcomes framework (Imbens and Rubin, 2015), each unit has a potential outcome  $Y_i(\mathbf{Z})$  for each assignment vector  $\mathbf{Z} \in \{0, 1\}^N$ . The causal estimand of interest is the *Total Treatment Effect* (TTE) defined by:

$$TTE := \frac{1}{N} \sum_{i=1}^N Y_i(\mathbf{Z} = \mathbf{1}) - \frac{1}{N} \sum_{i=1}^N Y_i(\mathbf{Z} = \mathbf{0}). \quad (1.2)$$

We denote by  $\mathbf{Y}(\mathbf{Z})$  the potential outcome vector under assignment  $\mathbf{Z}$ . For any vector  $\mathbf{u} \in \mathbb{R}^N$ , let  $\bar{\mathbf{u}} := \frac{1}{N} \sum_{i=1}^N u_i$  and  $\sigma^2(\mathbf{u}) := \frac{1}{N-1} \sum_{i=1}^N (u_i - \bar{u})^2$ . The definition of the Total Treatment Effect (TTE) can be rewritten using this notation:

$$TTE = \overline{\mathbf{Y}(\mathbf{1})} - \overline{\mathbf{Y}(\mathbf{0})}. \quad (1.3)$$

Two popular experimental designs, the Completely Randomized (CR) design and the Cluster-Based Randomized (CBR) design, are unbiased for the Total Treatment Effect when SUTVA is present. Recall that in a *Completely Randomized* experiment (CR), we sample the assignment vector  $\mathbf{Z}$  uniformly at random from the set  $\{\mathbf{z} \in \{0, 1\}^N : \sum z_i = n_t\}$ , where  $n_t$  is the number of units assigned to treatment and  $n_c := N - n_t$  is the number of units assigned to control. In a *Cluster-Based Randomized* assignment (CBR), we randomize over clusters of units in the graph, rather than individual units. We suppose that each unit is assigned to one of  $M$  clusters. We sample the cluster assignment vector  $\mathbf{z}$  uniformly at random from  $\{\mathbf{v} \in \{0, 1\}^M : \sum v_i = m_t\}$ , assigning units in cluster  $\mathcal{C}_j$  to the corresponding treatment:  $Z_i = 1 \Leftrightarrow z_j = 1$  if  $i \in \mathcal{C}_j$ , where  $m_t$  is the number of clusters assigned to treatment and  $m_c := M - m_t$  is the number of clusters assigned to control. We use the notation  $\mathbf{Z} \sim CR$  to denote an assignment vector sampled according to a completely randomized (CR) design, and  $\mathbf{Z} \sim CBR$  to denote an assignment vector sampled according to a cluster-based randomized (CBR) design.

Let  $\mathbf{Y}_t := \{Y_i : Z_i = 1\}$  be the outcome vector of all treatment units and  $\mathbf{Y}_c := \{Y_i : Z_i =$

0} be the outcome vector of all control units. We define the difference-in-means estimator  $\hat{\tau}_{cr}$  as

$$\hat{\tau}_{cr} := \bar{\mathbf{Y}}_t - \bar{\mathbf{Y}}_c \quad (1.4)$$

Recall that the variance  $\sigma_{cr}^2$  of the difference-in-means estimator under a completely randomized assignment is given by:

$$\sigma_{cr}^2 := \text{Var}_{\mathbf{Z} \sim \text{CR}}[\hat{\tau}_{cr}] = \frac{S_t}{n_t} + \frac{S_c}{n_c} - \frac{S_{tc}}{N} \quad (1.5)$$

where  $S_t := \sigma^2(\mathbf{Y}(1))$ ,  $S_c := \sigma^2(\mathbf{Y}(0))$  and  $S_{tc} := \sigma^2(\mathbf{Y}(1) - \mathbf{Y}(0))$ , where  $\mathbf{Y}(1) - \mathbf{Y}(0)$  is the element-wise difference between vectors  $\mathbf{Y}(1)$  and  $\mathbf{Y}(0)$ . See Section A.1 in the Appendix for more details.

Let  $Y_j^+ = \sum_{i \in \mathcal{C}_j} Y_i$  be the aggregated outcome of cluster  $j$ . Let  $\mathbf{Y}_t^+ := \{Y_j^+ : z_j = 1\}$ , and  $\mathbf{Y}_c^+ := \{Y_j^+ : z_j = 0\}$ . We define the Horvitz-Thompson estimator as:

$$\hat{\tau}_{cbr} := \frac{M}{N} \left( \bar{\mathbf{Y}}_t^+ - \bar{\mathbf{Y}}_c^+ \right) \quad (1.6)$$

Recall that the variance of  $\hat{\tau}_{cbr}$  under a cluster-based randomized assignment is given by:

$$\sigma_{cbr}^2 := \text{Var}_{\mathbf{Z} \sim \text{CBR}}[\hat{\tau}_{cbr}] = \frac{M^2}{N^2} \left( \frac{S_t^+}{m_t} + \frac{S_c^+}{m_c} - \frac{S_{tc}^+}{M} \right) \quad (1.7)$$

where  $S_t^+ := \sigma^2(\mathbf{Y}^+(1))$ ,  $S_c^+ := \sigma^2(\mathbf{Y}^+(0))$  and  $S_{tc}^+ := \sigma^2(\mathbf{Y}^+(1) - \mathbf{Y}^+(0))$ . See Section A.1 in the Appendix for more details.

It is a well-known result that the difference-in-means estimator is unbiased for the Total Treatment Effect under SUTVA for a completely-randomized assignment and that the Horvitz-Thompson estimator is also unbiased for the Total Treatment Effect under SUTVA for a cluster-based randomized assignment (Middleton and Aronow, 2011):

$$\mathbb{E}_{\mathbf{Z} \sim \text{CR}}[\hat{\tau}_{cr}] = \mathbb{E}_{\mathbf{Z} \sim \text{CBR}}[\hat{\tau}_{cbr}] = TTE \quad (1.8)$$

The unbiasedness of the cluster-based randomized (CBR) design does not require the clusters to be of equal size. When SUTVA does not hold, the equality result of Equation 1.8 is no longer guaranteed and we expect the estimate of the Total Treatment Effect to be

different under each design when interference is present.

For example, assume a network over the units of experimentation, such that the immediate neighborhood  $\mathcal{N}_i$  of a unit  $i$  are the units likely to interfere with unit  $i$ , and the following linear model of potential outcomes:

$$\forall i, Y_i(\mathbf{Z}) = \alpha + \beta Z_i + \gamma \rho_i + \epsilon_i \quad (1.9)$$

where  $\rho_i = \frac{1}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} Z_j$  is the average number of treated neighbors in unit  $i$ 's neighborhood  $\mathcal{N}_i$  and  $\epsilon_i \sim \mathcal{N}(0, \sigma^2) \perp \rho_i$ .  $\beta$  is often interpreted as a “direct treatment effect” parameter and  $\gamma$  is often interpreted as an “interference effect” parameter.

**Lemma 1.** *Under the model of interference in Eq. 1.9, the expectation of  $\hat{\tau}_{cr}$  under a completely-randomized (CR) assignment and  $\hat{\tau}_{cbr}$  under a cluster-based randomized (CBR) assignment are:*

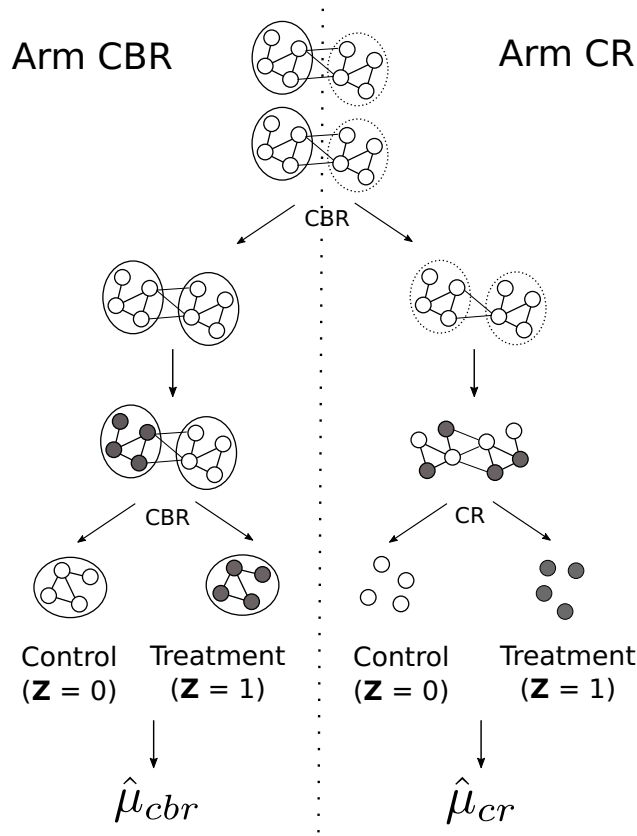
$$\mathbb{E}_{\mathbf{Z}, \epsilon}[\hat{\tau}_{cr}] = \beta - \frac{\gamma}{N - 1} \quad (1.10)$$

$$\mathbb{E}_{\mathbf{Z}, \epsilon}[\hat{\tau}_{cbr}] = \beta - \gamma \left( \frac{1 - \rho_C \cdot M}{M - 1} \right) \quad (1.11)$$

where  $\rho_C := \frac{1}{N} \sum_i \frac{|\mathcal{N}_i \cap \mathcal{C}(i)|}{|\mathcal{N}_i|}$  and  $\mathcal{C}(i)$  is the cluster unit  $i$  belongs to.

A proof is included in Section A.2 of the Appendix.  $\rho_C$  can be interpreted as the average number of edges not cut by the partitioning, weighted by each node’s degree. If the clustering is perfect (no edges cut), then  $\rho_C = 1$ . According to Lemma 1, when interference is present, neither estimator is unbiased for the Total Treatment Effect (TTE). Critically, the two estimators do not have the same expectation under the interference model in Equation 1.9. As a result, if it were possible to apply both the completely randomized and cluster-based randomized designs, we could test for interference by comparing the two estimates from each assignment strategy: if the two estimates are significantly different, there is interference; otherwise, we expect that there is no interference.

Unfortunately, just as we cannot observe each unit’s outcome under both treatment and control, we cannot observe every unit’s outcome under both designs. We solve this problem by randomly assigning units to one of two treatment *arms* and, within each treatment arm, following either a CR or a CBR design. In order to assign units to treatment *arms*, we



**Figure 1.1:** Illustration of the experiment-of-experiments design for testing for interference.

use a cluster-based randomized design. This allows us to maintain some of the network interference structure intact within each treatment arm without sacrificing covariate balance or introducing bias. Once units are assigned to treatment arms, we apply within each treatment arm either a cluster-based randomized design or a completely randomized design.

The procedure is described in pseudo-code in Algorithm 1 and illustrated in Figure 1.1. We group the units into  $M$  clusters  $\mathcal{C}$ , each cluster having the same number of units. We then assign units to treatment arms  $cr$  and  $cbr$  following a cluster-based design using  $\mathcal{C}$  as the clustering. Let  $\mathbf{W} \in \{0,1\}^N$  be the corresponding assignment vector of units to these treatment arms. Let  $m_{cr}$  (resp.  $n_{cr}$ ) be the number of clusters (resp. units) assigned to treatment arm  $cr$  ( $W_i = 1$ ) and let  $m_{cbr}$  (resp.  $n_{cbr}$ ) be the number of clusters (resp. units) assigned to treatment arm  $cbr$  ( $W_i = 0$ ). We assign all units such that  $W_i = 1$  to treatment

---

**Algorithm 1: Testing for interference**

---

Cluster the  $N$  units into  $M$  clusters  $\mathcal{C}$ ;  
Assign  $m_{cr}$  clusters to treatment arm  $cr$  uniformly at random;  
Assign the remaining  $m_{cbr}$  clusters to treatment arm  $cbr$ ;  
Assign  $n_{cr,t}$  units in treatment arm  $cr$  to treatment uniformly at random;  
Assign the remaining  $n_{cr,c}$  units in treatment arm  $cr$  to control;  
Assign  $m_{cbr,t}$  clusters within treatment arm  $cbr$  to treatment uniformly at random;  
Assign the remaining  $m_{cbr,c}$  clusters within treatment arm  $cbr$  to control;

---

and control using a completely randomized assignment, and all units such that  $W_i = 0$  to treatment and control using a cluster-based randomized assignment. We let  $\mathbf{Z} \in \{0, 1\}^N$  be the resulting assignment vector of units to treatment and control.

## 1.2 Theoretical results

In this section, we show how to conduct a statistical test for SUTVA from the Experiment-of-Experiments design presented in Section 1.1. We first define a difference-in-differences estimator. Let  $\mathbf{Y}_{cr,t} := \{Y_i : W_i = 1 \wedge Z_i = 1\}$  be the outcomes of the treated units in the  $cr$  arm and let  $\mathbf{Y}_{cr,c} := \{Y_i : W_i = 1 \wedge Z_i = 0\}$  be the outcomes of the control units in the  $cr$  arm. Recall that  $Y_j^+ = \sum_{i \in \mathcal{C}_j} Y_i$  is the aggregated outcome of cluster  $\mathcal{C}_j$ . Let  $\mathbf{Y}_{cbr,t}^+ := \{Y_j^+ : \forall i \in \mathcal{C}_j, W_i = 0 \wedge Z_i = 1\}$  be the aggregated cluster-outcomes of the clusters assigned to treatment in the  $cbr$  arm, and let  $\mathbf{Y}_{cbr,c}^+ := \{Y_j^+ : \forall i \in \mathcal{C}_j, W_i = 0 \wedge Z_i = 0\}$  be the aggregated cluster-outcomes of clusters assigned to control in the  $cr$  arm. We define two estimators for the Total Treatment Effect as well as the difference-in-differences estimator  $\Delta$ :

$$\hat{\tau}_{cr} := \bar{\mathbf{Y}}_{cr,t} - \bar{\mathbf{Y}}_{cr,c} \quad (1.12)$$

$$\hat{\tau}_{cbr} := \frac{m_{cbr}}{n_{cbr}} \left( \bar{\mathbf{Y}}_{cbr,t}^+ - \bar{\mathbf{Y}}_{cbr,c}^+ \right) \quad (1.13)$$

$$\Delta := \hat{\tau}_{cr} - \hat{\tau}_{cbr} \quad (1.14)$$

We examine the first moment of the difference-in-differences estimator under SUTVA in the following theorem.

**Theorem 1.** *If SUTVA holds, and each cluster in  $\mathcal{C}$  has the same size, then the expectation and variance of the difference-in-differences estimator is given by:*

$$\mathbb{E}_{\mathbf{W}, \mathbf{Z}}[\Delta] = 0 \quad (1.15)$$

$$\text{Var}_{\mathbf{W}, \mathbf{Z}}[\Delta] = \frac{n_{cr}}{n_{cr} - 1} \frac{M}{M - 1} \sigma_{cr}^2 + \left(1 - \frac{m_{cbr}}{N(n_{cr} - 1)}\right) \sigma_{cbr}^2 + \frac{M}{n_{cr} n_{cbr}} S_{tc}^+ \quad (1.16)$$

The definitions of  $\sigma_{cr}^2$  and  $\sigma_{cbr}^2$  can be found in Equations 1.5 and 1.7. A proof of Theorem 1 can be found in Section A.4. We conduct the following hypothesis test.

**Proposition 1.** *Let the Null hypothesis be that SUTVA holds and let  $\hat{\sigma}^2 \in \mathbb{R}_+$  be any computable quantity from the experimental data which upper-bounds the true variance:  $\hat{\sigma}^2 \geq \text{Var}_{\mathbf{W}, \mathbf{Z}}[\Delta]$ . We reject the null if and only if:*

$$\frac{|\hat{\tau}_{cr} - \hat{\tau}_{cbr}|}{\sqrt{\hat{\sigma}^2}} \geq \frac{1}{\sqrt{\alpha}}. \quad (1.17)$$

### 1.2.1 The Type I error rate

The following proposition controls the Type I error of our hypothesis test. It is a straightforward extension of Chebychev's inequality.

**Proposition 2.** *If SUTVA holds, and each cluster in  $\mathcal{C}$  has the same size, then we reject the null (incorrectly) with probability no greater than  $\alpha$ .*

A proof of Proposition 2 is included in the supplementary material. This result holds for any balanced clustering and for any model of interference. This is not surprising because the theorem states a result on the Type I error of our test, under which we can assume SUTVA. The tricky part of this hypothesis test is coming up with an appropriate upper-bound for the variance.

We now define a good candidate to upper-bound the variance inspired from Neyman's variance estimator. Let  $\hat{S}_{cr,t} := \sigma^2(Y_i : W_i = 1 \wedge Z_i = 1)$ ,  $\hat{S}_{cr,c} := \sigma^2(Y_i : W_i = 1 \wedge Z_i = 0)$ ,  $\hat{S}_{cbr,t}^+ := \sigma^2(Y_j^+ : \omega_j = 0 \wedge z_j = 1)$ , and  $\hat{S}_{cbr,c}^+ := \sigma^2(Y_j^+ : \omega_j = 0 \wedge z_j = 0)$ .

**Theorem 2.** Let  $\hat{\sigma}$  be our empirical variance upper-bound defined by:

$$\hat{\sigma}^2 := \frac{\hat{S}_{cr,t}}{n_{cr,t}} + \frac{\hat{S}_{cr,c}}{n_{cr,c}} + \frac{m_{cbr}^2}{n_{cbr}^2} \left( \frac{\hat{S}_{cbr,t}^+}{m_{cbr,t}} + \frac{\hat{S}_{cbr,c}^+}{m_{cbr,c}} \right) \quad (1.18)$$

If SUTVA holds and all the clusters of  $\mathcal{C}$  are the same size, then the previous quantity upper-bounds the theoretical variance of the  $\Delta$  estimator in expectation:

$$\mathbb{E}_{\mathbf{W},\mathbf{Z}} [\hat{\sigma}^2] \geq \text{Var}_{\mathbf{W},\mathbf{Z}}[\Delta]$$

In the case of a constant treatment effect,  $\exists \tau \in \mathbb{R}, \forall i, Y_i(1) = Y_i(0) + \tau$ , the above inequality is tight:  $\mathbb{E}_{\mathbf{W},\mathbf{Z}} [\hat{\sigma}^2] = \text{Var}_{\mathbf{W},\mathbf{Z}}[\Delta]$ .

In other words, summing the normalized variances of the observed outcomes in each treatment bucket of each treatment arm upper-bounds the variance of the difference-in-means estimator in expectation. The condition of Proposition 1 will be met only in expectation, however this is often deemed to be a sufficient condition in the literature (Imbens and Rubin, 2015).

Another possibility for upper-bounding the theoretical variance of the difference-in-differences estimator is to approximate it assuming Fisher's Null of no treatment effect:  $\forall i, Y_i(1) = Y_i(0)$ . Under Fisher's null, the theoretical formula for the variance becomes computable from the observed data. It is sometimes reasonable to assume that the variance of our estimator under Fisher's null is a good proxy for the variance under SUTVA, though this assumption may not always be appropriate. Let  $S := \sigma^2(\mathbf{Y})$  be the variance of all observed potential outcomes, and  $S^+ := \sigma^2(\mathbf{Y}^+)$  be the variance of all observed aggregated outcomes.

**Theorem 3.** Under Fisher's null hypothesis of no treatment effect, if all clusters of  $\mathcal{C}$  have the same size,

$$\text{Var}_{\mathbf{W},\mathbf{Z}}[\Delta] = \frac{n_{cr}}{n_{cr} - 1} \frac{M}{M - 1} \frac{n_{cr}}{n_{cr,t} n_{cr,c}} S + \left( 1 - \frac{m_{cbr}}{N(n_{cr} - 1)} \right) \frac{m_{cbr}}{m_{cbr,t} m_{cbr,c}} S^+ \quad (1.19)$$

A proof is included in Section A.6. Finally, a different way of rejecting the null from Proposition 1 is to approximate the test statistic  $T := \frac{\hat{\mu}_{cr} - \hat{\mu}_{cbr}}{\sqrt{\hat{\sigma}^2}}$  by a normal distribution  $\mathcal{N}(0, 1)$ .

In this case we obtain the following conservative  $(1 - \alpha) \times 100\%$  confidence intervals:

$$CI^{1-\alpha}(T) = \left( T - z_{\frac{\alpha}{2}}, T + z_{1-\frac{\alpha}{2}} \right) \quad (1.20)$$

where  $z_{\frac{\alpha}{2}}$  and  $z_{1-\frac{\alpha}{2}}$  are the  $\frac{\alpha}{2}$  quantile of the standard normal distribution.

### 1.2.2 The type II error rate

To paraphrase the result stated in Proposition 2, if we set our rejection region to  $\{T \geq \frac{1}{\sqrt{\alpha}}\}$  and the inequality  $\hat{\sigma}^2 \geq \sigma^2$  holds, then the probability of falsely rejecting the null is lower than  $\alpha$ . Computing the Type I error is straightforward because we work under the assumption that SUTVA holds. The same is not true of the type II error rate, where we must posit a model for the interference between units.

In Section 1.1, we saw that the expectation of the  $\hat{\mu}_{cr}$  and  $\hat{\mu}_{cbr}$  estimators for CR and CBR assignments differed under a linear model of interference described in Equation 1.9. We complete this analysis by computing the type II error of our test under this same model of interference. Recall that  $\rho_C := \frac{1}{N} \sum_{i=1}^N \frac{|\mathcal{N}_i \cap \mathcal{C}(i)|}{|\mathcal{N}_i|}$  is the average fraction of a unit's neighbors contained within its cluster  $\mathcal{C}(i)$ .

**Theorem 4.** *If all clusters of  $\mathcal{C}$  have the same size, then under the linear model of interference defined in Eq. 2.7, the expectation of the difference-in-differences  $\Delta$  estimator under our suggested hierarchical design is given by:*

$$E_{\mathbf{w}, \mathbf{z}}[\Delta] \approx \gamma \cdot \rho_C \quad (1.21)$$

*under the assumption that  $n_{cr} \gg 1$ ,  $m_{cbr} \gg 1$ , and  $m_{cr} \gg 1$ .*

A proof is included in Section A.7 of the appendix. The result of Theorem 4 is intuitive: the greater the interference parameter  $|\gamma|$  and the better the clustering (higher  $\rho_C$ ), the larger the expected difference between the estimators in each of the two arms is.

Knowing the type II error rate can help us determine which clustering of the graph is most appropriate. The selection of hyper-parameters in clustering algorithms, including the number of clusters to set, can be informed by minimizing the type II error under

plausible models of interference. The optimization program varies depending on the choice of variance estimator  $\hat{\sigma}_{\mathcal{C}}^2$  for a clustering  $\mathcal{C}$ :

$$\max_{M, \mathcal{C}} \frac{\rho_{\mathcal{C}}}{\sqrt{\hat{\sigma}_{\mathcal{C}}^2}},$$

where  $\mathcal{C}$  is composed of  $M$  *balanced* clusters. We discuss a reasonable heuristic in Section 1.4.2 to solving this optimization program, conjectured to be NP-hard.

## 1.3 Variations of the Experiment-of-Experiments design

### 1.3.1 Bernoulli randomization as an alternative assignment strategy

The completely randomized (CR) assignment is a well-understood assignment mechanism, which avoids degenerate cases where all units are assigned to treatment and control. However, experimentation platforms at major internet companies are rarely set up to run completely randomized experiments. Instead, these platforms run Bernoulli randomized (BR) assignments, which for large sample sizes, are intuitively equivalent. We provide a formal explanation as to why running a Bernoulli randomized assignment does not affect the validity of our test in practice: the variance of the difference-in-means estimator under the Bernoulli randomized mechanism and the completely randomized mechanism are equivalent up to  $O(1/N^2)$  terms.

**Theorem 5.** *Let BR be the re-randomized Bernoulli assignment, assigning units to treatment with probability  $p := n_t/N$  and to control with probability  $1 - p = n_c/N$ , rejecting the edge-assignments where all units are assigned to treatment or control. For all  $N \geq 2$  such that  $p^N + (1 - p)^N \leq 1/N^2$ , we have the following upper-bound:*

$$|\text{Var}_{\mathbf{Z} \sim \text{BR}}[\hat{\tau}] - \text{Var}_{\mathbf{Z} \sim \text{CR}}[\hat{\tau}]| \leq 5 \left( \frac{\sigma^2(\mathbf{Y}(1))}{n_t^2} + \frac{\sigma^2(\mathbf{Y}(0))}{n_c^2} \right)$$

A proof is included in Section A.8 in the Appendix. We did not seek to optimize the constant coefficient.

### 1.3.2 Stratification and subsampling considerations

One practical concern with our suggested hierarchical design is that if the chosen number of clusters is small, possibly much smaller than the number of units, we run the risk of having strong covariate imbalances between the two treatment arms. In this case, we recommend using a stratified treatment *arm* assignment. Let each graph cluster be assigned to one of  $L$  strata. Within each stratum  $s$ , we assign clusters completely at random to treatment arm  $cr(s)$  and treatment arm  $cbr(s)$ . Within each stratum  $s$ , units in treatment arm  $cr(s)$  are assigned completely at random to treatment and control, while clusters in treatment arm  $cbr(s)$  are assigned completely at random to treatment and control. Let  $\hat{\mu}_{cr}(s)$ ,  $\hat{\mu}_{cbr}(s)$ , and  $\Delta(s)$  be the restriction of  $\hat{\mu}_{cr}$ ,  $\hat{\mu}_{cbr}$ , and  $\Delta$  respectively to stratum  $s$  and  $M(s)$  be the total number of clusters in stratum  $s$ . The stratified  $\Delta'$  estimator can be expressed as an appropriately weighted average of the  $\Delta(s)$ :

$$\Delta' := \sum_{s=1}^L \frac{M(s)}{M} \Delta(s) \quad (1.22)$$

For a given empirical upper-bound  $\hat{\sigma}^2(s)$  of  $\text{Var}_{\mathbf{W}(s), \mathbf{Z}(s)}[\Delta(s)]$ , we define an empirical upper-bound  $\hat{\sigma}'^2$ :

$$\hat{\sigma}'^2 := \sum_{s=1}^L \left( \frac{M(s)}{M} \right)^2 \hat{\sigma}^2(s) \quad (1.23)$$

The results of Section 1.2 can easily be adapted to the stratified setting.

**Proposition 3.** *If SUTVA holds, and each cluster in  $\mathcal{C}$  has the same size, then the stratified empirical variance upper-bound upper-bounds the theoretical variance of the stratified difference-in-differences estimator in expectation:*

$$\mathbb{E}_{\mathbf{W}, \mathbf{Z}} [\hat{\sigma}'^2] \geq \text{Var}_{\mathbf{W}, \mathbf{Z}} [\Delta'] \quad (1.24)$$

An additional practical constraint is that online experimentation platforms need to run multiple experiments simultaneously, with multiple values of treatment and control. As a result, each experiment runs within a segment of the population chosen completely at random, leaving the other units available for other experiments. This subsampling, done completely at random at the unit-level, might negatively impact the quality of the clustering

in the *cbr* treatment arm. We therefore advise against subsampling in the cluster-based randomization arm, since not enough graph structure might be preserved post-subsampling. In other words, we recommend subsampling at the cluster-level, rather than at a unit-level, when deciding which units to include in the experiment.

## 1.4 Illustration on the *LinkedIn* experimentation platform<sup>2</sup>

Modern IT companies (e.g. Google (Tang *et al.*, 2010), Microsoft, (Kohavi *et al.*, 2013), Facebook (Bakshy *et al.*, 2014), LinkedIn (Xu *et al.*, 2015)) rely heavily on experimentation to understand the effect of each product decision, from minor UI changes to major product launches. Due to their extensive reliance on randomized experiments, these companies have each built mature experimentation platforms. It is an open question how many of these experiments violate SUTVA. By collaborating with the team in charge of LinkedIn’s experimentation platform, we were able to apply the previous theoretical framework to test for interference in one of LinkedIn’s many randomized experiments.

Users on LinkedIn interact with content posted by their “connections” through a personalized feed. Rather than presenting the content chronologically, LinkedIn’s feed ranks content by relevance. In order to improve the user experience, LinkedIn’s feed team seeks improvements to the ranking algorithm and to determine the impact of each modification on key user metrics through randomized experiments. Key user metrics include time spent on the site, engagement with content on feed, and original content creation.

Experimentation on feed ranking algorithms is a typical case where interference between units is present. If a user is assigned to a feed ranking algorithm that pushes more relevant content to the top than the default algorithm, they will interact more with their feed by liking, commenting on, or sharing more content. Each action can potential appear in their connection’s feed, thus affecting *their* user metrics. We seek to understand whether these network effects are significant.

---

<sup>2</sup>This section is the result of work done jointly with Martin Saveski.

### 1.4.1 Experimental set-up

To run the experiment, we

- (i) clustered the LinkedIn graph into balanced clusters (cf. Section 1.4.2),
- (ii) stratified the clusters by blocking on cluster covariates,
- (iii) assigned a subset of clusters to treatment and to control chosen at random. These clusters constitute the second treatment arm.
- (iv) Any unit not already assigned to treatment or control in step (iii) was given to the main experimentation pipeline: a sub-population of units is sub-sampled at random (cf. Section 1.3.2) and then assigned to treatment and control using a Bernoulli randomized assignment (cf. Section 1.3.1).

The number of units in each treatment arm was in the order of several million. Before applying treatment to units assigned to treatment, we ran covariate balance checks and measured outcomes 2 months prior to the experiment. The resulting outcomes can be found in Figure 1.3. As suggested in Section 1.3.2, we assigned each cluster to one of  $S$  strata in order to ensure balance of cluster-level covariates. We collected four covariates in each cluster: number of edges within the cluster, number of edges with an endpoint in another cluster, and two metrics that characterize users' engagement with the LinkedIn feed, averaged over all users in the cluster. We then stratified the clusters using balanced k-means clustering (Malinen and Fränti, 2014) to work best.

### 1.4.2 Clustering the *LinkedIn* graph

The main challenge of implementing the proposed test for interference is clustering the graph into clusters of equal size. In the last several years, there has been good progress in developing scalable balanced clustering algorithms for graphs with billions of edges (Tsourakakis *et al.*, 2014b; Aydin *et al.*, 2016). These algorithms have enabled practitioners to conduct large scale cluster-based randomized experiments (Ugander and Backstrom, 2013; Saveski *et al.*,

2017; Rolnick *et al.*, 2017). We performed an experimental evaluation of several of these balanced clustering algorithms.

### Comparison of clustering algorithms

Due to the scale of the LinkedIn graph—millions of nodes and billions of edges—we considered only streaming algorithms which were parallelizable while also enforcing balance:

- *METIS* (Karypis and Kumar, 1998) is a widely-used batch-clustering algorithm, and thus serves as our baseline to compare the quality of the clustering achieved by the streaming algorithms.
- *Balanced Label Propagation (BLP)* (Ugander and Backstrom, 2013) is an iterative algorithm that greedily minimizes edges-cut by solving a Linear Program to find an optimal relocation of nodes while maintaining balance
- *Restreaming Linear Deterministic Greedy (reLDG)* (Nishimura and Ugander, 2013) is a restreaming version of the Linear Deterministic Greedy (LDG) algorithm (Stanton and Kliot, 2012). LDG assigns each node  $i$  to a cluster  $j$  according to the following objective:

$$\arg \max_{j \in \{1, \dots, M\}} |\mathcal{C}_j^t \cap \mathcal{N}(i)| \left( 1 - \frac{|\mathcal{C}_j^t|}{H_j} \right), \quad (1.25)$$

where  $\mathcal{C}_j^t$  is the set of nodes assigned to cluster  $j$  at step  $t$  of the algorithm,  $H_j$  is the maximum capacity of cluster  $\mathcal{C}_j$  (usually set to  $\frac{N}{M}$  to achieve perfect balance), and  $\mathcal{N}(i)$  is the set of neighbors of node  $i$  in the graph. The first term maximizes the number of edges within clusters, while the second term enforces balance on the cluster sizes. Nishimura and Ugander then show that restreaming significantly increases the quality of the clusters compared to a single pass (Nishimura and Ugander, 2013).

- *Restreaming FUNNEL (reFUNNEL)* (Nishimura and Ugander, 2013) is a restreaming version of the FUNNEL algorithm (Tsourakakis *et al.*, 2014a), which is itself a streaming

Number of clusters ( $k$ )	BLP	reFENNEL	reLDG	METIS
100	26.7%	31.7%	<b>35.6%</b>	35.0%
300	22.7%	27.7%	<b>29.9%</b>	29.4%
500	-	26.1%	<b>27.7%</b>	27.0%
1000	-	23.9%	<b>24.7%</b>	23.8%

**Table 1.1:** Evaluation of the different balanced clustering algorithms. We report the percentage of within-cluster edges per clustering of the Netherlands LinkedIn graph. The values in bold represent the best performance. For BLP, we report results only for  $k = 100$  and  $k = 300$ , since the running times of one iteration for larger values of  $k$  were too long.

generalization of the modularity maximization. It assigns nodes to clusters as:

$$\arg \max_{i \in 1 \dots k} |\mathcal{C}_i \cap \mathcal{N}(u)| - \alpha |\mathcal{C}_i|,$$

where  $\alpha$  is a hyper-parameter. Unlike LDG, FUNNEL ensures only approximate balance, unless  $\alpha \geq \lceil \frac{N}{M} \rceil$ . Nishimura and Ugander (Nishimura and Ugander, 2013) suggest increasing  $\alpha$  in each restreaming pass to achieve best results. We run with linearly and logarithmically increasing schedules.

We first compared these algorithms on a smaller geographically well-isolated subset of the LinkedIn graph, the Netherlands, the results of which can be found in Table 1.1. We found the restreaming version of the Linear Deterministic Greedy algorithm (reLDG) to work best. To cluster the full LinkedIn graph, we ran the parallel version of reLDG, setting the number of clusters to  $k = \{1000, 3000, 5000, 7000, 10000\}$  and a leniency of 1% for the balance constraint, to slightly sacrifice balance for better clustering quality. In the experiment, we used the clustering obtained by setting  $k = 3000$  as it compromises between maximizing the fraction of edges within clusters (28.28%) and minimizing pre-intervention variance. The comparison of the clustering quality for various cluster sizes can be found in Table 1.2.

Number of clusters ( $k$ )	% of edges within clusters
1000	35.6
3000	28.5
5000	26.2
7000	22.8
10000	21.1

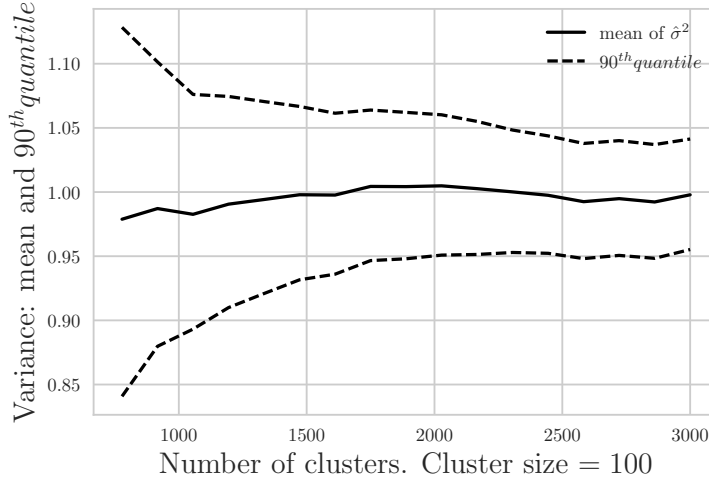
**Table 1.2:** Evaluation of different cluster-sizes for the ReLDG algorithm. We ran the parallel version of reLDG for 35 iterations.

### Simulation study

We ran two small-scale simulation studies to understand the effect of the clustering on the type I and type II error of our test.

Since we have a theoretical bound on the Type I error of our test under the assumption that our empirical upper-bound for the variance upper-bounds the theoretical variance for the realized assignment  $\hat{\sigma}^2(\mathbf{Z}) \geq \text{Var}_{\mathbf{Z}}[\Delta]$  — a property which is only guaranteed to hold true in expectation (cf. Theorem 2) — we examined this condition in the first simulation study. We fixed the cluster-size to 100 units but varied the number of clusters from 500 to 3000, effectively growing the graph from 50,000 to 300,000 units. In Figure 1.2, we report the expectation and the 10<sup>th</sup> and 90<sup>th</sup> quantiles of the ratio of the true variance of our estimator  $\text{Var}[\Delta]$  over 500,000 simulations for each value of the number of clusters. The upper-bound holds (tightly) in expectation but is not an upper-bound almost surely. However, despite the diminishing returns on confidence interval reduction from increasing the number of clusters, we see that for a number of cluster  $M \geq 2000$ , the upper-bound  $\hat{\sigma}^2$  concentrates 90% of the time in the  $[\cdot 95, 1.05] \times \text{Var}[\Delta]$  range.

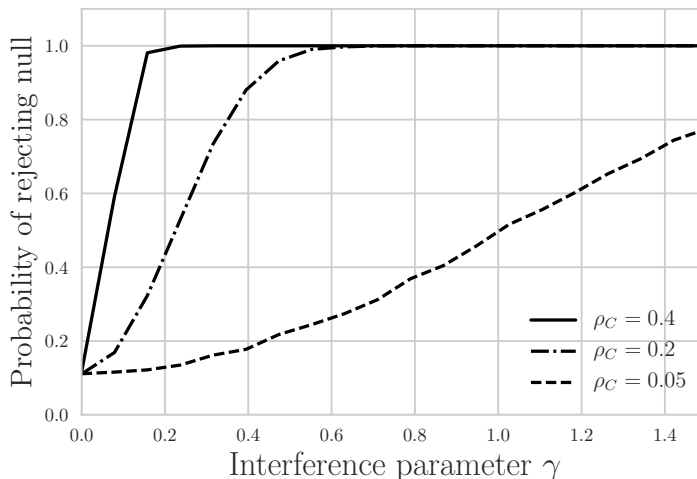
We examine the Type II error of our test in the second study. We considered a block-model graph with 40 balanced clusters of 1000 units. The edges of the graph were sampled such that the probability of an edge existing between two units in clusters  $\mathcal{C}_i$  and  $\mathcal{C}_j$  respectively is given by a constant cluster-level probability  $A_{ij}$  ( $= A_{ji}$ ). We chose  $(.01, .31)$ ,



**Figure 1.2:** Quality of the empirical variance upper-bound estimator as a function of graph size under SUTVA. We plot the expectation as well as the 10<sup>th</sup> and 90<sup>th</sup> quantiles of the ratio of the empirical upper-bound estimator  $\hat{\sigma}^2$  (cf. Eq 1.18) over the true variance  $\text{Var}[\Delta]$  as a function of graph size. We kept the cluster-size fixed to 100 units.

(.15, .45), and (.3, .6) as three tuples denoting the outer-diagonal and inner-diagonal probabilities, corresponding to values of the  $\rho_C$  parameter equal to (.05, .2, .4) respectively. Recall that  $\rho_C$  is a measure of graph cut quality — the higher  $\rho_C$  is, the fewer edges of the graph are cut. We varied the value of the interference parameter from 0 to 1.4 and computed the probability of rejecting the null under 1000 simulations. We assumed a linear interference model in Equation 1.9, fixing the value of the constant parameter to 0 ( $\alpha = 0$ ) and the direct treatment effect parameter to 1 ( $\beta = 1$ ).

We report the results in Fig. 1.3. Even with a low  $\rho_C$  coefficient of .05 inter-cluster edges, we correctly detect levels of interference that are of similar magnitude to the direct effect (equal to 1 here)  $\sim 75\%$  of the time. Furthermore, if  $\rho_C \geq 0.4$ , corresponding to a clustering which cuts fewer than 40% of all edges of a regular graph, we correctly detect levels of interference that are at least 1/5th of the magnitude of the direct treatment effect 99.9% of the time.



**Figure 1.3:** Impact of the interference parameter and clustering quality on the experimental power of the Experiment-of-Experiments test for interference. We plot the probability of rejecting the null hypothesis of SUTVA according to our suggested test for interference, for different values of  $\gamma$  and  $\rho_C$ . The interference model is taken from Eq. 1.9, with  $\alpha = 0$  and  $\beta = 1$ .  $\rho_C$  is the average number of edges cut, weighted by each node’s degree, as it appears in Eq. 1.21.

### 1.4.3 Experimental results

We launched our experimental design on the LinkedIn experimental platform, in August 2016. We considered a subset of the LinkedIn graph, containing several million users in treatment arm. We measured each user’s engagement with their LinkedIn feed at various points in time:  $y_i(t_{months})$ , which we expected to have strong interference effects. The primary outcome of interest was the change in a user’s engagement over time:  $Y_i(t_{months}) = y_i(t_{months}) - y_i(t_{months} - 2)$ , where  $t_{months} - 2$  takes place two months before date  $t$ . As a sanity check, we ran an A/A test on  $Y_{i,pre} = Y_i(-2) = y_i(-2) - y_i(-4)$ , where  $t = 0$  is the month the intervention was launched and  $t = -2$  takes place two months prior. As expected, we found that no significant interference in the A/A test, with a p-value of 0.68 using the gaussian assumption from Equation 1.20. We then evaluated the presence of interference on  $Y_i(2) = y_i(2) - y_i(0)$ , where  $t = 2$  takes place two months after the launch of the randomized experiment. We found a p-value less than 0.05 and hence conclude that interference is present in the graph. The results are reported in Table 1.3.

Statistic	Formula	Pre-treatment	Post-treatment
BR delta	$\sum_s \frac{M(s)}{M} \cdot \left( \overline{\mathbf{Y}_{br,t}(s)} - \overline{\mathbf{Y}_{br,c}(s)} \right)$	-0.016	0.107
CBR delta	$\sum_s \frac{M(s)}{M} \cdot \frac{m_{cbr}(s)}{n_{cbr}(s)} \cdot \left( \overline{\mathbf{Y}_{cbr,t}^+(s)} - \overline{\mathbf{Y}_{cbr,c}^+(s)} \right)$	-0.049	0.27
Delta of Deltas	$\sum_s \frac{M(s)}{M} \cdot \Delta(s)$	-0.033	-0.17
Upper bound of BR std	$\sqrt{\sum_s \left( \frac{M(s)}{M} \right)^2 \cdot \left( \frac{\hat{S}_{br,t}(s)}{n_{br,t}(s)} + \frac{\hat{S}_{br,c}(s)}{n_{br,c}(s)} \right)}$	0.028	0.022
Upper bound of CBR std	$\sqrt{\sum_s \left( \frac{M(s)}{M} \right)^2 \cdot \left( \frac{m_{cbr}(s)}{n_{cbr}(s)} \right)^2 \cdot \left( \frac{\hat{S}_{cbr,t}(s)}{m_{cbr,t}(s)} + \frac{\hat{S}_{cbr,c}(s)}{m_{cbr,c}(s)} \right)}$	0.076	0.082
Upper bound of std	$\sqrt{\sum_s \left( \frac{M(s)}{M} \right)^2 \cdot \hat{\sigma}^2(s)}$	0.081	0.085
2-tailed p-value		0.68	<b>0.048</b>

**Table 1.3:** The results of our suggested experiment-of-experiments test for interference on a subset of the LinkedIn graph. BR denotes the Bernoulli randomized treatment arm; CBR denotes the cluster-based randomized treatment arm. The final row displays the two-tailed p-value of the test statistic T under assumption of normality  $T \sim \mathcal{N}(0, 1)$ . Outcomes have been multiplied by an unspecified constant to avoid disclosing raw numbers. We reject the null of interference for the post-treatment outcomes, with a p-value of 0.048 < 0.05. As expected, we do not reject the null for the AA test on pre-experiment outcomes, with a p-value of 0.68

## 1.5 Conclusion

We have proposed an Experiment-of-Experiments test for interference, which we have validated on a live experiment on the LinkedIn platform. The test compares the outcome of a completely randomized assignment to a cluster-based randomized assignment. Our framework can easily be adapted to comparing different estimators or different designs altogether. Another possible improvement to the suggested design is to re-cluster the graph after the assignment of units to treatment arms, rather than re-using the initial clustering. Finally, further power analysis, such as exploring other parametrizations of interference, is beyond the scope of this chapter.

## Chapter 2

# Optimizing cluster-based randomized experiments

In this chapter, we tackle the second research question listed in the introduction, which we restate here:

*Can we determine which of two cluster-based randomized designs is most appropriate without explicit knowledge of the interference structure of the experiment?*

Recall that when SUTVA does not hold, a popular randomized design to mitigate interference is the cluster-based randomized design. A cluster-based randomized design assigns units in groups in the attempt to minimize the interaction between groups assigned to different values of the intervention. For a perfect clustering of units, with no interaction across groups, otherwise known as clusters, we recover many of the results stated under SUTVA.

In practice, however, such a grouping of units may not exist and A/B test practitioners often settle to find the best possible clustering. Finding the “best possible” clustering is often formulated as finding the balanced minimum cut of a weighted graph, where the nodes of the graph are the experimental units and the edges represent how liable two units are to interfere with one another. This is a challenging task, both algorithmically and empirically.

Algorithmically, clustering a graph into  $M$  balanced clusters is known to be NP-hard, even if we tolerate some unevenness in the clusters sizes (Andreev and Racke, 2006). Empirically, it is not always clear what the correct graph representation of the interference mechanism is: deciding which edge belongs to the graph and its edge weight is a non-trivial problem.

While the literature on finding balanced clustering of weighted graphs and analysing cluster-based randomized designs is extensive (Donner and Klar, 2004; Middleton and Aronow, 2011; Eckles *et al.*, 2017), there are relatively few prior works that tackle the question of determining which of two balanced clusterings is most appropriate without assuming that the exact structure of the interference mechanism is known. The objective of this paper is to show that we can, in fact, identify the better of two clusterings through experimentation under a specific assumption on the interference mechanism, which we call *monotonicity*.

We make the following contributions: we present an experiment-of-experiments design for comparing cluster-based randomized designs. We define a monotonicity assumption under which we can determine which clustering produces the least-biased estimates of the total treatment effect using this comparative design. We prove that pricing experiments in the context of ad exchanges verify this monotonicity assumption, and thus our framework applies to this illustrative example. In particular, we state results for the welfare of a single-item second-price auction and the Vickrey-Clarke-Groves auction in the positional ad setting. Finally, we report an empirical simulation study of our algorithms for a publicly-available dataset for online ads. While pricing experiments are done in the context of ad exchanges (AdE, 2018), we note that this chapter is a theoretical study of the subject and does not include any real treatments of ad campaigns.

In Section 2.1, we define the monotonicity assumption, describe the suggested experiment-of-experiments design, and propose a test for interpreting its results. In Section 2.2, we explain how this framework can be applied to a real-world setting, by showing that reserve-price experiments on advertising auctions are monotone. Finally, we validate these findings on a Yahoo! ad auction dataset in Section 2.3.

**Acknowledgements:** The results of this chapter are based on a research paper (Pouget-Abadie *et al.*, 2018), written in collaboration with Edoardo M. Airoidi, Vahab Mirrokni, and David C. Parkes.

## 2.1 An Experiment-of-Experiments design

In this section, we define the monotonicity assumption, introduce our experiment-of-experiments design, and suggest an approach to analysing its results. We will re-use much of the notation from Chapter 1.

Notably, let  $N$  be the number of experimental units,  $\mathbf{Y}$  the outcome vector, and  $\mathbf{Z}$  the assignment vector of units to treatment ( $Z_i = 1$ ) or control ( $Z_i = 0$ ). The estimand of interest is once again the *Total Treatment Effect* (TTE) (cf. Equation 1.3). We will use the notation  $\mathbb{E}_{\mathbf{Z} \sim \mathcal{C}}[X]$  to denote the expected value of estimator  $X$  under a cluster-based randomized design, which assigns the *clusters* of a clustering  $\mathcal{C}$  uniformly at random to treatment or control. We will use  $\mathcal{C}$ -CBR as a shorthand for such an assignment mechanism. Once again,  $M_t$  is the number of clusters assigned to treatment and  $M_c$  the number of clusters assigned to control.  $M = M_t + M_c$  is the total number of clusters.

Recall that the Horvitz-Thompson estimator (cf. Eq. 1.6) is an unbiased estimator of the Total Treatment Effect for any non-trivial  $\mathcal{C}$ -CBR assignment under SUTVA (cf. Equation 1.8). When SUTVA does not hold, unbiasedness is no longer guaranteed, and the Horvitz-Thompson estimator  $\hat{\tau}$  may be biased. Our objective is to minimize bias with respect to the clustering, without assuming any explicit knowledge of the interference mechanism or the estimand. This objective is summarized in the following equation:

$$\min_{\mathcal{C}} |\mathbb{E}_{\mathbf{Z} \sim \mathcal{C}}[\hat{\tau}] - TTE| \tag{2.1}$$

### 2.1.1 A monotonicity assumption

Choosing the clustering of our experimental units in a way that minimizes the bias of our estimators (cf. Eq. 2.1) when running a cluster-based experiment is a difficult task:

without the ground truth, we cannot observe the bias directly. However, under a specific monotonicity property— common to many randomized experiments —the task of choosing the better of two clusterings becomes straightforward.

**Definition 1.** For a domain  $\mathcal{P}$  of clusterings of our  $N$  units, we say that the interference model is  $\mathcal{P}$ -increasing if and only if

$$\forall \mathcal{C} \in \mathcal{P}, \mathbb{E}_{\mathbf{Z} \sim \mathcal{C}}[\hat{\tau}] \leq TTE, \quad (2.2)$$

and it is  $\mathcal{P}$ -decreasing if and only if

$$\forall \mathcal{C} \in \mathcal{P}, \mathbb{E}_{\mathbf{Z} \sim \mathcal{C}}[\hat{\tau}] \geq TTE \quad (2.3)$$

A model that is either  $\mathcal{P}$ -increasing or  $\mathcal{P}$ -decreasing for all clusterings of  $\mathcal{P}$  is  $\mathcal{P}$ -monotone.

A  $\mathcal{P}$ -monotone model is one for which the expectation of the Horvitz-Thompson estimator  $\hat{\tau}$  is either always a lower bound or always an upper-bound of the estimand under any  $\mathcal{C}$ -CBR design for  $\mathcal{C} \in \mathcal{P}$ . If a model is  $\mathcal{P}$ -increasing,  $\mathcal{P}$ -decreasing, or  $\mathcal{P}$ -monotone for the trivial set of all possible clusterings  $\mathcal{P}$ , then we simply say that the model is “increasing”, “decreasing”, or “monotone” without specifying  $\mathcal{P}$ . Before delving into examples of monotone interference mechanisms, we introduce the following proposition, which highlights why monotonicity is useful for reasoning about bias.

**Proposition 4.** If the interference model is  $\mathcal{P}$ -increasing, then for all  $\mathcal{C}_1, \mathcal{C}_2 \in \mathcal{P}$ , it holds that

$$\mathbb{E}_{\mathbf{Z} \sim \mathcal{C}_1}[\hat{\tau}] \leq \mathbb{E}_{\mathbf{Z} \sim \mathcal{C}_2}[\hat{\tau}] \implies |\mathbb{E}_{\mathbf{Z} \sim \mathcal{C}_1}[\hat{\tau}] - TTE| \geq |\mathbb{E}_{\mathbf{Z} \sim \mathcal{C}_2}[\hat{\tau}] - TTE| \quad (2.4)$$

If the interference model is  $\mathcal{P}$ -decreasing, then for all  $\mathcal{C}_1, \mathcal{C}_2 \in \mathcal{P}$ , it holds that

$$\mathbb{E}_{\mathbf{Z} \sim \mathcal{C}_1}[\hat{\tau}] \leq \mathbb{E}_{\mathbf{Z} \sim \mathcal{C}_2}[\hat{\tau}] \implies |\mathbb{E}_{\mathbf{Z} \sim \mathcal{C}_1}[\hat{\tau}] - TTE| \leq |\mathbb{E}_{\mathbf{Z} \sim \mathcal{C}_2}[\hat{\tau}] - TTE| \quad (2.5)$$

*Proof.* If the model is  $\mathcal{P}$ -increasing, for  $k \in \{1, 2\}$ , and  $\mathcal{C}_k \in \mathcal{P}$ ,

$$\mathbb{E}_{\mathbf{Z} \sim \mathcal{C}_k}[\hat{\tau}] - TTE = -|\mathbb{E}_{\mathbf{Z} \sim \mathcal{C}_k}[\hat{\tau}] - TTE| \quad (2.6)$$

Hence, the inequality sign is flipped when the model is  $\mathcal{P}$ -increasing. A similar reasoning

applies for  $\mathcal{P}$ -decreasing models.  $\square$

Proposition 4 is a simple consequence of Definition 1: if we know that two cluster-based estimates are both lower bounds of the estimand, then the greater of the two must be less biased. The same reasoning applies if they both upper-bound the estimand. It is sufficient to compare the expectation of our estimators to determine which is less biased.

The crux of our framework therefore hinges on monotonicity. Many commonly studied parametric models of interference are in fact monotone. Consider the following *linear model of interference*, generalized from Equation 1.9 in Chapter 1. Let  $(\alpha_i, \beta_i, \gamma) \in \mathbb{R}^3$ , the outcome of unit  $i$  is given by:

$$\forall i, Y_i(\mathbf{Z}) = \alpha_i + \beta_i Z_i + \gamma \rho_i + \epsilon_i, \quad (2.7)$$

where  $\rho_i = \frac{1}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} Z_j$  is the proportion of  $i$ 's neighborhood that is treated and  $\epsilon_i \sim \mathcal{N}(0, 1) \perp \rho_i$ . Equation 2.7 expresses each unit's outcome as a linear function of a fixed effect, a heterogeneous treatment effect, and a network effect proportional to the fraction of  $i$ 's neighborhood that is treated. As shown in the following proposition, this interference model is monotone.

**Proposition 5.** *For a given clustering  $\mathcal{C}$ , let  $\rho_{\mathcal{C}} = \frac{1}{N} \sum_i \frac{|\mathcal{N}_i \cap \mathcal{C}(i)|}{|\mathcal{N}_i|}$  be the average proportion of unit  $i$ 's neighborhood  $\mathcal{N}_i$  that is included in its assigned cluster  $\mathcal{C}(i)$ . Then,*

$$TTE - \mathbb{E}_{\mathbf{Z} \sim \mathcal{C}}[\hat{\tau}] = \frac{\gamma M}{M-1} (1 - \rho_{\mathcal{C}}) \quad (2.8)$$

*It follows that if  $\gamma \geq 0$ , the interference model is increasing, otherwise it is decreasing.*

We can also extend the above for heterogeneous network effect parameters  $\gamma_i$ . A proof can be found in Section B.1 of the Appendix.

**Proposition 6.** *For a clustering  $\mathcal{C}$ , let  $\rho_{\mathcal{C},i} = \frac{|\mathcal{N}_i \cap \mathcal{C}(i)|}{|\mathcal{N}_i|}$ . For all possible clusterings  $\mathcal{C}$ ,*

$$TTE - \mathbb{E}_{\mathbf{Z} \sim \mathcal{C}}[\hat{\tau}] = \frac{M}{N(M-1)} \sum_i \gamma_i (1 - \rho_{\mathcal{C},i}) \quad (2.9)$$

*It follows that if  $\sum_i \gamma_i (1 - \rho_{\mathcal{C},i}) \geq 0$ , then the interference model is increasing, otherwise it is decreasing.*

It follows that if  $\gamma_i \geq 0, \forall i$ , then the interference mechanism is increasing, and if  $\gamma_i \leq 0, \forall i$ , then it is decreasing. If the sign of  $\gamma_i$  is not consistent, then the monotonicity depends on the clustering: if all units with a given sign are perfectly clustered ( $\rho_{C,i} = 1$ ), e.g. all units with  $\gamma_i \geq 0$ , then the mechanism is once again monotone. For complex interference mechanisms, it can be easier to establish the following sufficient (but not necessary) condition:

**Proposition 7.** *We say an interference mechanism verifies the self-excitation property for a set of clusterings  $\mathcal{P}$ , if for all units  $i$  and clustering  $C \in \mathcal{P}$ ,*

$$\mathbb{E}_{\mathbf{Z} \sim C}[Y_i(\mathbf{Z}) : Z_i = 0] \geq Y_i(\mathbf{0}) \quad (2.10)$$

$$\mathbb{E}_{\mathbf{Z} \sim C}[Y_i(\mathbf{Z}) : Z_i = 1] \leq Y_i(\mathbf{1}) \quad (2.11)$$

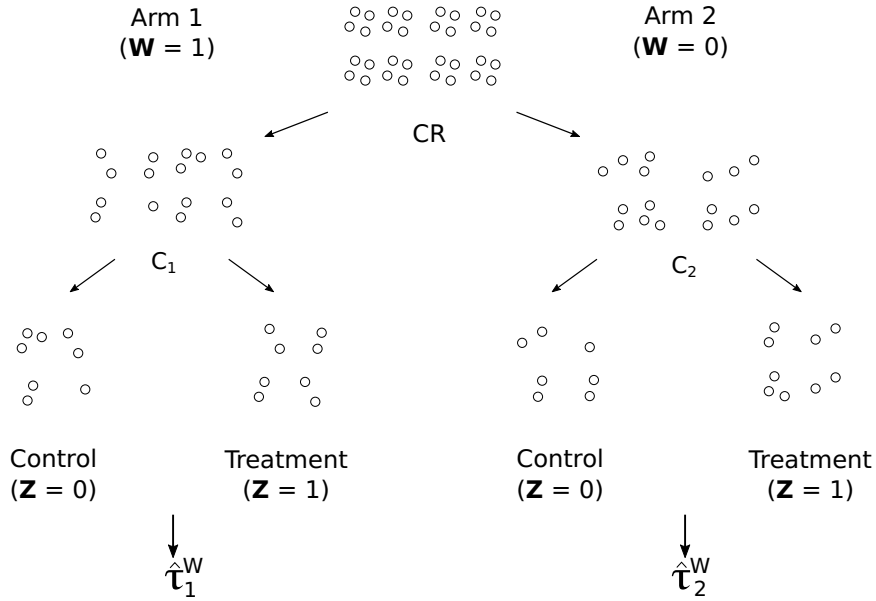
*A  $\mathcal{P}$ -self-exciting process is  $\mathcal{P}$ -increasing. A  $\mathcal{P}$ -self-deexciting mechanism, with flipped inequalities, is  $\mathcal{P}$ -decreasing.*

The proof is included in Section B.2 of the Appendix. The two inequalities capture the following phenomenon: conditioned on  $i$ 's treatment status, if  $i$ 's outcome is greatest when  $i$ 's neighborhood is entirely in treatment, and lowest when  $i$ 's neighborhood is entirely in control, then an experiment always under-estimates the true treatment effect. This only needs to be true in expectation over the assignments  $\mathbf{Z}$ . For example, we will show that the interference mechanism present in certain reserve price experiments in an advertiser auction setting is self-exciting (cf. Section 2.2).

We say the interference mechanism is self-exciting because these inequalities are verified when units benefit from being surrounded by units in treatment. A successful messaging feature launch is a straightforward example of a self-exciting process, as is any pricing mechanism that penalizes any treated bidders and boosts the utility of their competitors.

### 2.1.2 Design and analysis

Under monotonicity, Proposition 4 states that we can determine the least-biased of two  $\mathcal{P}$ -increasing or  $\mathcal{P}$ -decreasing cluster-based designs, without knowledge of the estimand, by comparing the expectation of their estimates. However, only one cluster-based design



**Figure 2.1:** A hierarchical experimental design, which assigns the experimental units to one of two cluster-based randomized designs,  $C_1$  and  $C_2$ , completely at random (CR).  $\hat{\tau}_1^W$  and  $\hat{\tau}_2^W$  represent the treatment effect estimates under each design respectively.

can ever be applied to the set of experimental units in its entirety, and the comparison of  $\mathbb{E}_{\mathbf{Z} \sim C_1}[\hat{\tau}]$  with  $\mathbb{E}_{\mathbf{Z} \sim C_2}[\hat{\tau}]$  cannot be done directly.

This resembles the problem we faced in Chapter 1 of causal inference: we wish to compare two randomized designs but can only run the experiment once. Much like in the Experiment-of-Experiments design introduced in Section 1.1, we suggest to randomly assign different units to either clustering, resulting in a 2-step hierarchical randomized design. The procedure, described in pseudo-code in Algorithm 2, is as follows:

- Assign units completely at random to two treatment arms, one for each clustering. Let  $\mathbf{W} \in \{1, 2\}^N$  be the vector representing that assignment.
- Within each treatment arm, cluster the remaining units together according to the appropriate cluster: if  $W_i = W_j = k$  and  $C_k(i) = C_k(j)$ , then  $i$  and  $j$  belong to the same cluster in treatment arm  $k \in \{1, 2\}$ . The resulting clusterings are  $C_1^W$  and  $C_2^W$ .
- Within each treatment arm, assign the resulting clusters to treatment and control

uniformly at random. Let  $\mathbf{Z}$  be the resulting assignment vector. This is possible because no unit belongs to both  $\mathcal{C}_1^{\mathbf{W}}$  and  $\mathcal{C}_2^{\mathbf{W}}$ .

---

**Algorithm 2:** Comparing cluster-based randomized assignments

---

Choose  $\mathbf{W} \in \{1, 2\}^N$  uniformly at random, encoding the assignment of units to treatment arms 1 and 2;  
**for**  $k \in \{1, 2\}$  **do**  
    Let  $\mathcal{C}_k^{\mathbf{W}}$  be the clustering on  $\{i \in [1, N] : W_i = k\}$  such that  
     $\mathcal{C}_k^{\mathbf{W}}(i) = \mathcal{C}_k^{\mathbf{W}}(j)$  iff  $C_k(i) = C_k(j)$ ;  
    Assign units in treatment arm  $k$  to treatment and control with a  $\mathcal{C}_k^{\mathbf{W}}$ -cluster-based design;  
**end**  
**return** the resulting assignment vector  $\mathbf{Z}$ ;

---

Algorithm 2 provides us with two estimates,  $\hat{\tau}_1^{\mathbf{W}}$  and  $\hat{\tau}_2^{\mathbf{W}}$ , of the causal effect, one from each treatment arm. The resulting clusterings  $\mathcal{C}_1^{\mathbf{W}}$  and  $\mathcal{C}_2^{\mathbf{W}}$  may be unbalanced. This is of minor importance as the Horvitz-Thompson estimator (cf. Eq. 1.6) is unbiased under SUTVA for unbalanced clusterings, and balancedness is required only to control its variance. In practice,  $\mathcal{C}_1$  and  $\mathcal{C}_2$  are not required to have the same number of clusters, but we expect the clusters sizes to be large enough for each cluster to have at least one unit in each treatment arm after the first stage with high probability.

From the comparison of  $\hat{\tau}_1^{\mathbf{W}}$  and  $\hat{\tau}_2^{\mathbf{W}}$ , we seek to order  $\mathbb{E}_{\mathbf{Z} \sim \mathcal{C}_1}[\hat{\tau}_1]$  and  $\mathbb{E}_{\mathbf{Z} \sim \mathcal{C}_2}[\hat{\tau}_2]$ . Under arbitrary interference structures, these proxy estimates are not guaranteed to have the same ordering, the key condition for Proposition 4. Intuitively,  $\hat{\tau}_1^{\mathbf{W}}$  and  $\hat{\tau}_2^{\mathbf{W}}$  represent the treatment effect estimates for two “weakened” versions of each clustering  $\mathcal{C}_1$  and  $\mathcal{C}_2$ . Because the assignment of units to treatment arms is done completely at random, it affects each clustering in the same way, and we expect the ordering to stay the same. For the linear model of interference in Prop. 6, we have:

**Property 1.** *An interference mechanism is said to be  $\mathcal{P}$ -transitive if  $\forall \mathcal{C}_1, \mathcal{C}_2 \in \mathcal{P}$ ,*

$$\mathbb{E}_{\mathbf{W}, \mathbf{Z} \sim \mathcal{C}_1^{\mathbf{W}}} [\hat{\tau}_1^{\mathbf{W}}] \leq \mathbb{E}_{\mathbf{W}, \mathbf{Z} \sim \mathcal{C}_2^{\mathbf{W}}} [\hat{\tau}_2^{\mathbf{W}}] \Leftrightarrow \mathbb{E}_{\mathbf{Z} \sim \mathcal{C}_1} [\hat{\tau}] \leq \mathbb{E}_{\mathbf{Z} \sim \mathcal{C}_2} [\hat{\tau}] \quad (2.12)$$

If an interference mechanism is transitive for all possible clusterings  $\mathcal{P}$ , we simply say

that it is “transitive” without specifying  $\mathcal{P}$ . As a sanity check, we can also confirm that the property holds for SUTVA. The property can also be shown for the linear interference mechanisms introduced in Prop. 6:

**Proposition 8.** *Under SUTVA, for all  $\mathcal{C}_1, \mathcal{C}_2$  and  $k \in \{1, 2\}$ , it holds that*

$$\mathbb{E}_{\mathbf{W}, \mathbf{Z} \sim \mathcal{C}_k^{\mathbf{W}}} [\hat{\tau}_k^{\mathbf{W}}] = \mathbb{E}_{\mathbf{Z} \sim \mathcal{C}_k} [\hat{\tau}] = TTE. \quad (2.13)$$

Hence, the no-interference case is trivially transitive. Furthermore, the linear model of interference in Prop. 6 is transitive if the same number of units is assigned to each treatment arm in the first stage of the experiment-of-experiment design:  $\sum[W_i = 1] = \frac{N}{2}$ .

A full proof can be found in Section B.3 of the Appendix. For more complex mechanisms of interference, as is the case for reserve price experiments, we use simulations to confirm the intuition that transitivity holds (cf. Section 2.3).

As is common with A/B tests, we do not have access to the expectation of our estimators, and rely on approximations to the variance, such as Neyman’s variance estimator. In order to meaningfully compare the estimates we obtain, we must apply our method of choice to determine when their ordering is significant. For example, we can make a normal approximation to the distribution of the estimates — using Neyman’s estimator to upper-bound the variance — to estimate the probability that one estimate is greater than the other with a certain significance level:

**Proposition 9.** *Let  $\mathcal{C}_1, \mathcal{C}_2$  be two clusterings in  $\mathcal{P}$ . For  $k \in \{1, 2\}$ , recall the definition of the Neymanian variance estimator for cluster-based randomized designs:*

$$\hat{\sigma}_k^{\mathbf{W}} = \frac{M_k}{N_k} \left( \frac{\hat{S}_{k,t}}{M_{k,t}} + \frac{\hat{S}_{k,c}}{M_{k,c}} \right), \quad (2.14)$$

where  $M_k$  (resp.  $N_k$ ) is the number of clusters (resp. units) in  $\mathcal{C}_k^{\mathbf{W}}$ ,  $\hat{S}_{k,t} = \text{var}\{Y_{j,k}^+ : z_j = 1\}$  and  $\hat{S}_{k,c} = \text{Var}\{Y_{j,k}^+ : z_j = 0\}$ . Recall that  $Y_{j,k}^+ = \sum_{\mathcal{C}_k^{\mathbf{W}}(i)=j} Y_i$  is the aggregated outcome for cluster  $j$  in arm  $k$ . Assume that the interference mechanism is transitive and  $\mathcal{P}$ -increasing. If  $\alpha$  is the level of

significance chosen, we state that  $C_1$  is a significantly better clustering than  $C_2$  if and only if

$$\Phi \left( \frac{\hat{\tau}_1^{\mathbf{W}} - \hat{\tau}_2^{\mathbf{W}}}{\sqrt{\hat{\sigma}_1^{\mathbf{W}} + \hat{\sigma}_2^{\mathbf{W}}}} \right) < \alpha, \quad (2.15)$$

where  $\Phi$  is the cumulative distribution function of the normal distribution.

A similar reasoning applies to  $\mathcal{P}$ -decreasing mechanisms. If the Gaussian approximation is not appropriate, the distribution of the estimators can equally be approximated by a bootstrap analysis, or a more sophisticated model-based imputation method (Imbens and Rubin, 2015). More details can be found in Section B.4.

## 2.2 Application to reserve price experiments

Online advertising exchanges provide an interface for bidders to participate in a set of auctions for advertising online. These ads can appear within the company’s own content, in a social feed, below a search query, or on the webpage of an affiliated publisher. These auctions provide the vast majority of revenue to these platforms, and are thus the subject of experimentation and optimization. Platforms run experiments and monitor different metrics including of revenue and estimates of bidders’ welfare.

One possible parameter subject to optimization is the method of determining reserve prices. Online marketplaces can choose to implement a reserve price, which sets the minimum bid required for a bid to be valid and compete with others. It may vary from bidder to bidder, and from auction to auction. A higher reserve may improve revenue, but if it is too high, then too many bids are discarded and ad opportunities can go unsold.

Modifications to a reserve price rule are prime examples of experiments where SUTVA does not hold. A change in reserve price to one bidder affects the bidding problem facing another bidder, even when her reserve is unchanged (e.g., reducing competition when the reserve to the first bidder is higher). We establish conditions under which the resulting interference mechanism within reserve price experiments is monotone, both in the case of a single-item second price auction setting and in the Vickrey-Clarke-Groves auction setting

for positional ads (Varian and Harris, 2014).

### 2.2.1 Single-item second price auctions

We consider a single-item second-price auction with  $N$  bidders  $B = \{B_i\}_{i \in N}$  without budget constraints: the highest bidder wins the auction and is charged the maximum of her reserve price and the second-highest bid. The second price auction is truthful (bidding true values is a dominant-strategy equilibrium), and we will assume that the bidders are rational.

Consider two reserve price mechanisms  $(r_i)_{i \in B}$  (control) and  $(r'_i)_{i \in B}$  (treatment). Suppose that the reserve price mechanism corresponding to treatment always sets a higher reserve price than the reserve price mechanism corresponding to control:  $\forall i, r'_i > r_i$ . By symmetry, the following argumentation would also work if the treatment and control labels were switched.

We suppose the bidders have values  $(v_i)$  for winning the auction. We randomly assign bidders to either the treatment or control reserve price mechanism, with  $\mathbf{Z}$  the resulting assignment. The chosen metric of interest is a bidder's utility, denoted by  $Y_i(\mathbf{Z})$ . For a second-price auction,  $Y_i = 0$  if bidder  $i$  does not win the auction, and  $Y_i = v_i - p$  when she wins the auction and pays price  $p$ . The bidder welfare of an auction is the sum of each bidder's utility,  $\sum_i Y_i(\mathbf{Z})$ , and the estimand is given by:

$$S = \sum_i Y_i(\mathbf{1}) - \sum_i Y_i(\mathbf{0}) \quad (2.16)$$

The reserve price experiment for second price auctions verifies the self-excitation property (cf. Prop. 7). The idea is that assigning a unit to the intervention can only make them less competitive by discarding their bid from the auction. Thus, the higher the number of treated units, the lower the competition for the remaining bidders, and the higher their utility.

**Theorem 6.** *Consider a set of rational agents with no budget-constraints. Let the outcome of interest be each agent's welfare. The interference mechanism of a reserve price experiment, assigning treated units to a higher personalized reserve price, for a single-item second-price auction is self-exciting,*

and thus monotone.

*Proof.* Consider bidder  $i$ 's outcome under  $\mathbf{Z} = \mathbf{0}$  and under any assignment  $\mathbf{Z}'$  such that  $Z_i = 0$ . There are three possible cases:

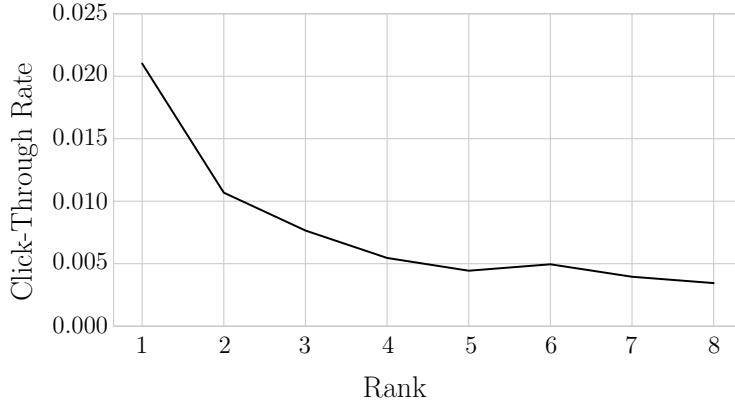
- Bidder  $i$  wins the auction in neither assignment. Her utility is therefore constant.
- Bidder  $i$  wins the auction in only one assignment. It must be that bidder  $i$  wins under  $\mathbf{Z}'$  but not  $\mathbf{Z}$ . Her utility is 0 under  $\mathbf{Z}$  and greater than 0 under  $\mathbf{Z}'$ .
- Bidder  $i$  wins the auction under both assignments. If the second highest bid is the same under both assignments, bidder  $i$ 's utility is constant. Otherwise, the second highest bid under  $\mathbf{Z}'$  can only be lower than the second highest bid under  $\mathbf{Z}$ . Thus bidder  $i$ 's payment is lower and her utility is higher under assignment  $\mathbf{Z}'$  than under assignment  $\mathbf{Z}$ .

By symmetry, we reach a similar conclusion when comparing assignments  $\mathbf{Z} = \mathbf{1}$  and any assignment  $\mathbf{Z}'$  such that  $Z'_i = 1$ . □

It follows that the reserve price experiment is increasing, and any cluster-based randomized design underestimates the bidder welfare estimand.

## 2.2.2 Positional ad auctions

In practice, ad auctions are multi-item, used for selling more than one ad position on a user's view. We now extend the previous results to a multi-item setting, with  $m$  items (or "slots"). We assume the common positional ad setting, where each slot has an inherent click-through rate  $pos_j$ , which we can suppose is ordered:  $pos_1 > pos_2 > \dots > pos_m$  (Varian, 2007). Each bidder  $i$  is only ever allocated at most one item, with value  $v_i$  for getting a click. We assume for simplicity that all bidders have the same ad quality, and thus the same click-through rate for a given ad slot. As a result, bidder  $i$ 's utility for winning slot  $j$  is  $v_i \cdot pos_j - p_i$ , where  $p_i$  is the required payment of bidder  $i$ .



**Figure 2.2:** The average click-through rate (CTR) observed in the Yahoo! Search Auction dataset, described in Section 2.3, can be observed to be an approximately decreasing and convex function of the slot rank. The confidence intervals were too small to be meaningfully reported in the figure.

The Vickrey-Clarke-Groves (VCG) auction takes place in two parts. First, a value-maximising allocation is chosen (based on bids). Here, the highest bids win the highest slots. Bidders are then charged the externality they impose on all other bidders. In other words, assuming that bidder  $k$  obtains the  $k^{\text{th}}$  slot, bidder  $k$  pays:

$$p_k = \sum_{j=k+1}^m (pos_{j-1} - pos_j) \cdot v_j \cdot \mathbf{1}_{[v_j \geq r_j]}$$

where  $r_j$  is the reserve imposed on bidder  $j$  with value  $v_j$ . We can prove that the self-excitation property holds under a convexity assumption.

**Theorem 7.** Consider a set of rational agents with no budget-constraints. Let the outcome of interest be each agent’s welfare. The interference mechanism of a reserve price experiment, assigning treated units to a higher personalized reserve price, for a VCG auction in the positional ad setting with no quality effects is self-exciting, and thus monotone if the click-through rate function  $pos : i \mapsto pos_i$  is convex:

$$\forall i > j, pos_{i+1} - pos_i \leq pos_{j+1} - pos_j,$$

This convexity assumption is verified empirically in the literature and in the Yahoo! auction dataset<sup>1</sup> introduced in Section 2.3 (cf. Figure 2.2). The intuition behind the proof is

<sup>1</sup>Our own dataset could potentially suffer from endogeneity, where weaker bidders are consistently assigned

similar to the single-item setting: for a bidder  $i$ , the greater the number of  $i$ 's competitors are treated, the fewer are able to compete, and thus the higher  $i$ 's utility. We prove this through a case-by-case analysis. Let  $r_i^Z$  be the reserve that bidder  $k$  faces under assignment vector  $Z$ :  $r_i^Z = r_i$  if  $Z_i = 0$  and  $r_i'$  otherwise.

*Proof.* Consider the outcomes of bidder  $i$  and  $j$  under  $\mathbf{Z}$  and  $\mathbf{Z}'$  such that for all  $k \neq j$ ,  $Z_k = Z'_k$ ,  $Z_i = Z'_i = 0$ , and  $Z_j = 0 < Z'_j = 1$ . By transitivity, if we can show  $Y_i(\mathbf{Z}) \leq Y_i(\mathbf{Z}')$ , then it follows that  $Y_i(\mathbf{0}) \leq \mathbb{E}_{\mathcal{C}}[Y_i(\mathbf{Z}) : Z_i = 0]$ . There are three possible cases:

- The allocation of bidders to slots does not change and thus prices do not change. Bidder  $i$ 's utility is constant.
- Bidder  $i$  is allocated to slot  $i$  for both  $\mathbf{Z}$  and  $\mathbf{Z}'$  assignments, but bidder  $j$ 's ( $j < i$ ) bid is discarded when  $j$  is treated ( $\mathbf{Z}'$ ):  $r_j' > v_j > r_j$ . The difference of bidder  $i$ 's outcome under the two treatment assignments is:

$$Y_i(\mathbf{Z}) - Y_i(\mathbf{Z}') = - \sum_{k \geq j} (pos_{k-1} - pos_k) (v_k \mathbf{1}_{v_k > r_k^Z} - v_{k+1} \mathbf{1}_{v_{k+1} > r_{k+1}^Z})$$

This quantity is always negative, hence  $Y_i(\mathbf{Z}) \leq Y_i(\mathbf{Z}')$ .

- Bidder  $j$ 's ( $j < i$ ) bid is discarded when  $j$  is treated and thus bidder  $i$  is allocated to slot  $i - 1$ . In that case, bidder  $i$ 's utility under  $\mathbf{Z}$  is:

$$Y_i(\mathbf{Z}) = pos_i v_i - \sum_{k \geq i+1} (pos_{k-1} - pos_k) v_k \mathbf{1}_{v_k > r_k^Z}$$

The same bidder  $i$ 's utility under  $\mathbf{Z}'$  is:

$$Y_i(\mathbf{Z}') = pos_{i-1} v_i - \sum_{k \geq i+1} (pos_{k-2} - pos_k) v_k \mathbf{1}_{v_k > r_k^Z}$$

It follows that the difference of bidder  $i$ 's outcomes is equal to:

$$Y_i(\mathbf{Z}) - Y_i(\mathbf{Z}') = (pos_i - pos_{i-1}) v_i - \sum_{k \geq i+1} (pos_{k-2} + pos_k - 2pos_{k-1}) v_k$$

---

to lower slots. The assumption is, however, supported elsewhere in the literature (Brooks, 2004; Richardson *et al.*, 2007).

Per keyphrase			Per bidder		
number of bids	min	1	number of bids	min	1
	median	2		median	9
	max	7041		max	$2.1 \cdot 10^4$
bid value	min	.3¢	bid value	min	.5¢
	median	66¢		median	60¢
	max	\$320		max	\$4700
impressions	min	1	impressions	min	1
	median	3		median	31
	max	$5 \cdot 10^6$		max	$1.4 \cdot 10^6$
clicks	min	0	clicks	min	0
	$cdf(1)$	91.4		$cdf(1)$	93.3
	max	7041		max	$1.1 \cdot 10^4$

**Table 2.1:** Summary statistics for the Yahoo! dataset, aggregated by keyphrase or by bidder, per day for the entire 4 month period. Bid values are given in USD unless specified otherwise.  $cdf(1)$  is the value of the cumulative distribution function of impressions for a single impression.

where the  $\mathbf{1}_{v_k > r_k^z}$  terms are implicit. Each individual term of the sum is positive by convexity, such that  $Y_i(\mathbf{Z}) \leq Y_i(\mathbf{Z}')$ .

□

## 2.3 Illustration on the Yahoo! bid dataset

In this section, we validate our design strategy for comparing two given graph clusterings for the purpose of experimentation under interference to an advertising auction dataset. For this purpose, we make use of a Yahoo! auction dataset.

The *Yahoo! Search Marketing Advertiser Bid-Impression-Click data on competing Keywords* dataset is a publicly-available dataset released by Yahoo!<sup>2</sup>, containing bid, impression, click, and revenue data between advertiser-keyphrase pairs over a period of 4 months. The

<sup>2</sup>Available for download at <https://webscope.sandbox.yahoo.com/>

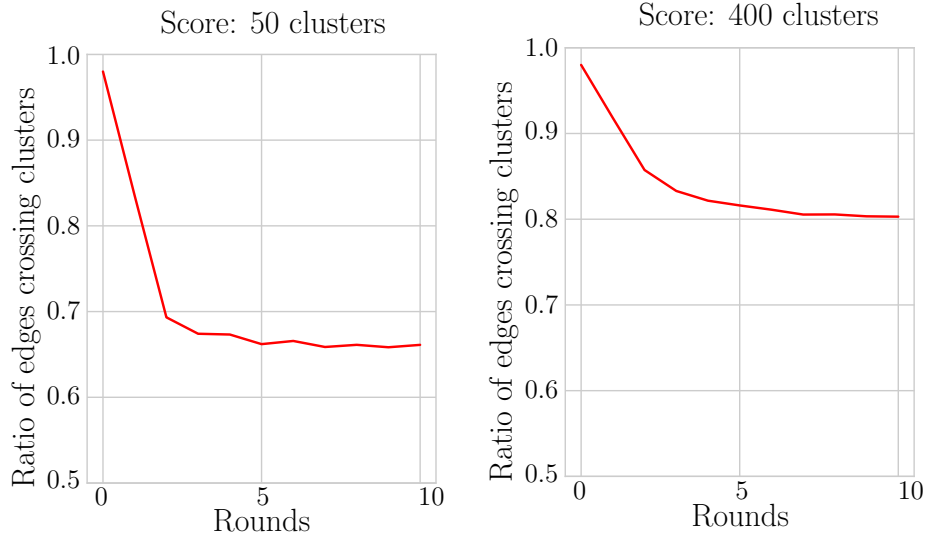
advertiser and keyphrase are anonymized, represented as a randomly-chosen string. A sample line of the dataset is reproduced<sup>3</sup> below:

day	id	rank	keyphrase	bid	impress.	clicks
1	a3d2	2	f3e4, j6r3, ...	100.0	1.0	0.0

The dataset contains 77,850,272 bidding activities of 16,268 different bidders. There are a total of 75,359 keywords represented, for a total of unique 648,515 keyphrases (or list of keywords). Table 2.1 contains a series of summary statistics computed over keyphrase-day pairs and bidder-day pairs, namely the total number of bids, the total bid value, the total number of impressions, and the total number of clicks per keyword (or per bidder) and per day.

We can represent the *Yahoo!* dataset by a set of bipartite graphs between bidders, identified by their `account_id`, and the keyphrases. The *bid* bipartite graph on day  $t$  draws a weighted edge of weight  $w_{ij}$  between every bidder-keyphrase pair such that bidder  $i$  bids  $w_{ij}$  on keyphrase  $j$  on day  $t$ . We can aggregate these graphs over the entire time period (4 months) by summing their edge weights together. We can also consider the impression, rank, and clicks graphs, where the weight of the edge is given by the number of impressions, the rank, or the number of clicks respectively received by bidder  $i$  on keyphrase  $j$ .

The dataset only provides data aggregated at the granularity of a single day, reporting the average bid and total number of impressions and clicks for each bidder, keyphrase day triplet. Hence, we define a keyphrase-day pair as a single auction, where each bidder's bid is set to the reported average bid for that keyphrase-day pair. For the sake of simplicity, we will only consider a setting with the top four ad positions, which account for the majority of clicks.



**Figure 2.3:** Weighted ratio of edges across clusters for successive runs of the reLDG algorithm on the weighted bid graph into 50 clusters and 400 clusters respectively.

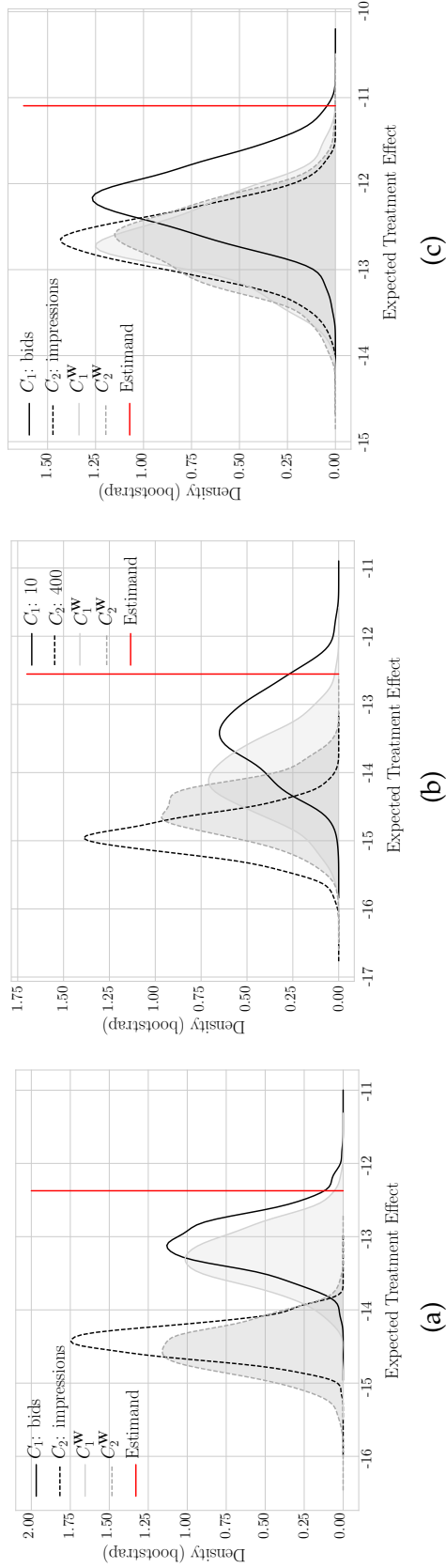
### 2.3.1 Simulating a reserve price experiment

While the *Yahoo! Search Auction* dataset provides us with a set of bidders, keyphrases, and the bids, impressions, and clicks that link them, it does not provide us with an actual intervention on the auction ecosystem. We must therefore simulate the impact of a change in the reserve price given to each bidder.

While many possible units of randomization exist for an auction experiment (keyphrases, bidders, browsers, users, various pairings of these units, etc.), the reserve price experiment we consider randomizes on bidders. On large auction platforms, the reserve price might be set through the application of machine learning methods. In our context, we choose a random non-zero reserve price for each bidder, calibrating the spread of the distribution such that some bidders will not be able to match the reserve price for all auctions. All bidders assigned to the intervention will face their non-zero reserve price, fixed for every auction for simplicity. All bidders assigned to the control bucket will not face a reserve price.

---

<sup>3</sup>The account ID and keyword ID's have been shortened for the sake of exposition in this sample line. The bid value is given in 1/100¢.



**Figure 2.4:** We compare the distribution of the expectation of our Horvitz-Thompson estimator for the total treatment effect (in red) under several cluster-based randomized assignments. In each plot, the solid and dotted lines represent the expectation of the estimator under  $C_1$  and  $C_2$  respectively — the two estimate distributions we wish to compare but cannot simultaneously observe. The shaded distributions correspond to the observed distributions of the expectation of the estimator under the induced clusterings  $C_1^W$  and  $C_2^W$ , resulting from our Experiment-of-Experiments design. The red segment represents the total treatment effect estimand. Each plot establishes a comparison of two different clusterings: (a)  $C_1$  is a reLDG clustering,  $C_2$  is a random clustering ( $M_1 = M_2 = 50$ ); (b)  $C_1$  is a reLDG clustering with 10 clusters,  $C_2$  is a reLDG clustering with 400 clusters; (c)  $C_1$  is a reLDG clustering of the bid graph, whereas  $C_2$  is a reLDG clustering of the impressions graph. ( $M_1 = M_2 = 50$ ). Monotonicity is verified because every distribution is on the same side of the estimand; transitivity is verified because the ordering of the solid and dotted distributions is preserved when going from the unshaded plots to the shaded plots. The loss of power is quantified by the increase in overlap between the solid and dotted distributions, when comparing the unshaded plots with the shaded plots.

Within the same auction for a given keyphrase, two participating bidders may face distinct reserves and be assigned to different treatment buckets. A bidder-cluster-based randomized experiment is thus used to mitigate the possible interference between bidders, our units of randomization, within a single auction.

To validate our experiment-of-experiments design, we must find candidate balanced graph clusterings to compare, a problem known to be NP-hard — even when we slightly relax the balancedness assumption (Andreev and Racke, 2006).

In Chapter 1, we briefly reviewed scalable balanced clustering algorithms. Of the numerous algorithms for finding such clusterings, the *Restreaming Linear Deterministic Greedy* (reLDG) algorithm (Nishimura and Ugander, 2013) is a popular choice, which was effective for running our *LinkedIn* experiments. We can apply this clustering algorithm to any of the bipartite graphs introduced in Section 2.2, aggregated over the entire time period, resulting in a set of mixed bidder-keyphrase clusters. The bidder-only clusters are obtained from the previous clustering by simply removing the keyphrase nodes from consideration.

Recall the algorithm’s objective from Equation 2.17,

$$\arg \max_{j \in \{1, \dots, M\}} \left| \mathcal{C}_j^t \cap \mathcal{N}(i) \right| \left( 1 - \frac{|\mathcal{C}_j^t|}{H_j} \right), \quad (2.17)$$

where  $\mathcal{C}_j^t$  is the set of nodes assigned to cluster  $j$  at step  $t$  of the algorithm,  $H_j$  is the maximum capacity of cluster  $\mathcal{C}_j$ .

The algorithm’s objective must be slightly modified to accommodate edge-weighted graphs. Let  $w_{il}$  be the weight of the edge between node  $i$  and node  $l$ . We therefore extend the  $|\mathcal{C}_j^t \cap \mathcal{N}(i)|$  term to the edge-weighted setting by redefining it as  $\sum_{l \in \mathcal{C}_j^t \cap \mathcal{N}(i)} w_{il}$ . Furthermore, we must also modify the balance requirement, since only the bidder side of the bipartite graph clustering is required to be balanced! We therefore replace  $\left( 1 - |\mathcal{C}_j^t|/H_j \right)$  with  $\left( 1 - |\mathcal{C}_{j,c}^t|/H_{j,c} \right)$  where  $\mathcal{C}_{j,c}^t$  is the set of bidder nodes in cluster  $\mathcal{C}_j^t$  and  $H_{j,c}$  is the maximum number of allowed bidder nodes in cluster  $\mathcal{C}_j^t$ . The final objective is given by:

$$\arg \max_{j \in \{1, \dots, M\}} \left| \sum_{l \in \mathcal{C}_j^t \cap \mathcal{N}(i)} w_{il} \right| \left( 1 - \frac{|\mathcal{C}_{j,c}^t|}{H_{j,c}} \right) \quad (2.18)$$

Figure 2.3 plots the proportion of edges cut, weighted by the bid amount, over consecutive runs of the reLDG algorithm for 50 and 100 clusters. We adopt three main vectors of comparison between candidate clusterings to determine the efficacy of our proposed experiment-of-experiment design:

- *Quality*: comparing clusterings of the graph that differ in their estimated quality, for example by looking at the number of edges cut, for a fixed number of clusters: we compare a random graph clustering to a clustering obtained by running the reLDG algorithm to convergence.
- *Number of clusters*: comparing two clusterings of the graph obtained by running the same clustering algorithm for a different number of clusters: we consider a reLDG clustering with 10 clusters and a reLDG clustering with 400 clusters.
- *Metric*: comparing clusterings of the graph that are obtained by applying the same algorithm on different bipartite graphs: we compare a reLDG clustering of the *bid* graph with an reLDG clustering of the *impressions* graph.

The dataset does not provide the budgets of the bidders or their perceived ad quality, hence we will adopt the same simplifying assumptions as Section 2.2 of no quality effects between bidders and no budget constraints. Furthermore, we assume bids are unchanged as a result of the experiment (which would be valid for rational, non budget-limited bidders).

### 2.3.2 Experimental results

We first compare a clustering of the graph obtained by running the modified reLDG algorithm (cf. Section 2.2) against a completely random balanced clustering of the graph. We fix a subset of auctions with few bidders per auction, in order to showcase the framework and establish the monotonicity and transitivity properties by allowing a setting for which there is a clear difference between the two clusterings. The reduction in cut size — measured by the ratio of the weighted sum of edges inter-clusters over the sum of all edge weights — over the iterations of the algorithm is shown in Figure 2.3. While the weighted cut of

the graph for a random clustering is around 98%, the clustering obtained with the reLDG algorithm approaches 66% within a few iterations.

We validate the monotonicity assumption, as well as the transitivity assumption, for reserve price experiments. In Figure 2.4 (a), we plot four distributions as well as the Total Treatment Effect estimand (cf. Eq. 1.2), obtained by taking the difference between assigning all units to a higher reserve price and assigning none. Namely, we plot the distribution of the Horvitz-Thomson estimator’s expectation (cf. Eq 1.6) under each cluster-based design:  $\mathbb{E}_{\mathbf{Z} \sim \mathcal{C}_k}[\hat{\tau}]$  where  $k = 1$  for the reLDG clustering and  $k = 2$  for the random clustering. We also plot the distribution of the expectation of the Experiment-of-Experiments (EoE) estimators:  $\mathbb{E}_{\mathbf{W}, \mathbf{Z} \sim \mathcal{C}_k^{\mathbf{W}}}[\hat{\tau}_k^{\mathbf{W}}]$ .

We find that they all under-estimate the true treatment effect, as expected from the increasing property. As expected, the Horvitz-Thompson estimator is more biased under a random clustering than under the reLDG clustering. Furthermore, we find that the property of transitivity holds (cf. Eq. 1), namely the Experiment-of-Experiments estimate of the “random estimator” under-estimates the total treatment effect more severely than the Experiment-of-Experiments estimate of the “reLDG estimator”.

We repeat the experiment to compare a reLDG clustering with 10 clusters with another reLDG clustering with 400 clusters (cf. Figure 2.4 (b)). We find that the clustering with 10 clusters is less biased but exhibits higher variance, and that the transitivity property holds. Finally, in Figure 2.4 (c), we compare a clustering of the impressions bipartite graph with a clustering of the bid bipartite graph. The transitivity property is again verified. Moreover, we see that clustering the bid bipartite graph may be a better heuristic in this setting, but the difference in the two clusterings is very slight. The code is available for download at <https://jean.pouget-abadie.com/kdd2018code.html>.

## 2.4 Conclusion

We have introduced two properties, monotonicity and transitivity, under which the estimation of causal effects in the presence of interference can be improved by selecting the

least-biased of two clusterings. We proved that certain parametric models of interference are monotone and transitive. A more exhaustive examination of other parametric models of interference (e.g. Biswas and Airoidi (2018); Basse and Airoidi (2015)) for these properties was beyond the scope of this work. Furthermore, while we were able to prove monotonicity for certain reserve price experiments, transitivity was established only in simulations. A natural question arising from this work is whether monotonicity and transitivity can be established through empirical means, using an observational method or through a randomized experiment. Finally, while our Experiment-of-Experiment design can improve the bias of subsequent randomized experiments—by selecting which of two clusterings should be used for the cluster-based randomized assignment—the reduction in bias comes at a cost of reduced power in the current experiment: half the units belonging to the more biased clustering are discarded in the final analysis. Hence, an important direction of future work is quantifying and bounding this loss of power, as well as exploring alternate means of choosing a clustering with a smaller power reduction, either through observational data or a less intrusive experimental design.

## Chapter 3

# Randomized and Optimized Saturation designs

The results of chapters 1 and 2 are primarily centered around two standard experimental designs: the completely-randomized design and the cluster-based randomized design. Both are unbiased when the standard unit treatment value assumption holds. When interference is present however, it is generally believed that a cluster-based randomized design will be less biased (Eckles *et al.*, 2017), but will have higher variance than a completely-randomized design that assigns the same proportion of units to treatment. The complexity of finding balanced partitionings of a large set of experimental units (Andreev and Racke, 2006) is another aspect to take into consideration when choosing between both of these standard designs.

A third design category is the *randomized saturation* design. It first assigns clusters of units to treatment proportions, and then assigns the units within each group to treatment and control at random according to the assigned treatment proportion. Randomized Saturation designs are often used in the context of interference because they allow the experimenter to infer a unit's reaction to varying levels of treatment (Banerjee *et al.*, 2012; Sinclair *et al.*, 2012; Crépon *et al.*, 2013). This is especially true if we are willing to make an anonymous interference assumption (Manski, 2013) or an assumption of no peer-effect-

heterogeneity (Athey *et al.*, 2015). For an excellent reference on randomized saturation designs, we refer the reader to (Baird *et al.*, 2016).

Randomized saturation designs offer an interesting middle ground between completely-randomized and cluster-based randomized designs. Both can be conceptualized as a randomized saturation design: the completely randomized designs corresponds to a randomized saturation design with identical treatment proportions across all clusters<sup>1</sup>; the cluster-based randomized corresponds to a randomized saturation design with full treatment or full control proportions. In this chapter, we aim to quantify the bias and variance tradeoffs we can obtain with a randomized saturation designs over the completely randomized design and the cluster-based randomized design.

*What and when do we stand to gain from running randomization saturation designs in the presence of interference?*

Randomized saturation designs are an example of *model-assisted* designs (Basse and Airoidi, 2015). Indeed, the treatment-proportions distribution can be chosen to optimize a particular objective under a set of model assumptions, without sacrificing the validity if the model is misspecified. With high confidence in our modelling assumptions, we can further optimize the assignment of treatment-proportions to each cluster of experimental units. We call these designs *Optimized saturation designs* and show that they yield additional improvements over their randomized saturation design counterparts under certain assumptions.

In Section 3.1, we define randomized saturation designs, and explore the bias and variance of the standard difference-in-means estimator under the stable unit treatment value assumption, as well as under a well-studied linear model of interference. We extend these results to random-graph model setting and other model-assisted estimators. In Section 3.2, we introduce and define optimized saturation designs and show that they can yield additional improvements over randomized saturation designs, but can also lead to

---

<sup>1</sup>In fact, this would correspond to a stratified completely randomized design

increased bias under model misspecification. Finally, in Section 3.3, we run a small-scale simulation study to validate the results of this chapter.

**Acknowledgements:** The results of this chapter were obtained in collaboration with Edoardo M. Airoidi, in preparation for a submission at a peer-reviewed journal.

### 3.1 Randomized saturation designs

A randomized saturation design is any two-stage design that first assigns clusters of the experimental units at random to treatment proportions, and then assigns the units within each cluster to treatment and control, respecting the assigned treatment proportion for each cluster. We re-use most of the notation from Chapters 1 and 2. Let  $N$  be the number of experimental units, let  $\mathbf{Y}$  be their outcome vector, and let  $\mathbf{Z} \in \{0, 1\}^N$  be the assignment vector stating whether each unit  $i$  is in treatment ( $Z_i = 1$ ) or control ( $Z_i = 0$ ). Let  $M$  be the number of clusters of the experimental units. There are many possible kinds of randomization saturation designs. We list two below, and show that they are equivalent when the number of clusters is large.

**Definition 2.** *The independently-sampled randomized saturation design is a two-stage design defined by a probability distribution  $\mathcal{D}$  on  $[0, 1]$  and the following procedure: for each cluster  $\mathcal{C}_j$ , sample  $\pi_j \sim \mathcal{D}$  and assign  $n_j = \lfloor \pi_j \cdot N_j \rfloor$  randomly-chosen units of cluster  $\mathcal{C}_j$  to treatment and the remainder  $N_j - n_j$  units of cluster  $\mathcal{C}_j$  to control.*

The independently-sampled randomized saturation design is entirely characterized by its distribution  $\mathcal{D}$ . The total number of treated units is a random variable, given by  $n_t = \sum_{j=1}^M \lfloor \pi_j \cdot N_j \rfloor$ . Assuming the size of each cluster is large ( $N_j \gg 1$ ), the expected number of treated units over the sampling of  $\boldsymbol{\pi} \sim D^M$  is the expectation of  $\mathcal{D}$  times the total number of experimental units  $N$ .

$$\mathbb{E}_{\boldsymbol{\pi} \sim D^M} [n_t] \approx N \cdot \mathbb{E}_{\pi \sim \mathcal{D}} [\pi] \tag{3.1}$$

**Definition 3.** The permutation-based randomized saturation design is a two-stage randomized design defined by a fixed vector  $\boldsymbol{\pi} \in [0, 1]^M$  of length  $M$  and the following procedure: sample a random permutation  $P$  of  $[1, M]$ , letting  $\mathbf{P}$  be the corresponding permutation matrix of  $P$ . For each block  $\mathcal{C}_j$ , assign  $n_j = \lfloor (\mathbf{P}\boldsymbol{\pi})_j N_j \rfloor$  randomly-chosen units of  $\mathcal{C}_j$  to treatment, and the remainder  $N_j - n_j$  units of  $\mathcal{C}_j$  to control, where  $(\mathbf{P}\boldsymbol{\pi})_j$  is the  $j^{\text{th}}$  coordinate of the permuted vector  $\mathbf{P}\boldsymbol{\pi}$ .

The permutation-based design is entirely characterized by its vector  $\boldsymbol{\pi}$ . The total number of treated units is fixed when the clusters are of equal size:

$$n_t = \sum_{j=1}^M \left\lfloor \pi_j \frac{N}{M} \right\rfloor \quad (3.2)$$

For this reason, we tend to prefer the second implementation of randomized saturation designs, and the one we will refer to unless otherwise explicitly stated. For ease of exposition, we will assume that the number of units in each cluster is large enough to ignore the flooring function.

The treatment-proportions vector  $\boldsymbol{\pi}$  can be chosen explicitly by the experimenter or be the result of an optimization program; it can also be randomly sampled from a probability distribution. In the latter case, the independently-sampled and permutation-based randomized saturation designs are equivalent when the number of clusters is large. Assuming that the treatment proportions vector is sampled from a probability distribution  $\boldsymbol{\pi} \sim \mathcal{D}^M$ , the  $k^{\text{th}}$  moment of the number of units assigned to treatment in each the permutation-based design is equal asymptotically to its  $k^{\text{th}}$  moment under the independently-sampled design by the law of large numbers.

$$\forall k \in \mathbb{N}, \mathbb{E}_P \left[ (\mathbf{P}\boldsymbol{\pi})_j^k \right] = \sum_{j=1}^M \frac{\pi_j^k}{M} \xrightarrow{M \rightarrow +\infty} \mathbb{E}_{\pi_j \sim \mathcal{D}} [\pi_j^k] \quad (3.3)$$

In the equation above,  $\mathbb{E}_P \left[ (\mathbf{P}\boldsymbol{\pi})_j^k \right]$  is the  $k^{\text{th}}$  moment of the  $j^{\text{th}}$  coordinate of  $(\mathbf{P}\boldsymbol{\pi})$ , and is shown to be equivalent to the  $k^{\text{th}}$  moment of the  $j^{\text{th}}$  coordinate of  $\boldsymbol{\pi}$  sampled according to  $\mathcal{D}$ ,  $\mathbb{E}_{\pi_j \sim \mathcal{D}} [\pi_j^k]$ , when the number of clusters is large.

The two standard randomized designs that we have studied thus far are in fact instantiations of randomized saturation designs. The cluster-based randomized design is an example

of a randomized saturation design where  $\boldsymbol{\pi} \in \{0, 1\}^M$ , assigning either all of cluster to treatment or to control. The randomized saturation design with constant vector  $\boldsymbol{\pi} = \left(\frac{n_t}{N}\right)^M$ , corresponds to a stratified completely randomized assignment, where the same proportion of units are assigned to treatment in every cluster.

### 3.1.1 Bias and variance under no interference

Recall that  $\hat{\tau}$  denotes the difference-in-means estimator, defined in Equation 1.4. By the law of iterated expectations, the shorthand  $\mathbb{E}_{\mathbf{Z}}[\hat{\tau}]$  denotes  $\mathbb{E}_{\boldsymbol{\pi}}[\mathbb{E}_{\mathbf{Z}}[\hat{\tau}|\boldsymbol{\pi}]]$ , i.e. the expectation taken with respect to the permutation of the treatment proportions assignment to clusters  $\boldsymbol{\pi}$  and with respect to the assignment of units to treatment and control  $\mathbf{Z}$ , conditioned on the assignment of  $\boldsymbol{\pi}$ . When the standard unit treatment value assumption holds, the difference-in-means estimator  $\hat{\tau}$  is unbiased under a randomized saturation design for the total treatment effect  $TTE$ , defined by

$$TTE := \frac{1}{N} \sum_{i=1}^N Y_i(1) - Y_i(0) \quad (3.4)$$

**Proposition 10.** *Assume the standard unit treatment value assumption holds,*

$$\mathbb{E}_{\mathbf{Z}}[\hat{\tau}|\boldsymbol{\pi}] = \frac{1}{n_t} \sum_{j=1}^M \pi_j Y_j^+(1) - \frac{1}{n_c} \sum_{j=1}^M (1 - \pi_j) Y_j^+(0) \quad (3.5)$$

where  $Y_j^+ := \sum_{i \in \mathcal{C}_j} Y_i$  is the cluster-level outcome of cluster  $\mathcal{C}_j$ . In expectation over the randomized saturation assignment,

$$\mathbb{E}_{\mathbf{Z}}[\hat{\tau}] = TTE \quad (3.6)$$

A proof is included in Section C.2. From Proposition 10, the difference-in-means estimator is *not* guaranteed to be unbiased if we condition on a specific assignment of clusters to treatment proportions, and only by randomizing over the assignment of treatment proportion can we guarantee unbiasedness.

$$\exists \boldsymbol{\pi}, \mathbb{E}_{\mathbf{Z}}[\hat{\tau}|\boldsymbol{\pi}] \neq TTE \quad (3.7)$$

We can also give a concise formula of the variance of the difference-in-means estimator

under the standard unit treatment value assumption and a randomized saturation design. Consider the outcome vector of cluster  $\mathcal{C}_j$  denoted by  $\mathbf{Y}^{(j)}$ . Let  $S_{tj} = \sigma^2(\mathbf{Y}^{(j)}(\mathbf{1}))$ ,  $S_{cj} = \sigma^2(\mathbf{Y}^{(j)}(\mathbf{0}))$ ,  $S_{tcj} = \sigma^2(\mathbf{Y}^{(j)}(\mathbf{1}) - \mathbf{Y}^{(j)}(\mathbf{0}))$ . Recall the definition of the cluster-aggregated outcomes vector  $\mathbf{Y}^+$ , with coordinates  $Y_j^+ = \sum_{i \in \mathcal{C}_j} Y_i$  for every cluster  $\mathcal{C}_j$ . Let  $S_t^+ = \sigma^2(\mathbf{Y}^+(\mathbf{1}))$ ,  $S_c^+ = \sigma^2(\mathbf{Y}^+(\mathbf{0}))$ ,  $S_{tc}^+ = \sigma^2(\mathbf{Y}^+(\mathbf{1}) - \mathbf{Y}^+(\mathbf{0}))$ .

**Proposition 11.** *When the standard treatment value assumption holds, the conditional variance of the difference-in-means estimator with respect to the assignment of treatment proportions to clusters  $\boldsymbol{\pi}$  is*

$$\text{Var}_{\mathbf{Z}}[\hat{\tau} | \boldsymbol{\pi}] = \frac{N^2}{n_t n_c} \sum_{j=1}^M \frac{N_j}{N} \pi_j (1 - \pi_j) \left( \frac{S_{tj}}{n_t} + \frac{S_{cj}}{n_c} - \frac{S_{tcj}}{N} \right) \quad (3.8)$$

The variance of the difference-in-means estimator under a randomized saturation design is

$$\begin{aligned} \text{Var}_{\mathbf{Z}}[\hat{\tau}] &= \frac{n_t n_c}{N^2} \sum_{j=1}^M N_j \left( \frac{S_{tj}}{n_t^2} + \frac{S_{cj}}{n_c^2} + \frac{S_{tcj}}{n_t n_c} \right) \\ &+ \text{Var}[\boldsymbol{\pi}] \left[ M \left( \frac{S_t^+}{n_t^2} + \frac{S_c^+}{n_c^2} + \frac{S_{tc}^+}{n_c n_t} \right) - \sum_{j=1}^M N_j \left( \frac{S_{tj}}{n_t^2} + \frac{S_{cj}}{n_c^2} + \frac{S_{tcj}}{n_t n_c} \right) \right] \end{aligned} \quad (3.9)$$

A proof can be found in Section C.3. The important takeaway from Proposition 11 is that the variance of difference-in-means estimator for a randomized saturation design under the stable treatment value assumption is linear in the empirical variance of the treatment-proportions vector  $\boldsymbol{\pi}$ . Optimizing the variance of this estimator, or any function thereof, under this no-interference assumption reduces to choosing the optimal variance of the treatment proportions vector  $\boldsymbol{\pi}$ . The following lemma characterizes the two extrema of  $\text{Var}[\boldsymbol{\pi}]$ , with the treatment-proportions vector  $\boldsymbol{\pi}$  constrained to verify  $\forall i, \pi_i \in [0, 1]$  and  $\bar{\pi} = \frac{n_t}{N}$ .

**Lemma 2.** *The minimum of the variance of the treatment-proportions vector  $\boldsymbol{\pi}$  is 0, attained if and only if the vector  $\boldsymbol{\pi}$  is constant and equal to  $(\frac{n_t}{N})_M$ . Constrained to verify  $\bar{\pi} = \frac{n_t}{N}$ , the variance of the treatment-proportions vector is maximized only for  $\boldsymbol{\pi} \in \{0, 1\}^M$ , assigning either all of a cluster to treatment or control, assuming such a solution exists. The variance of  $\boldsymbol{\pi}$  is then  $\text{Var}_{\mathbf{Z}}[\boldsymbol{\pi}] = \frac{n_t n_c}{N^2}$ .*

A proof can be found in Section C.1. As we seek to optimize the variance in Eq. 3.9,

it is important to note that the coefficient in front of  $\text{Var}_{\mathbf{Z}}[\boldsymbol{\pi}]$  is not always positive. For example, suppose that the aggregated outcomes of each cluster are identical, but with some inter-cluster variance.

$$\exists \mathbf{Y}, \forall \mathcal{C}_k, \mathcal{C}_j, Y_k^+ = Y_j^+ \text{ and } S_{tj} > 0 \text{ and } S_{cj} > 0 \text{ and } S_{tcj} > 0$$

In this case, Eq. 3.9 becomes:

$$\text{Var}_{\mathbf{Z}}[\hat{\tau}] = \left( \frac{n_t n_c}{N^2} - \text{Var}[\boldsymbol{\pi}] \right) \sum_{j=1}^M N_j \left( \frac{S_{tj}}{n_t^2} + \frac{S_{cj}}{n_c^2} + \frac{S_{tcj}}{n_t n_c} \right) \quad (3.10)$$

From Lemma 2, it holds that  $\text{Var}[\boldsymbol{\pi}] \leq \frac{n_t n_c}{N^2}$ , with equality when  $\boldsymbol{\pi} \in \{0, 1\}^M$ , corresponding to a cluster-based randomized assignment. We then have  $\text{Var}_{\mathbf{Z}}[\hat{\tau}] = 0$ , which is expected since the cluster-level outcomes are identical. Thus, the assignment which minimises the variance or the mean squared error of the difference-in-means estimator is not always the stratified completely randomized assignment, assigning the same treatment-proportions to each cluster.

Finally, for a constant vector  $\boldsymbol{\pi} = \{\frac{n_t}{M}\}_M$  and for a vector  $\boldsymbol{\pi} \in \{0, 1\}^M$ , we recover the standard variance formulas (Imbens and Rubin, 2015) of a stratified completely-randomized assignment and a cluster-based randomized assignment.

**Corollary 1.** *For a constant vector  $\boldsymbol{\pi} = \{\frac{n_t}{N}\}_M$ , Eq. 3.9 becomes:*

$$\text{Var}_{\mathbf{Z}}[\hat{\tau}] = \sum_{j=1}^M \frac{N_j}{N} \left( \frac{S_{tj}}{n_t} + \frac{S_{cj}}{n_c} - \frac{S_{tcj}}{N} \right) \quad (3.11)$$

*For a cluster-based randomized assignment,  $\boldsymbol{\pi} \in \{0, 1\}^M$ , Eq. 3.9 becomes*

$$\text{Var}_{\mathbf{Z}}[\hat{\tau}] = \frac{M}{N} \left( \frac{S_t^+}{n_t} + \frac{S_c^+}{n_c} - \frac{S_{tc}^+}{N} \right) \quad (3.12)$$

A proof can be found in Section C.4.

### 3.1.2 Bias under a linear interference model

In the previous section, we explored the bias and variance of the difference-in-means estimator under the stable unit treatment value assumption. In this section, we seek to

extend these results to a setting where interference is present. For the sake of exposition, we will focus on the linear model of interference, introduced in a simplified form in Equation 1.9 of Chapter 1, and then in a more generalized form in Equation 2.7 of Chapter 2. We generalize it again slightly here.

Consider a network over the units of experimentation, such that an edge between two units indicates they are likely to interfere with one another. Let the neighborhood  $\mathcal{N}_i$  of unit  $i$  be the set of all units linked by a direct edge to unit  $i$  and let  $(\alpha_i, \beta_i, \gamma_i) \in \mathbb{R}^3$ . The outcome of unit  $i$  is

$$Y_i(\mathbf{Z}) = \alpha_i + \beta_i Z_i + \gamma_i \rho_i + \epsilon_i \quad (3.13)$$

where  $\rho_i = \frac{1}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} Z_j$  is the proportion of  $i$ 's neighborhood that is treated and  $\epsilon_i \sim \mathcal{N}(0, 1)$  is a unit-specific random effect. The  $\beta_i$  coefficient can be interpreted as a direct effect parameter, while the  $\gamma_i$  coefficient can be interpreted as an interference parameter: if  $\forall i, \gamma_i = 0$ , then the standard unit treatment value assumption holds.

Unlike the linear model in Equation 2.7, we consider individualistic direct treatment and interference parameters, though we will occasionally make the assumption that the interference effect parameters  $\gamma$  are the same within each block of the graph. In that case, we say that the interference effects are *block-fixed*:

$$\forall k, \forall (i, j) \in \mathcal{C}_k, \gamma_i = \gamma_j = \gamma^{(k)}$$

The linear model of interference in Equation 3.13 is an example of an anonymous interaction model (Manski, 2013) for which randomized saturation designs are appropriate. For any two assignment vectors  $\mathbf{Z}$  and  $\mathbf{Z}'$  such that the number of treated neighbors of unit  $i$  is identical, unit  $i$ 's outcome is held constant.

$$\sum_{j \in \mathcal{N}_i} Z_j = \sum_{j \in \mathcal{N}_i} Z'_j \implies Y_i(\mathbf{Z}) = Y_i(\mathbf{Z}')$$

We begin by quantifying the total treatment effect for this linear model of interference. Let  $\bar{\beta} := \frac{1}{N} \sum_i \beta_i$  and  $\bar{\gamma} := \frac{1}{N} \sum_i \gamma_i$  be the average of the direct treatment effect and interference parameters respectively.

**Proposition 12.** *Under the linear model of interference in Equation 3.13, the total treatment effect is*

$$TTE = \bar{\beta} + \bar{\gamma} \quad (3.14)$$

We explore the bias of the classic difference-in-means estimator, assuming block-fixed interference effects. We introduce the following weighted average of the individualistic direct treatment effect vector  $\beta$ , which weighs each component  $\beta_i$  by the treatment proportion  $\pi_j$  of the corresponding cluster  $\mathcal{C}(i)$  that unit  $i$  belongs to:

$$\bar{\beta} := \frac{1}{n_t} \sum_{j=1}^M \pi_j \sum_{i \in \mathcal{C}_j} \beta_i \quad (3.15)$$

We introduce the following linear combination of the different components of  $\gamma$ , where each component is downweighted by the proportion of inter-cluster edges to intra-cluster edges per cluster:

$$\gamma' := \frac{1}{M} \sum_j \frac{\gamma^{(j)}}{N_j} \sum_{i \in \mathcal{C}_j} \frac{|\mathcal{N}_i \cap \mathcal{C}_j|}{|\mathcal{N}_i|} \quad (3.16)$$

$\gamma'$  is a measure of clustering quality and is contained in the segment  $[0, \bar{\gamma}]$ . For a perfect clustering,  $\gamma' = \bar{\gamma}$ . For a random clustering,  $\gamma' \approx \frac{\bar{\gamma}}{M}$ . For a clustering which places no unit in the same cluster as one of its neighbors,  $\gamma' = 0$ .

**Theorem 8.** *Assuming block-fixed interference effects, the expectation of the difference-in-means estimator conditioned on the assignment of clusters to treatment-proportions is*

$$\mathbb{E}_{\mathbf{Z}} [\hat{\tau} | \boldsymbol{\pi}] = \bar{\beta} + \sum_{j,l=1}^M \gamma^{(j)} \pi_j \left( \frac{\pi_l}{n_t} - \frac{1 - \pi_l}{n_c} \right) \sum_{i \in \mathcal{C}_j} \frac{|\mathcal{N}_i \cap \mathcal{C}_l|}{|\mathcal{N}_i|} \quad (3.17)$$

*Taking the expectation with respect to the random assignment of treatment-proportions to clusters:*

$$\mathbb{E}_{\mathbf{Z}} [\hat{\tau}] = \bar{\beta} + \frac{N^2}{n_t n_c} \left( \gamma' - \frac{\bar{\gamma} - \gamma'}{M - 1} \right) \text{Var}[\boldsymbol{\pi}] \quad (3.18)$$

The important takeaway of Theorem 8 is that the expectation, and by extension the bias, of the difference-in-means estimator under a randomized saturation design is linear in the empirical variance of the treatment-proportions vector  $\boldsymbol{\pi}$ . Similarly to optimizing  $\text{Var}[\hat{\tau}]$  under the stable treatment value assumption (cf. Prop. 11), optimizing the bias

of a randomized saturation design under the linear model of interference above reduces to choosing the optimal variance of the treatment-proportions vector. In the following corollaries, we examine two extreme cases of Theorem 8: one of a perfect clustering and one of a random clustering.

**Corollary 2.** *Suppose that the clusters of units are the same size and have no edges between them, and that the interference effects are block-fixed, then the bias is*

$$|TTE - \mathbb{E}_{\mathbf{Z}}[\hat{\tau}]| = \left(1 - \frac{N^2}{n_t n_c} \text{Var}[\boldsymbol{\pi}]\right) \bar{\gamma} \quad (3.19)$$

*The bias of the difference-in-means estimator is 0 if and only if the assignment  $\boldsymbol{\pi}$  assigns either all of a cluster to treatment or none:  $\boldsymbol{\pi} \in \{0, 1\}^M$ . If we also assume constant direct treatment and interference effect parameters, then the conditional bias is also linear in  $\text{Var}[\boldsymbol{\pi}]$ :*

$$|TTE - \mathbb{E}_{\mathbf{Z}}[\hat{\tau}|\boldsymbol{\pi}]| = \left(1 - \frac{N^2}{n_t n_c} \text{Var}[\boldsymbol{\pi}]\right) \gamma \quad (3.20)$$

*The conditional bias is then equal to 0 if and only if the assignment  $\boldsymbol{\pi}$  assigns either all of a cluster to treatment or none.*

A proof can be found in Section C.9. As expected, for a perfectly-clustered graph, the bias is minimized for a cluster-based randomized design. We now explore the other extreme case where the units are clustered at random.

**Corollary 3.** *Suppose that the units are randomly placed in clusters of equal size, and that the interference effects are block-fixed, then the bias is constant asymptotically:*

$$|TTE - \mathbb{E}_{\mathbf{Z}}[\hat{\tau}]| \approx \bar{\gamma}$$

*If we also make the assumption of constant direct and interference effects, then the conditional bias is also constant:*

$$|TTE - \mathbb{E}_{\mathbf{Z}}[\hat{\tau}|\boldsymbol{\pi}]| \approx \gamma$$

A proof can be found in Section C.10. In the completely random case, with constant direct and interference effects, the treatment-proportions vector  $\boldsymbol{\pi}$  has no influence on the

bias. As in the case of the perfectly-clustered assignment, if we relax the assumption of constant direct and interference effects to block-fixed interference effects, this is no longer true pointwise for every vector  $\boldsymbol{\pi}$ , but holds in expectation. The previous two corollaries can be generalized by the following theorem.

**Theorem 9.** *Assume that the interference effects are block-fixed and that the linear interference model from Equation 3.13 holds. We can distinguish two cases. If  $\gamma' \geq \frac{\bar{\gamma}}{M}$ , then the bias of the difference-in-means estimator under a randomized saturation design is minimized for a cluster-based randomized assignment  $\boldsymbol{\pi} \in \{0, 1\}^M$ , where  $\gamma'$  is defined in Eq. 3.16. The bias is then equal to*

$$|TTE - \mathbb{E}_{\mathbf{Z}}[\hat{\tau}]| = \frac{M}{M-1} (\bar{\gamma} - \gamma')$$

*If  $\gamma' \leq \frac{\bar{\gamma}}{M}$ , then the bias of the difference-in-means estimator under a randomized saturation design is minimized for a constant treatment-proportions vector  $\boldsymbol{\pi} = \left(\frac{n_i}{N}\right)_M$  and is equal to*

$$|TTE - \mathbb{E}_{\mathbf{Z}}[\hat{\tau}]| = \bar{\gamma}$$

*If  $\gamma' = \frac{\bar{\gamma}}{M}$ , then both results hold: the bias is constant, such that every randomized saturation design minimizes the bias.*

The significance of  $\frac{\bar{\gamma}}{M}$  as the cut-off point is intuitive: when the graph is randomly-clustered,  $\gamma' \approx \frac{\bar{\gamma}}{M}$ . Hence, the first regime corresponds to a “better-than-random” clustering of the experimental units, while the second regime corresponds to a “worse-than-random” or “adversarially-clustered” clustering. A proof can be found in Section C.11.

In conclusion, to optimize the bias of the difference-in-means estimator for a randomized saturation design under the linear interference model in Equation 3.13, the optimal randomized saturation design is either a stratified completely randomized design or a cluster-based randomized design—the parameter  $\gamma'$ , an indicator of the quality of the clustering, being the deciding factor between the two.

### 3.1.3 Extension to a random graph model

As shown in Theorems 8 and 9, the bias of the difference-in-means estimator under a randomized saturation design for the linear model of interference depends on the quality of the clustering, as indicated by  $\gamma'$ . While we can compute  $\gamma'$  if the interaction graph of units is known exactly, it is sometimes useful to extend the definition of  $\gamma'$  as well as the results of Theorems 8 and 9 to a setting where only a random graph model of the interference mechanism is known,

One of the simplest and well-studied random graph models is the stochastic block model (Holland *et al.*, 1983; Anderson *et al.*, 1992; Wasserman and Faust, 1994; Goldenberg *et al.*, 2010). It states that the probability that an edge exists between two units in a graph  $G$  depends only on the clusters they belong to. In other words, two units belonging to clusters  $\mathcal{C}_k$  and  $\mathcal{C}_l$ —with  $l$  and  $k$  possibly equal—are linked by an edge with probability  $A_{kl}$ . We define  $\mathbf{A} := (A_{kl}) \in \mathbb{R}^{M^2}$  the *block-matrix* of the graph  $G$ , such that

$$\forall(i, j, k, l), i \in \mathcal{C}_k, j \in \mathcal{C}_l \implies \mathbb{P}((i, j) \in G) = A_{kl} \quad (3.21)$$

As shown in (Airoldi, 2016), in the asymptotic regime of many large clusters, the expectation of  $\gamma'$  with respect to the stochastic block model is

$$\mathbb{E}_A[\gamma'] \approx \frac{1}{M} \sum_{j=1}^M \frac{\gamma^{(j)}}{N_j} \frac{A_{jj}}{\sum_k A_{jk} N_k} \quad (3.22)$$

Theorem 8 and 9, as well as Corollary 2 and 3, can be reinterpreted with this probabilistic model of the interference graph.

**Theorem 10.** *Assuming block-fixed interference effects, the expectation of the difference-in-means estimator with respect to the random block model is*

$$\mathbb{E}_{\mathbf{Z}, \mathbf{A}}[\hat{\tau}] = \bar{\beta} + \frac{N^2}{n_t n_c} \left( \mathbb{E}_A[\gamma'] - \frac{\bar{\gamma} - \mathbb{E}_A[\gamma']}{M-1} \right) \text{Var}[\boldsymbol{\pi}] \quad (3.23)$$

Suppose that the block-model matrix  $\mathbf{A}$  is balanced and diagonal, then the bias is

$$|TTE - \mathbb{E}_{\mathbf{Z}, \mathbf{A}}[\hat{\tau}]| = \left( 1 - \frac{N^2}{n_t n_c} \text{Var}[\boldsymbol{\pi}] \right) \bar{\gamma} \quad (3.24)$$

The bias of the difference-in-means estimator is 0 if and only if the assignment  $\pi$  assigns either all of a cluster to treatment or none:  $\pi \in \{0, 1\}^M$ . Suppose that the block-model matrix is balanced and with constant entries equal to  $\frac{1}{M}$ , then the bias is constant:

$$|TTE - \mathbb{E}_{\mathbf{Z}, \mathbf{A}}[\hat{\tau}]| = \bar{\gamma}$$

When  $\mathbb{E}_A[\gamma'] \geq \frac{\bar{\gamma}}{M}$ , the optimal randomized saturation design is a cluster-based randomized design.

When  $\mathbb{E}_A[\gamma'] \leq \frac{\bar{\gamma}}{M}$ , the optimal design is a stratified completely-randomized design.

### 3.1.4 Extension to a stratified estimator

The difference-in-means estimator is agnostic to any model assumptions or validity of the clustering, and is unbiased if the standard unit treatment value assumption holds. Its simplicity and robustness make it an important estimator to study in the context of saturation designs. However, if we believe that the suggested clustering of units is representative in some way, then we can benefit from using a stratified estimator  $\hat{\tau}^s$ , defined in the following equation. Recall that  $N_j$  is the total number of units in cluster  $\mathcal{C}_j$  and  $n_j$  is the number of units assigned to treatment in that cluster. Let  $\lambda \in \mathbb{R}_+^M$  be a vector of positive coefficients, usually chosen to sum to 1.

$$\hat{\tau}^s := \sum_{j=1}^M \lambda_j \hat{\tau}_j = \sum_{j=1}^M \lambda_j \left( \sum_{i \in \mathcal{C}_j} \frac{Z_i}{n_j} Y_i(\mathbf{Z}) - \sum_{i \in \mathcal{C}_j} \frac{1 - Z_i}{N_j - n_j} Y_i(\mathbf{Z}) \right) \quad (3.25)$$

In general, we recommend choosing  $\lambda_j = \frac{N_j}{N}$ , as evidenced by the following result on the bias of the stratified estimator under the stable treatment value assumption:

**Proposition 13.** *Assume that the standard unit treatment value assumption holds. The conditional expectation of the stratified estimator is*

$$\forall \pi, \mathbb{E}_{\mathbf{Z}}[\hat{\tau}^s | \pi] = \mathbb{E}_{\mathbf{Z}}[\hat{\tau}^s] = \sum_{j=1}^M \frac{\lambda_j}{N_j} \sum_{i \in \mathcal{C}_j} Y_i(1) - Y_i(0) \quad (3.26)$$

If  $\lambda_j = \frac{N_j}{N}$ , then the stratified estimator is unbiased for the total treatment effect, conditionally on the assignment of treatment proportions to clusters. The same holds true in expectation over a

randomized saturation assignment.

A proof can be found in Section C.5. Similarly, the variance has an easily interpretable closed-form under the standard unit treatment value assumption. Let  $S_{tj} := \sigma^2(Y_i(1) : i \in \mathcal{C}_j)$  be the variance of the outcomes of the units in  $\mathcal{C}_j$  if they were all placed in treatment. Let  $S_{cj} := \sigma^2(Y_i(0) : i \in \mathcal{C}_j)$  be the variance of the outcomes of the units in  $\mathcal{C}_j$  if they were all placed in control. Let  $S_{tcj} := \sigma^2(Y_i(1) - Y_i(0) : i \in \mathcal{C}_j)$  be the variance of the difference of potential outcomes of the units in  $\mathcal{C}_j$ .

**Proposition 14.** *Assume that the standard treatment value assumption holds. The conditional variance of the stratified estimator is*

$$\text{Var}_{\mathbf{Z}}[\hat{\tau}|\boldsymbol{\pi}] = \sum_{j=1}^M \lambda_j^2 \left( \frac{S_{tj}}{n_j} + \frac{S_{cj}}{N_j - n_j} - \frac{S_{tcj}}{N_j} \right) \quad (3.27)$$

The variance of the stratified estimator under a randomized saturation design is

$$\text{Var}_{\mathbf{Z}}[\hat{\tau}^s] = \sum_{j=1}^M \lambda_j^2 \frac{N}{N_j} \left( \frac{S_{tj}}{\pi^\dagger N} + \frac{S_{cj}}{(1 - \pi)^\dagger N} - \frac{S_{tcj}}{N} \right) \quad (3.28)$$

where  $\pi^\dagger := \left( \frac{1}{M} \sum_{j=1}^M \pi_j^{-1} \right)^{-1}$  is the harmonic mean of  $\pi$  and  $(1 - \pi)^\dagger$  is the harmonic mean of  $1 - \pi$ . Constrained to maintain  $\bar{\pi} = \frac{n_t}{N}$ , the stratified completely randomized design with  $\boldsymbol{\pi} = \left( \frac{n_t}{N} \right)_M$  minimizes the variance of the stratified estimator in Eq. 3.28, which is then equal to

$$\text{Var}_{\mathbf{Z}}[\hat{\tau}^s] = \sum_{j=1}^M \lambda_j^2 \frac{N}{N_j} \left( \frac{S_{tj}}{n_t} + \frac{S_{cj}}{n_c} - \frac{S_{tcj}}{N} \right) \quad (3.29)$$

A proof can be found in Section C.6. In contrast to the variance of the difference-in-means estimator  $\text{Var}[\hat{\tau}]$ , which is linear in the variance of the treatment-proportions vector  $\text{Var}[\boldsymbol{\pi}]$  (cf. Prop 11), the variance of the stratified estimator  $\text{Var}[\hat{\tau}^s]$  depends on the variance of the treatment-proportions vector only through the inverse of the harmonic mean of  $\boldsymbol{\pi}$ , as stated in Proposition 14. Since any mean-preserving spread decreases the harmonic mean (Mitchell, 2004), when holding  $n_t$  constant, any increase in the variance of the treatment-proportions vector increases the variance of the stratified estimator under the stable unit treatment value assumption.

As expected, we recover the variance formula of Equation 3.11 when we set  $\forall j, \lambda_j = \frac{N_j}{N}$ . We can also express the bias of the stratified estimator under the linear interference model of Equation 3.13. Recall that  $\rho_C = \frac{1}{N} \sum_{i=1}^N \frac{|\mathcal{N}_i \cap \mathcal{C}_j|}{|\mathcal{N}_i|}$  is the proportion of a unit's neighborhood that also belongs to its cluster, averaged over all units.

**Proposition 15.** *Under the linear model of interference in Eq. 3.13, the expectation of the stratified estimator conditioned on the assignment of clusters to treatment-proportions is:*

$$\mathbb{E}_{\mathbf{Z}}[\hat{\tau}^s | \boldsymbol{\pi}] = \mathbb{E}_{\mathbf{Z}}[\hat{\tau}^s] = \bar{\beta} + \frac{1}{N} \sum_{j=1}^M \sum_{i \in \mathcal{C}_j} \gamma_i \frac{|\mathcal{N}_i \cap \mathcal{C}_j|}{|\mathcal{N}_i|} \quad (3.30)$$

If the interference effects are constant, then the formula becomes:

$$\mathbb{E}_{\mathbf{Z}}[\hat{\tau}^s] = \bar{\beta} + \rho_C \gamma \quad (3.31)$$

A proof can be found in Section C.8. In conclusion, the bias of the stratified estimator does not depend on the treatment-proportions vector  $\boldsymbol{\pi}$  under the standard unit treatment value assumption or under the suggested linear model of interference. Its variance, when the standard unit treatment value assumption can be assumed, decreases with  $\text{Var}[\boldsymbol{\pi}]$ .

## 3.2 Optimized saturation designs

In the previous section, we analysed the bias and variance of the difference-in-means estimator and the stratified estimator for a randomized saturation design. From these results, we determined which randomized saturation design optimized these objectives, under a regime where the standard unit treatment value assumption holds and a regime where a linear model of interference holds. Since the bias and variance of these estimators can be expressed in terms of the variance of the treatment-proportions vector  $\boldsymbol{\pi}$ , finding the “optimal randomized saturation designs” reduces to optimizing over  $\text{Var}[\boldsymbol{\pi}]$ .

This optimization is limited by the random assignment of coordinates of  $\boldsymbol{\pi}$  to each cluster. *Optimized saturation designs*, which we introduce below, go one step further in their optimization by removing the permutation step and choosing the optimal treatment

proportion for each cluster.

**Definition 4.** Let  $f$  be an objective function, taking as input a treatment-proportions vector  $\boldsymbol{\pi}$ , a clustering  $\mathcal{C}$  of the experimental units, and a set of parameters  $\Theta$ . Let  $\mathcal{S}$  be an allowable set of treatment-proportions vectors. An optimized saturation design selects  $\boldsymbol{\pi}^* \in \mathcal{S}$  that minimizes  $f$

$$\boldsymbol{\pi}^* \in \arg \min_{\boldsymbol{\pi} \in \mathcal{S}} f(\boldsymbol{\pi}, \mathcal{C}, \Theta) \quad (3.32)$$

and assigns  $n_j = \lfloor \pi_j N_j \rfloor$  randomly-chosen units of cluster  $\mathcal{C}_j$  to treatment and the remaining  $N_j - n_j$  units of cluster  $\mathcal{C}_j$  to control.

There are many possible optimized saturation designs one could choose from. We list some examples below:

**Example 1.** Let  $f$  be the bias of the difference-in-means estimator  $\hat{\tau}$  under the standard unit treatment value assumption. Let  $\mathcal{S} := \{\boldsymbol{\pi} \in [0, 1]^M : \bar{\pi} = \frac{n_t}{N}\}$  be the set of treatment proportion vectors with fixed average  $\frac{n_t}{N}$ , where  $n_t \in (0, N)$  is some fixed number of treated units.

$$f : (\boldsymbol{\pi}, \mathcal{C}, \{\mathbf{Y}(0), \mathbf{Y}(1)\}) \mapsto |\text{TTE} - \mathbb{E}_{\mathbf{Z}}[\hat{\tau} | \boldsymbol{\pi}]| \quad (3.33)$$

The constant vector  $\boldsymbol{\pi}^* := \left(\frac{n_t}{N}\right)_M$  minimizes the objective function  $f$  and belongs to  $\mathcal{S}$ :

$$\forall \mathcal{C}, \mathbf{Y}(0), \mathbf{Y}(1), \left(\frac{n_t}{N}\right)_M \in \arg \min_{\boldsymbol{\pi} \in \mathcal{S}} f(\boldsymbol{\pi}, \mathcal{C}, \{\mathbf{Y}(0), \mathbf{Y}(1)\})$$

In other words, the stratified completely randomized assignment is an optimized saturation design for  $f$ , the bias of the difference-in-means estimator under the standard unit treatment value assumption.

A proof can be found in Section C.12. Optimized saturation designs can seek to optimize more than simply the bias of an estimators, but the variance as well. A very reasonable objective is to optimize them jointly in the form of the mean-squared error, as is done in the following example.

**Example 2.** Let  $f$  be the mean-squared error of the difference-in-means estimator  $\hat{\tau}$  under the standard unit treatment value assumption. Let  $\mathcal{S} := \{\boldsymbol{\pi} \in [0, 1]^M : \bar{\pi} = \frac{n_t}{N}\}$  be the set of treatment

proportion vectors with fixed average  $\frac{n_t}{N}$ , where  $n_t$  is some fixed number of treated units.

$$f : (\boldsymbol{\pi}, \mathcal{C}, \{\mathbf{Y}(1), \mathbf{Y}(0)\}) \mapsto \text{MSE}_{\mathbf{Z}}[\hat{\tau}|\boldsymbol{\pi}] = (\text{TTE} - \mathbb{E}_{\mathbf{Z}}[\hat{\tau}|\boldsymbol{\pi}])^2 + \text{Var}_{\mathbf{Z}}[\hat{\tau}|\boldsymbol{\pi}] \quad (3.34)$$

If we make the assumption that all cluster-level outcomes  $Y_j^+ := \sum_{i \in \mathcal{C}_j} Y_i$  are equal,

$$\forall \mathcal{C}_k, \mathcal{C}_j, Y_k^+(1) = Y_j^+(1) \text{ and } Y_k^+(0) = Y_j^+(0) \quad (3.35)$$

Then, assuming without loss of generality that the coefficients  $\alpha_j := \frac{S_{tj}}{n_t} + \frac{S_{cj}}{n_c} - \frac{S_{tcj}}{N}$  are sorted in increasing order, the vector  $\boldsymbol{\pi}^*$ , defined as

$$\forall k, \pi_{k+1}^* = \max \left( 0, \min \left( 1, \frac{Mn_t}{N} - \sum_{j=1}^k \pi_j^* \right) \right)$$

minimizes the mean-squared error of the difference-in-means estimator under the stable treatment value assumption and the assumption that all cluster-level outcomes are equal. If  $\boldsymbol{\alpha}$  is constant, then the constant vector  $\boldsymbol{\pi}^*$ , equal to  $\left(\frac{n_t}{N}\right)_M$ , also minimizes the above objective.

A proof can be found in Section C.13. Beyond operating under the standard unit treatment value assumption, optimized saturation design can postulate a parametric model of potential outcomes with interference, like the one in Equation 3.13. They can also leverage other estimators than the difference-in-means estimator. In the following example, we postulate a block-model graph over the experimental units, such that the probability that two units—belonging to clusters  $\mathcal{C}_j$  and  $\mathcal{C}_k$  respectively—are linked by an edge, is  $A_{jk}$ . We seek to optimize the variance of different regression estimators for the direct treatment effect  $\beta$  and the interference effect  $\gamma$  from the following linear regression:

$$\forall i, y_i = \alpha + \beta Z_i + \gamma \rho_i + \epsilon_i \quad (3.36)$$

where  $\epsilon_i \sim \mathcal{N}(0, \sigma_j^2)$ , with  $i \in \mathcal{C}_j$  is a cluster-level random effect, and  $\rho_i := \frac{1}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} Z_j$  is the proportion of unit  $i$ 's neighbors that are treated.

**Example 3.** Assume that the response function is given by the linear model in Eq. 3.36. The

asymptotic Fisher information matrix associated with model parameters  $\alpha$ ,  $\beta$ , and  $\gamma$  is

$$I = - \begin{pmatrix} L & C & D \\ C & C & E \\ D & E & F \end{pmatrix} \quad (3.37)$$

where the entries are  $L := \frac{1}{N} \sum_{j=1}^M \frac{N_j}{\sigma_j^2}$ ,  $C := \frac{1}{N} \sum_{j=1}^M \frac{N_j}{\sigma_j^2} \pi_j$ ,  $D := \frac{1}{N} \sum_{j=1}^M \frac{N_j}{\sigma_j^2} \left( \sum_{k=1}^M \pi_k W_{jk} \right)$ , and  $F := \frac{1}{N} \sum_{j=1}^M \frac{N_j}{\sigma_j^2} \left( \sum_{k=1}^M \pi_k W_{jk} \right)^2$ , where  $W_{jk} := \frac{A_{jk} N_j}{\sum_l A_{jl} N_l}$  is an entry of the row-normalized block-model matrix. A complete proof can be found in (Airoidi, 2016)<sup>2</sup>. The variance of the interference parameter estimator  $\hat{\gamma}$  corresponds to the element in position (3,3) of  $I^{-1}$ . If we wish to optimize the estimation of the interference parameter, a reasonable objective function  $f$  to minimize is

$$f : (\boldsymbol{\pi}, \mathcal{C}, \{\mathbf{W}, \sigma^2\}) \mapsto \frac{LC - C^2}{\det(I)} \quad (3.38)$$

with  $S := \{\boldsymbol{\pi} \in [0, 1]^M : \bar{\pi} = \frac{n_t}{N}\}$ , the allowable set of treatment-proportions vectors and  $\sigma^2 \in \mathbb{R}^M$ , the cluster-level random effects. The variance of the direct treatment effect estimator  $\hat{\beta}$  corresponds to the element in position (2,2) of  $I^{-1}$ . If we want to optimize the estimation of the direct treatment effect, a reasonable objective function  $f$  to minimize is

$$f : (\boldsymbol{\pi}, \mathcal{C}, \{\mathbf{W}, \sigma^2\}) \mapsto \frac{LF - D^2}{\det(I)} \quad (3.39)$$

The determinant of  $I$ ,  $\det(I)$ , appearing in the denominator of both objectives, is equal to  $LCF + 2CDE - CD^2 - C^2F - LE^2$ .

### 3.3 Simulation study

In order to validate the results of the previous section, we implement a small-scale simulation study to validate the bias-variance tradeoffs available to randomized saturation designs. We also illustrate the possible benefits of optimization saturation designs.

---

<sup>2</sup>The referenced article is an unpublished manuscript. Contact [jeanpougetabadie@g.harvard.edu](mailto:jeanpougetabadie@g.harvard.edu) for more information.

### 3.3.1 The bias-variance tradeoff of randomized saturation designs

The first claim we wish to examine is the tradeoff in bias-variance offered by the randomized saturation design over the completely randomized design and the cluster-based randomized design. In Figure 3.2, we examine this trade-off under the standard unit treatment value assumption and in Figure 3.3, we examine this trade-off under a linear model of interference.

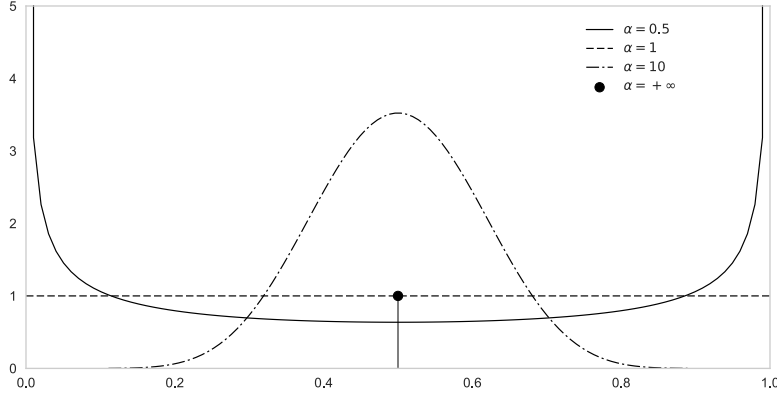
In the case of Figure 3.2, we consider 10,000 units, and group them into 20 clusters of equal size, each containing 500 units. We then vary the variance of the treatment-proportions vector by sampling it from a beta-distribution, with parameters  $(\alpha, \beta) = (\alpha, \alpha)$  for  $\alpha \in \mathbb{R}_+$ . The mean of this distribution is  $1/2$  and its variance is approximately equal to  $1/(16\alpha)$ , which decreases as  $\alpha$  increases. A plot of the probability density functions for this distribution under sample values of  $\alpha$  can be found in Figure 3.1.

Under the standard unit treatment value assumption, we let  $\beta, \sigma^2, \mu \in \mathbb{R}_+$  and consider the following response model:

$$\begin{aligned} \forall j \in [1, M], \mu_j &\sim \mathcal{U}([0, \mu]) \\ \forall i \in [1, N], Y_i(0) &\sim \mathcal{N}(0, 1) \\ \forall i \in [1, N], Y_i(1) &= Y_i(0) + \beta + \mu_{\mathcal{C}(i)} \end{aligned} \tag{3.40}$$

This response model is the sum of a direct treatment effect  $\beta$  and a uniformly-distributed cluster-level random effect  $\mu_{\mathcal{C}(i)}$ , where  $\mathcal{C}(i)$  corresponds to unit  $i$ 's cluster. The total treatment effect is equal to  $\beta$  asymptotically, which we set to 10 for this first simulation. In order to invert the sign in front of  $\text{Var}[\boldsymbol{\pi}]$  in Eq. 3.9 and explore different regimes of optimality, we first set  $\mu = 0$  in Figure 3.2.a, hence removing all cluster-level effects, and  $\mu = 0.1$  in Figure 3.2.b. Since the difference-in-means estimator  $\hat{\tau}$  is unbiased under the standard unit treatment value assumption for any randomized saturation design, we compare only the standard deviation of  $\hat{\tau}$  for different values of  $\alpha$ . We plot the error bars of the estimated standard deviation  $\hat{\tau}$ , obtained by bootstrapping the 10,000 simulations, each corresponding to a different assignment  $\mathbf{Z}$ .

The plots of Figure 3.2 have the form  $x \mapsto \pm 1/\sqrt{x}$  as expected. Indeed, as shown in



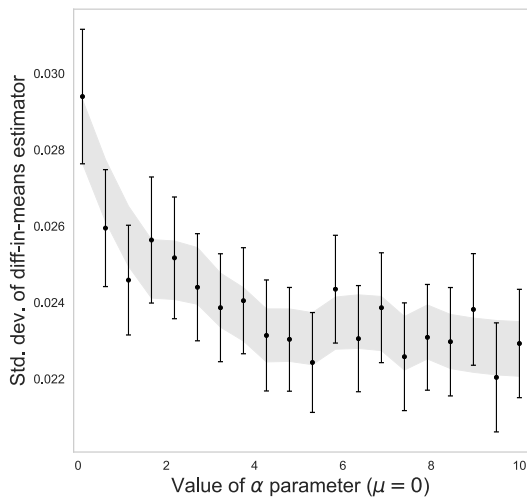
**Figure 3.1:** Various forms of the beta distribution

Eq. 3.9,  $\text{Var}_{\mathbf{Z}}[\hat{\tau}]$  is linear in  $\text{Var}[\pi]$  and the standard deviation of  $\pi$  is thus approximately proportional to  $\alpha^{-1}$ . The sign is decided by the relative magnitude of inner-cluster and cluster-level variances.

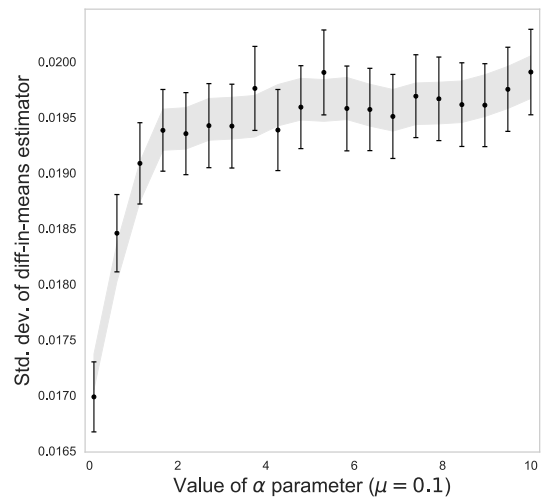
When  $\alpha$  goes to infinity, we approach a completely-randomized assignment; when  $\alpha$  goes to 0, we approach a cluster-based randomized assignment (cf. Figure 3.1). By comparing the two figures, we confirm the result of Proposition 11: the optimal assignment type depends on the relative magnitude of the inner-cluster and cluster-level variances. As an argument in favor of randomized saturation designs, setting  $\alpha = 5$ , which corresponds to neither a completely randomized assignment or a cluster-based randomized assignment, ensures that the randomized saturation design has close-to-optimal variance in both cases.

In Figure 3.3, we examine this same bias-variance tradeoff under a linear model of interference. We consider 2000 units, grouped into 40 clusters of equal size. We let  $\beta, \gamma \in \mathbb{R}_+$  and we simulate the following response model:

$$\begin{aligned}
 \forall j \in [1, M], \mu_j &\sim \mathcal{U}([0, \mu]) \\
 \forall i \in [1, N], \epsilon_i &\sim \mathcal{N}(0, 1) \\
 \forall i \in [1, N], Y_i &= \beta Z_i + \rho_i + \epsilon_i + \mu_{\mathcal{C}(i)}
 \end{aligned} \tag{3.41}$$



(a)



(b)

**Figure 3.2:** Standard deviation of the difference-in-means estimator under the stable unit treatment value assumption for a randomized saturation design. The treatment proportions are sampled according to a beta-distribution  $(\alpha, \alpha)$ , where we vary the parameter  $\alpha$  from 0.01 to 10. We plot the 90% error bars obtained by bootstrapping, at a rejection rate of 10%, 10 times the 10,000 simulations. The response model is given in Equation 3.40. We set  $\beta = 10$ . In (a), we set  $\mu = 0$ , corresponding to no cluster-level random effects. In (b), we set  $\mu = 0.1$ .

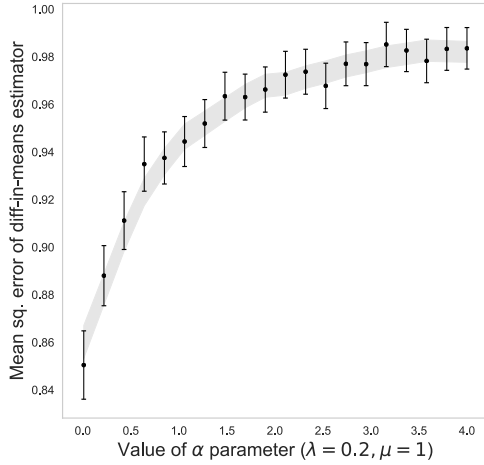
We set the direct treatment effect to  $\beta = 1$ . Like in the previous simulations,  $\mu_{\mathcal{C}(i)}$  is a uniformly-distributed cluster-level random effect, where  $\mathcal{C}(i)$  corresponds to unit  $i$ 's cluster. Recall that  $\rho_i := \frac{1}{|\mathcal{N}_i|} \sum_{j \in \mathcal{C}(i)} Z_j$  is the average proportion of a unit's neighbors that also belong to its cluster  $\mathcal{C}(i)$ , and is a measure of clustering quality. The higher  $\rho_i$  is for all units, the better clustered the graph is. The graph is sampled from a block model, using the following block matrix, where  $\lambda \in [0, 1]$  is a parameter of our choosing.

$$\begin{pmatrix} 1 & \lambda & \dots & \lambda \\ \lambda & 1 & \dots & \vdots \\ \vdots & \vdots & \ddots & \lambda \\ \lambda & \lambda & \dots & 1 \end{pmatrix}$$

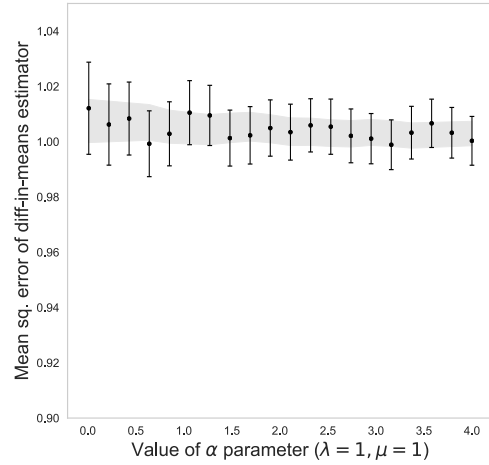
When  $\lambda = 0$ , the graph is perfectly clustered and  $\gamma' = \gamma \geq \frac{\gamma}{M}$ . When  $\lambda = 1$ , the graph corresponds to a random clustering and we are in the regime where  $\gamma' = \frac{\gamma}{M}$ . If we were to set the diagonal to 0 and  $\lambda \neq 0$ , the graph would correspond to an adversarial clustering, where  $\gamma' \leq \frac{\gamma}{M}$ .

In Figure 3.3, we explore four possible regimes of the mean-squared error of the difference-in-means estimator  $\hat{\tau}$  under the linear interference model, specified in Eq. 3.41, for different randomized saturation designs. The proportion of each block is selected according to a beta distribution of parameters  $(\alpha, \alpha)$  for  $\alpha \in \mathbb{R}_+$ . We vary  $\alpha$  from 0, corresponding to a cluster-based randomized assignment, to  $\alpha \geq 4$ , corresponding approximately to a stratified completely randomized assignment.

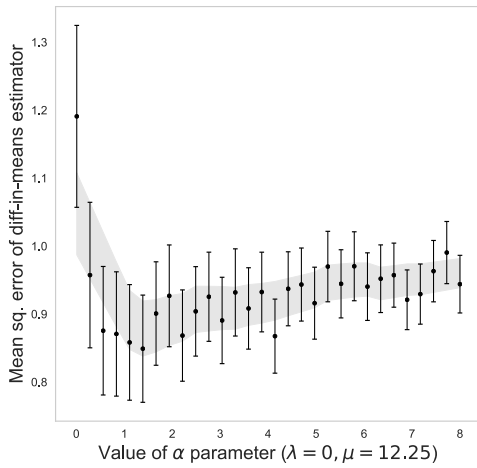
- In Figure 3.3.a, we consider an approximately well-clustered graph ( $\lambda = 0.2$ ) with low cluster-level random effects ( $\mu = 1$ ). As expected, the cluster-based randomized design outperforms all others in terms of the mean-squared error.
- In Figure 3.3.b, we consider a randomly-clustered graph ( $\lambda = 1$ ) and low cluster-level random effects ( $\mu = 1$ ). As expected, the mean-squared error does not vary with the distribution of the randomized saturation design, and both the completely randomized design and cluster-based randomized design perform equally well.



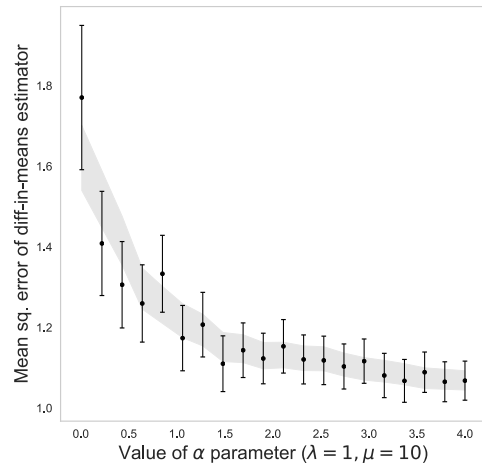
(a)



(b)



(c)



(d)

**Figure 3.3:** Mean-squared error of the difference-in-means estimator  $\hat{\tau}$  under the linear interference model in Eq. 3.41. The treatment proportions are sampled according to a beta-distribution, where we vary the parameter  $\alpha$ . When  $\alpha$  is close to 0, the randomized saturation design corresponds to a cluster-based randomized design, when  $\alpha$  approaches  $+\infty$ , the design corresponds to a stratified completely randomized design. We plot the 90% error bars, obtained by bootstrapping at a rejection rate of 10%, 10 times the 1000 simulations. We set  $\beta = 1$ . We explore four different regimes. Recall that the closer  $\lambda$  is to 0, the better the clustering and that the higher  $\mu$  is, the stronger the cluster-level random effects are. In (a), we set  $\lambda = 0.2$  and  $\mu = 1$ . In (b), we set  $\lambda = 1$  and  $\mu = 1$ . In (c), we set  $\lambda = 0$  and  $\mu = 12.25$ . In (d), we set  $\lambda = 1$  and  $\mu = 10$ .

- In Figure 3.3.c, the graph is well-clustered ( $\lambda = 0$ ) but the cluster-level random effects are significant ( $\mu = 12.25$ ). A cluster-based randomized design suffers from high variance despite being unbiased for the total treatment effect; a completely randomized design suffers from high bias despite having lower variance than its cluster-based randomized counterpart. The plot shows that the optimal design is obtained closer to  $\alpha = 1$ , corresponding to a uniform sampling of the treatment proportions within each cluster. We found that  $\mu = 12.25$  was the value under which the bias and variance had similar magnitudes, allowing us to obtain this U-shaped plot.
- Finally, in Figure 3.3.d, the graph is randomly clustered ( $\lambda = 1$ ) with high cluster-level effects ( $\mu = 10$ ). As expected, the completely randomized design performs better than the cluster-based randomized design: the higher  $\alpha$  is, the lower the mean-squared error.

### 3.3.2 The benefits of optimized saturation designs

We now wish to examine how optimized saturation designs improve over randomized saturation designs when their underlying modeling assumptions hold. In particular, we examine Example 3, where the objective is to choose the un-permuted vector  $\pi$  to minimize the variance of the interference effect parameter  $\gamma$  in the following linear regression:

$$\forall i, y_i = \alpha + \beta Z_i + \gamma \rho_i + \epsilon_i \quad (3.42)$$

where  $\epsilon_i \sim \mathcal{N}(0, \sigma_j^2)$  with  $i \in \mathcal{C}_j$  a cluster-level random effect, and  $\rho_i := \frac{1}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} Z_j$ , the proportion of unit  $i$ 's treated neighbors. We consider 5 clusters of 100 units and simulate the toy block-model graph for these units from the following  $5 \times 5$  block-matrices.

$$\begin{array}{ccc}
\begin{pmatrix} 0.5 & 0.1 & \dots & \dots & 0.1 \\ 0.1 & 0.5 & 0.1 & \ddots & \vdots \\ \vdots & 0.1 & 0.5 & 0.1 & \vdots \\ \vdots & \ddots & 0.1 & 0.5 & 0.1 \\ 0.1 & \dots & \dots & 0.1 & 0.5 \end{pmatrix} & 
\begin{pmatrix} 0.1 & 0 & 0 & 0 & 0 \\ 0 & 0.1 & 0 & 0 & 0 \\ 0 & 0 & 0.1 & 0 & 0 \\ 0 & 0 & 0 & 0.1 & 0 \\ 0 & 0 & 0 & 0 & 0.1 \end{pmatrix} & 
\begin{pmatrix} 0.8 & 0 & 1 & 0.5 & 1 \\ 0 & 0 & 0 & 0.5 & 0 \\ 1 & 0 & 0.2 & 1 & 1 \\ 0.5 & 0.5 & 1 & 0.5 & 0.5 \\ 1 & 0 & 1 & 0.5 & 0.1 \end{pmatrix} \\
(a) & (b) & (c)
\end{array}$$

**Table 3.1:** Block-model matrices used in simulation

These values were selected somewhat at random, but were chosen to be representative of three cases: a reasonable clustering, a perfect clustering, and an adversarially-chosen clustering, corresponding respectively to matrices (a), (b), and (c) in Table 3.1. We compared the completely-randomized design, the cluster-based randomized design, and the optimized saturation design for 4 objectives: the standard deviation of the estimator of the interference effect and of the estimator of the direct treatment effect, as well as the bias and standard deviation of the difference-in-means estimator. The results are reported in Table 3.2.

The optimized saturation design is chosen to minimize the variance of the interference effect and we report these values in bold in the table. As expected, and confirming the effectiveness of our optimization algorithm, the optimized saturation design reports the lowest variance of the interference effect estimator  $\hat{\gamma}$  in all three cases. Moreover, the design is relatively robust to the other objectives, with a comparable variance of the direct treatment effect estimator  $\hat{\beta}$  to its variance under the completely randomized design. Furthermore, in the perfect clustering (b), the variance of both estimators under the cluster-based assignment explodes, due to the strong covariance of the two covariates. The optimized saturation design does not suffer from this issue. Finally, while the cluster-based design is the least biased in cases where the clustering is better-than-random, i.e (a) and (b), the optimized saturation design, while it does not optimize for bias, is less biased and has lower variance than the completely randomized design in all three cases.

Block-model	Design	$\text{std}(\hat{\gamma}) \cdot 10^{-2}$	$\text{std}(\hat{\beta}) \cdot 10^{-3}$	$ TTE - \mathbb{E}_Z[\hat{\tau}] $	$\text{std}[\hat{\tau}] \cdot 10^{-1}$
(a)	Completely randomized	2.7 ( $7 \cdot 10^{-4}$ )	2.6 ( $\pm 7 \cdot 10^{-5}$ )	3.0 ( $\pm 4 \cdot 10^{-4}$ )	0.14 ( $\pm \cdot 10^{-4}$ )
	Cluster-based randomized	3.1 ( $\pm 7 \cdot 10^{-4}$ )	14 ( $\pm 3 \cdot 10^{-4}$ )	1.7 ( $\pm 4 \cdot 10^{-4}$ )	0.12 ( $\pm 3 \cdot 10^{-4}$ )
	Optimized saturation	<b>1.2</b> ( $\pm 3 \cdot 10^{-4}$ )	3.8 ( $\pm 9 \cdot 10^{-5}$ )	2.3 ( $\pm 4 \cdot 10^{-4}$ )	0.13 ( $\pm 3 \cdot 10^{-4}$ )
(b)	Completely randomized	0.77 ( $\pm 2 \cdot 10^{-4}$ )	2.6 ( $\pm 6 \cdot 10^{-5}$ )	3.0 ( $\pm 1 \cdot 10^{-3}$ )	0.44 ( $\pm 1 \cdot 10^{-3}$ )
	Cluster-based randomized	$10^4$	$10^4$	0.01 ( $\pm 1 \cdot 10^{-4}$ )	0.029 ( $\cdot 1 \cdot 10^{-4}$ )
	Optimized saturation	<b>0.50</b> ( $\pm 1 \cdot 10^{-4}$ )	3.6 ( $\pm 8 \cdot 10^{-5}$ )	1.5 ( $\pm 1 \cdot 10^{-3}$ )	0.35 ( $\pm 8 \cdot 10^{-4}$ )
(c)	Completely randomized	4.5 ( $\pm 1 \cdot 10^{-4}$ )	2.7 ( $\pm 7 \cdot 10^{-5}$ )	3.0 ( $\pm 3 \cdot 10^{-4}$ )	0.10 ( $\pm 3 \cdot 10^{-4}$ )
	Cluster-based randomized	2.1 ( $\pm 2 \cdot 10^{-4}$ )	7.0 ( $\pm 4 \cdot 10^{-4}$ )	3.4 ( $\pm 3 \cdot 10^{-2}$ )	7.8 ( $\pm 2 \cdot 10^{-2}$ )
	Optimized saturation	<b>0.93</b> ( $\pm 2 \cdot 10^{-5}$ )	3.0 ( $\pm 7 \cdot 10^{-5}$ )	2.6 ( $\pm 1 \cdot 10^{-4}$ )	0.038 ( $\pm 9 \cdot 10^{-5}$ )

**Table 3.2:** Standard deviation of the interference effect estimator  $\hat{\gamma}$  and the direct treatment effect estimator  $\hat{\beta}$  obtained from the linear regression in Eq. 3.42 as well as the bias and standard deviation of the difference-in-means estimator. These values are compared across multiple designs: the completely randomized design, the cluster-based randomized design, and the optimized saturation design specified in Example 3. The simulations consider 100 units, grouped into 5 clusters of equal size. The response models are computed according to Example 3 and the  $5 \times 5$  block-model matrices listed in Table 3.1. The standard deviations of each estimate indicated by parentheses, are obtained by bootstrapping 10 times the 1,000 simulations, with a rejection rate of 10%.

# Conclusion

Causal inference, as specified by the Neyman-Rubin potential outcomes framework, strongly relies on the standard unit treatment value assumption. When interference is present and this assumption is violated, the main results of causal inference no longer hold, short of assuming an explicit parametric model of interference. Extending these results under a minimal set of assumptions is key to deploying randomized designs and drawing causal inferences in practice.

When interference is present, designs must be tailored to the assumed interference structure in order to mitigate bias. Experiment-of-experiment designs are an intuitive diagnostic tool for determining the sensitivity of causal estimates to our choice of design. It is therefore unsurprising that experiment-of-experiments designs have allowed us to prove two results on causal inference with interference, which rely on very few assumptions. In Chapter 1, we showed that the presence of interference can be tested by comparing the outcomes of a completely randomized and cluster-based randomized assignment strategy. In Chapter 2, we proved that the relative effectiveness of two cluster-based randomized designs can be evaluated under an assumption of monotonicity and transitivity. Testing the validity of these two assumptions is a natural next research step that was not pursued here.

Randomized saturation designs are a simple type of experiment-of-experiment design, which compare completely randomized designs with different treatment proportions. In Chapter 3, we sought to better understand the properties of these designs, under the stable treatment value assumption and under interference. An area for future work is the extension of this analysis to other models of interference and estimators, commonly used by

practitioners.

Finally, while focusing on the setting of experimentation platforms at major tech companies has allowed us to alleviate many of causal inference's primary concerns—number of experimental units, compliance, collection of outcome data, blindness—there remain cases where technical constraints limit the possible units of randomization. (Roalson *et al.*, 2016; Zigler and Papadogeorgou, 2018) give examples of a bipartite structure of the interference graph. This causal set-up is relatively under-explored and yet altogether common at tech companies, raising a number of interesting new research directions.

# References

- (2018). Ad exchange auction model. Retrieved from [https://support.google.com/adxseller/answer/152039?hl=en&ref\\_topic=2904831](https://support.google.com/adxseller/answer/152039?hl=en&ref_topic=2904831) in February, 2018.
- AIROLDI, E. M. (2016). Optimal block randomized designs for causal inference on large networks. *Unpublished manuscript*.
- ANDERSON, C. J., WASSERMAN, S. and FAUST, K. (1992). Building stochastic blockmodels. *Social networks*, **14** (1-2), 137–161.
- ANDREEV, K. and RACKE, H. (2006). Balanced graph partitioning. *Theory of Computing Systems*, **39** (6), 929–939.
- ARAL, S. and WALKER, D. (2011). Creating social contagion through viral product design: A randomized trial of peer influence in networks. *Management science*, **57** (9), 1623–1639.
- ARONOW, P. M. (2012). A general method for detecting interference between units in randomized experiments. *Sociological Methods & Research*, **41** (1), 3–16.
- ATHEY, S., ECKLES, D. and IMBENS, G. W. (2015). *Exact P-values for Network Interference*. Tech. rep., National Bureau of Economic Research.
- AYDIN, K., BATENI, M. H. and MIRROKNI, V. S. (2016). Distributed balanced partitioning via linear embedding. In *WSDM*.
- BACKSTROM, L. and KLEINBERG, J. (2011). Network bucket testing. In *WWW*, pp. 615–624.
- BAIRD, S., BOHREN, J. A., MCINTOSH, C. and ÖZLER, B. (2016). Optimal design of experiments in the presence of interference. *Review of Economics and Statistics*.
- BAKSHY, E., ECKLES, D. and BERNSTEIN, M. S. (2014). Designing and deploying online field experiments. In *WWW*, pp. 283–292.
- BANERJEE, A. V., CHATTOPADHYAY, R., DUFLO, E., KENISTON, D. and SINGH, N. (2012). Can institutions be reformed from within? evidence from a randomized experiment with the rajasthan police. *CEPR Discussion Paper DP8869*.
- BASSE, G. and AIROLDI, E. (2017). Limitations of design-based causal inference and a/b testing under arbitrary and network interference. *arXiv:1705.05752*.
- , FELLER, A. and TOULIS, P. (2017). Exact tests for two-stage randomized designs in the presence of interference. *arXiv:1709.08036*.

- BASSE, G. W. and AIROLDI, E. M. (2015). Model-assisted design of experiments in the presence of network correlated outcomes. *arXiv:1507.00803*.
- BISWAS, N. and AIROLDI, E. M. (2018). Estimating peer-influence effects under homophily: Randomized treatments and insights. In *Complex Networks IX*, Springer International Publishing, pp. 323–347.
- BROOKS, N. (2004). The atlas rank report: How search engine rank impacts traffic. *Insights, Atlas Institute Digital Marketing*.
- CHAO, M.-T. and STRAWDERMAN, W. (1972). Negative moments of positive random variables. *Journal of the American Statistical Association*, **67** (338), 429–431.
- CHOI, D. S. (2014). Estimation of monotone treatment effects in network experiments. *arXiv:1408.4102*.
- COMMIT (1991). Community intervention trial for smoking cessation (commit): summary of design and intervention. *Journal of the National Cancer Institute*, **83** (22), 1620–1628.
- CORNFIELD, J. (1978). Symposium on chd prevention trials: Design issues in testing life style intervention randomization by group: A formal analysis. *American Journal of Epidemiology*, **108** (2), 100–102.
- CRÉPON, B., DUFLO, E., GURGAND, M., RATHELOT, R. and ZAMORA, P. (2013). Do labor market policies have displacement effects? evidence from a clustered randomized experiment. *The Quarterly Journal of Economics*, **128** (2), 531–580.
- DATTA, S., HALLORAN, M. E. and LONGINI, I. M. (1999). Efficiency of estimating vaccine efficacy for susceptibility and infectiousness: randomization by individual versus household. *Biometrics*, **55** (3), 792–798.
- DONNER, A. and KLAR, N. (2004). Pitfalls of and controversies in cluster randomization trials. *American Journal of Public Health*, **94** (3), 416–422.
- DURBIN, J. (1954). Errors in variables. *Revue de l'institut International de Statistique*, pp. 23–32.
- ECKLES, D., KARRER, B. and UGANDER, J. (2017). Design and analysis of experiments in networks: Reducing bias from interference. *Journal of Causal Inference*, **5** (1).
- , KIZILCEC, R. F. and BAKSHY, E. (2016). Estimating peer effects in networks with peer encouragement designs. *PNAS*, **113** (27), 7316–7322.
- FISHER, R. A. (1919). The causes of human variability. *International Journal of Epidemiology*, **10** (4), 213–220.
- (1925). *Statistical methods for research workers*. 1st ed, Oliver and Boyd.
- (1935). *Design of experiments*. Oliver and Boyd.
- GOLDENBERG, A., ZHENG, A. X., FIENBERG, S. E., AIROLDI, E. M. *et al.* (2010). A survey of statistical network models. *Foundations and Trends® in Machine Learning*, **2** (2), 129–233.

- GUI, H., XU, Y., BHASIN, A. and HAN, J. (2015). Network a/b testing: From sampling to estimation. In *WWW*.
- HAUSMAN, J. A. (1978). Specification tests in econometrics. *Econometrica: Journal of the Econometric Society*, pp. 1251–1271.
- HOLLAND, P. W., LASKEY, K. B. and LEINHARDT, S. (1983). Stochastic blockmodels: First steps. *Social networks*, **5** (2), 109–137.
- HONG, G. and RAUDENBUSH, S. W. (2012). Evaluating kindergarten retention policy. *Journal of the American Statistical Association*.
- HUDGENS, M. G. and HALLORAN, M. E. (2008). Toward causal inference with interference. *Journal of the American Statistical Association*, pp. 832–842.
- IMBENS, G. W. and RUBIN, D. B. (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press.
- KARYPIS, G. and KUMAR, V. (1998). Multilevelk-way partitioning scheme for irregular graphs. *Journal of Parallel and Distributed computing*, **48** (1), 96–129.
- KATZIR, L., LIBERTY, E. and SOMEKH, O. (2012). Framework and algorithms for network bucket testing. In *WWW*, pp. 1029–1036.
- KEMPTON, R. (1997). Interference between plots. In *Statistical methods for plant variety evaluation*, Springer, pp. 101–116.
- KOHAVI, R., DENG, A., FRASCA, B., WALKER, T., XU, Y. and POHLMANN, N. (2013). Online controlled experiments at large scale. In *KDD*, pp. 1168–1176.
- MALINEN, M. I. and FRÄNTI, P. (2014). Balanced k-means for clustering. In *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*, Springer, pp. 32–41.
- MANSKI, C. F. (2013). Identification of treatment response with social interactions. *The Econometrics Journal*, **16** (1), S1–S23.
- MIDDLETON, J. A. and ARONOW, P. M. (2011). Unbiased estimation of the average treatment effect in cluster-randomized experiments. *SSRN:1803849*.
- MITCHELL, D. W. (2004). 88.27 more on spreads and non-arithmetic means. *The Mathematical Gazette*, **88** (511), 142–144.
- MURRAY, D. M., VARNELL, S. P. and BLITSTEIN, J. L. (2004). Design and analysis of group-randomized trials: a review of recent methodological developments. *American Journal of Public Health*, **94** (3), 423–432.
- NISHIMURA, J. and UGANDER, J. (2013). Restreaming graph partitioning: simple versatile algorithms for advanced balancing. In *KDD*.

- POUGET-ABADIE, J., MIRROKNI, V., PARKES, D. C. and AIROLDI, E. M. (2018). Optimizing cluster-based randomized experiments under monotonicity. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, ACM, pp. 2090–2099.
- , SAVESKI, M., SAINT-JACQUES, G., DUAN, W., XU, Y., GHOSH, S. and AIROLDI, E. M. (2017). Testing for arbitrary interference on experimentation platforms. *arXiv:1704.01190*.
- RICHARDSON, M., DOMINOWSKA, E. and RAGNO, R. (2007). Predicting clicks: estimating the click-through rate for new ads. In *WWW*.
- ROLNICK, D., AYDIN, K., KAMALI, S., MIRROKNI, V. and NAJMI, A. (2016). Geocuts: Geographic clustering using travel statistics. *arXiv preprint arXiv:1611.03780*.
- , —, —, MIRROKNI, V. S. and NAJMI, A. (2017). Geocuts: Geographic clustering using travel statistics. *arxiv:1611.03780*.
- ROSENBAUM, P. R. (2007). Interference between units in randomized experiments. *Journal of the American Statistical Association*, **102** (477).
- RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, **66** (5), 688–701.
- (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of statistics*, pp. 34–58.
- (1980). Discussion of ‘randomization analysis of experimental data in the fisher randomization test’ by basu. *The Journal of the American Statistical Association*, **75** (371), 591–593.
- 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 *KDD*.
- SHAKYA, H., STAFFORD, D., HUGHES, D., KEEGAN, T., NEGRON, R., BROOME, J., MCKNIGHT, M., NICOLL, L., NELSON, J., IRIARTE, E., FERRERA, M., AIROLDI, E. M., FOWLER, J. and CHRISTAKIS, N. A. (2017). Exploiting social influence to magnify population-level behavior change in maternal and child health: Study protocol for a randomized controlled trial of network targeting algorithms in rural Honduras. *BMJ Open*, p. e012996.
- SINCLAIR, B., MCCONNELL, M. and GREEN, D. P. (2012). Detecting spillover effects: Design and analysis of multilevel experiments. *American Journal of Political Science*, **56** (4), 1055–1069.
- SOBEL, M. E. (2006). 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.
- SPLAWA-NEYMAN, J., DABROWSKA, D. M. and SPEED, T. (1923, 1990). On the application of probability theory to agricultural experiments. essay on principles. section 9. Translated in *Statistical Science*, pp. 465–472.

- STANTON, I. and KLIOT, G. (2012). Streaming graph partitioning for large distributed graphs. In *KDD*.
- STRUCHINER, C. J., HALLORAN, M. E., ROBINS, J. M. and SPIELMAN, A. (1990). The behaviour of common measures of association used to assess a vaccination programme under complex disease transmission patterns—a computer simulation study of malaria vaccines. *International Journal of Epidemiology*, **19** (1), 187–196.
- TANG, D., AGARWAL, A., O'BRIEN, D. and MEYER, M. (2010). Overlapping experiment infrastructure: More, better, faster experimentation. In *KDD*, pp. 17–26.
- TCHETGEN, E. J. T. and VANDERWEELE, T. J. (2012). On causal inference in the presence of interference. *Statistical Methods in Medical Research*, **21** (1), 55–75.
- TOULIS, P. and KAO, E. (2013). Estimation of Causal Peer Influence Effects. In *ICML*.
- TSOURAKAKIS, C., GKANTSIDIS, C., RADUNOVIC, B. and VOJNOVIC, M. (2014a). Fennel: Streaming graph partitioning for massive scale graphs. In *WSDM*, pp. 333–342.
- TSOURAKAKIS, C. E., GKANTSIDIS, C., RADUNOVIC, B. and VOJNOVIC, M. (2014b). FENNEL: streaming graph partitioning for massive scale graphs. In *WSDM*.
- UGANDER, J. and BACKSTROM, L. (2013). Balanced label propagation for partitioning massive graphs. In *WSDM*.
- , KARRER, B., BACKSTROM, L. and KLEINBERG, J. (2013). Graph cluster randomization: Network exposure to multiple universes. In *KDD*.
- VARIAN, H. R. (2007). Position auctions. *International Journal of Industrial Organization*, **25** (6), 1163–1178.
- and HARRIS, C. (2014). The vcg auction in theory and practice. *American Economic Review*, **104** (5), 442–45.
- WALKER, D. and MUCHNIK, L. (2014). Design of randomized experiments in networks. *Proceedings of the IEEE*, **102** (12), 1940–1951.
- WASSERMAN, S. and FAUST, K. (1994). *Social network analysis: Methods and applications*, vol. 8. Cambridge university press.
- WU, D.-M. (1973). Alternative tests of independence between stochastic regressors and disturbances. *Econometrica: Journal of the Econometric Society*, pp. 733–750.
- XU, Y., CHEN, N., FERNANDEZ, A., SINNO, O. and BHASIN, A. (2015). From infrastructure to culture: A/B testing challenges in large scale social networks. In *KDD*, pp. 2227–2236.
- ZIGLER, C. M. and PAPADOGEORGOU, G. (2018). Bipartite causal inference with interference. *arXiv preprint arXiv:1807.08660*.

# Appendix A

## Appendix to Chapter 1

### A.1 Review of the completely and cluster-based randomized assignments

We briefly review the notation and main results for the completely randomized (CR) design and the cluster-based randomized design (CBR). Recall that for a vector  $\mathbf{u}$  of length  $N$ , we let  $\bar{u} = \frac{1}{N} \sum_i u_i$  and we let  $\sigma^2(\mathbf{u}) = \frac{1}{N-1} \sum_i (u_i - \bar{u})^2$ .

In a Completely Randomized (CR) experiment over  $N$  units of experimentation, we assign  $n_t$  units to treatment and  $n_c = N - n_t$  units to control at random. Let the vector  $\mathbf{Z} \in \{0, 1\}^N$  be the indicator vector for whether a unit is in treatment ( $Z_i = 1$ ) or control ( $Z_i = 0$ ). Let  $\mathbf{Y}_t := \{Y_i : Z_i = 1\}$  and  $\mathbf{Y}_c := \{Y_i : Z_i = 0\}$ . The difference-in-means estimator is defined as:

$$\hat{\tau}_{cr} := \bar{\mathbf{Y}}_t - \bar{\mathbf{Y}}_c \tag{A.1}$$

It is an unbiased estimator of the causal estimand TTE under SUTVA:

$$\mathbb{E}_{\mathbf{Z} \sim cr}[\hat{\tau}_{cr}] = TTE \tag{A.2}$$

where we use the notation  $\mathbf{Z} \sim cr$  to denote a completely randomized assignment and TTE

for the total treatment effect:

$$TTE := \frac{1}{N} \sum_i Y_i(1) - Y_i(0) \quad (\text{A.3})$$

We introduce the following variance quantities  $S_t := \sigma^2(\mathbf{Y}(1))$ ,  $S_c := \sigma^2(\mathbf{Y}(0))$  and  $S_{tc} := \sigma^2(\mathbf{Y}(1) - \mathbf{Y}(0))$ , where  $\mathbf{Y}(1) - \mathbf{Y}(0)$  is the element-wise difference between vectors  $\mathbf{Y}(1)$  and  $\mathbf{Y}(0)$ . The true variance is given by:

$$\sigma_{cr}^2 := \text{Var}_{\mathbf{Z} \sim cr}[\hat{\tau}_{cr}] = \frac{S_t}{n_t} + \frac{S_c}{n_c} - \frac{S_{tc}}{N} \quad (\text{A.4})$$

We let  $S := \sigma^2(\mathbf{Y})$  be the variance of the observed outcomes. The variance under Fisher's null of no treatment effect is given by:

$$\text{Var}_{\mathbf{Z} \sim cr}[\hat{\tau}_{cr}] := \frac{N}{n_c n_t} S \quad (\text{A.5})$$

We let  $\hat{S}_t := \sigma^2(\mathbf{Y}_t)$  and  $\hat{S}_c := \sigma^2(\mathbf{Y}_c)$ . They are unbiased estimators of  $S_t$  and  $S_c$  respectively, under the completely randomized assignment distribution:

$$\begin{aligned} \mathbb{E}_{\mathbf{Z} \sim cr}[\hat{S}_t] &= S_t \\ \mathbb{E}_{\mathbf{Z} \sim cr}[\hat{S}_c] &= S_c \end{aligned}$$

If we consider the following quantity,

$$\hat{\sigma}_{cr}^2 := \frac{\hat{S}_t}{N_t} + \frac{\hat{S}_c}{N_c} \quad (\text{A.6})$$

then, if SUTVA holds,

$$\mathbb{E}_{\mathbf{Z} \sim cr}[\hat{\sigma}_{cr}^2] \geq \text{Var}_{\mathbf{Z}}[\hat{\tau}_{cr}] \quad (\text{A.7})$$

The cluster-based randomized assignment differs from the completely randomized assignment in that it randomizes over clusters of units in the graph, rather than individual units. We suppose that the experimental units are partitioned into  $M$  clusters. We assign  $m_t$  clusters to treatment and  $M - m_c$  to control completely at random. Each unit is given the intervention assigned to its cluster. Let  $\mathbf{z} \in \{0, 1\}^M$  be the indicator vector of whether a cluster is assigned to treatment ( $\mathbf{z} = 1$ ) or control ( $\mathbf{z} = 0$ ). Furthermore, let  $\mathbf{Y}^+ \in \mathbb{R}^M$  be

the vector of *aggregated* outcomes, defined as:

$$Y_j^+ := \sum_{i \in \mathcal{C}_j} Y_i \quad (\text{A.8})$$

Let  $\mathbf{Y}_t^+ := \{Y_j^+ : z_j = 1\}$ ,  $\mathbf{Y}_c^+ := \{Y_j^+ : z_j = 0\}$ . An unbiased estimator of the causal estimand  $\tau$  under SUTVA is given by:

$$\hat{\tau}_{cbr} := \frac{M}{N} \left( \overline{\mathbf{Y}_t^+} - \overline{\mathbf{Y}_c^+} \right) \quad (\text{A.9})$$

In other words, the following equality holds under SUTVA:

$$\mathbb{E}_{\mathbf{Z} \sim cbr} [\hat{\tau}_{cbr}] = TTE \quad (\text{A.10})$$

where we let  $\mathbf{Z} \sim cbr$  denote a cluster-based randomized assignment. Note that this holds true regardless of whether the clustering is balanced. We introduce  $S_t^+ = \sigma^2(\mathbf{Y}^+(1))$ ,  $S_c^+ = \sigma^2(\mathbf{Y}^+(0))$  and  $S_{tc}^+ = \sigma^2(\mathbf{Y}^+(1) - \mathbf{Y}^+(0))$ . The true variance, under SUTVA, is given by:

$$\sigma_{cbr}^2 := \text{Var}_{\mathbf{Z} \sim cbr} [\hat{\tau}_{cbr}] = \frac{M^2}{N^2} \left( \frac{S_t^+}{m_t} + \frac{S_c^+}{m_c} - \frac{S_{tc}^+}{M} \right) \quad (\text{A.11})$$

We let  $S^+ = \sigma^2(\mathbf{Y}^+)$  be the variance of the observed *aggregated* outcomes. The variance under Fisher's Null of no treatment effect is given by:

$$\text{Var}_{\mathbf{Z} \sim cbr} [\hat{\tau}_{cbr}] := \frac{M^2}{N^2} \frac{M}{M_c M_t} S^+ \quad (\text{A.12})$$

We introduce the following quantities:  $\hat{S}_t^+ = \sigma^2(\mathbf{Y}_t^+)$  and  $\hat{S}_c^+ = \sigma^2(\mathbf{Y}_c^+)$ . If we consider the following quantity:

$$\hat{\sigma}_{cbr}^2 := \frac{M^2}{N^2} \left( \frac{\hat{S}_t^+}{M_t} + \frac{\hat{S}_c^+}{M_c} \right) \quad (\text{A.13})$$

then, if SUTVA holds,

$$\mathbb{E}_{\mathbf{Z} \sim cbr} [\hat{\sigma}_{cbr}^2] \geq \text{Var}_{\mathbf{Z}} [\hat{\tau}_{cbr}] \quad (\text{A.14})$$

## A.2 Proof of Lemma 1

For each unit  $i$ , let  $\rho_i := \frac{1}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} Z_j$ . Recall that the potential outcomes model is taken to be

$$\forall i, Y_i(\mathbf{Z}) = \alpha + \beta Z_i + \gamma \rho_i + \epsilon_i$$

The difference-in-means estimator can be rewritten as

$$\mathbb{E}_{\mathbf{Z} \sim cr} [\hat{\tau}_{cr}] = \mathbb{E}_{\mathbf{Z} \sim cr} \left[ \sum_{i \in G} \frac{(-1)^{Z_i}}{n_t^{Z_i} n_c^{(1-Z_i)}} \left( \alpha + \beta Z_i + \frac{\gamma}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} Z_j \right) + \epsilon_i \right]$$

We have  $\mathbb{E}_{\mathbf{Z} \sim cr} [\epsilon_i] = 0$ . Furthermore,

$$\begin{aligned} \mathbb{E}_{\mathbf{Z} \sim cr} \left[ \sum_{i=1}^N \frac{(-1)^{Z_i}}{n_t^{Z_i} n_c^{(1-Z_i)}} \alpha \right] &= \alpha \sum_{i=1}^N \left( \frac{1}{n_t} \frac{n_t}{N} - \frac{1}{n_c} \frac{n_c}{N} \right) = 0 \\ \mathbb{E}_{\mathbf{Z} \sim cr} \left[ \sum_{i=1}^N \frac{(-1)^{Z_i}}{n_t^{Z_i} n_c^{(1-Z_i)}} \beta Z_i \right] &= \beta \sum_{i=1}^N \frac{1}{n_t} \frac{n_t}{N} = \beta \\ \mathbb{E}_{\mathbf{Z} \sim cr} \left[ \sum_{i=1}^N \frac{(-1)^{Z_i}}{n_t^{Z_i} n_c^{(1-Z_i)}} \gamma \frac{1}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} Z_j \right] &= \gamma \sum_{i=1}^N \frac{1}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} \frac{1}{n_t} \frac{n_t}{N} \frac{n_t - 1}{N - 1} - \frac{1}{n_c} \frac{n_c}{N} \frac{n_t}{N - 1} \\ &= -\frac{\gamma}{N - 1} \end{aligned}$$

Hence, we obtain Equation 1.10. We repeat the analysis with the Horvitz-Thompson estimator:

$$\begin{aligned} \mathbb{E}_{\mathbf{Z} \sim cbr} [\hat{\tau}_{cbr}] &= \mathbb{E}_{\mathbf{Z}} \left[ \frac{M}{N} \sum_{i=1}^N \frac{(-1)^{Z_i}}{m_t^{Z_i} m_c^{(1-Z_i)}} \left( \alpha + \beta Z_i + \frac{\gamma}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} Z_j \right) \right] \\ &= \beta + \frac{\gamma}{N} \sum_{i=1}^N \frac{|\mathcal{N}_i \cap \mathcal{C}(i)|}{|\mathcal{N}_i|} - \frac{\gamma}{N(M-1)} \sum_{i=1}^N \frac{|\mathcal{N}_i \setminus \mathcal{C}(i)|}{|\mathcal{N}_i|} \\ &= \beta - \gamma \left( \rho_C - \frac{1}{M-1} (1 - \rho_C) \right) \\ &= \beta - \gamma \frac{1 - \rho_C \cdot M}{M - 1} \end{aligned}$$

where we have defined  $\rho_C = \frac{1}{N} \sum_{i \in G} \frac{|\mathcal{N}_i \cap \mathcal{C}(i)|}{|\mathcal{N}_i|}$ .

### A.3 Proof of Proposition 2

Under SUTVA, the test statistic  $T := \frac{\hat{\tau}_{br} - \hat{\tau}_{cbr}}{\sqrt{\hat{\sigma}^2}}$  is a random variable with mean 0 and variance smaller than 1 if  $\hat{\sigma}^2 \geq \text{Var}_{\mathbf{W}, \mathbf{Z}}[\Delta]$ . Chebyshev's inequality implies:

$$\forall t \in \mathbb{R}^+, \mathbb{P}(|T| \geq t) \leq \frac{1}{t^2}$$

It follows that, if SUTVA holds, and we reject with  $\{|T| \geq \frac{1}{\sqrt{\alpha}}\}$ , then we reject with probability less than  $\alpha$ .

### A.4 Proof of Theorem 1

Let  $m_{cr}$  and  $n_{cr}$  be the number of clusters and units respectively in treatment arm  $cr$  ( $W_i = 1$ ), and let  $m_{cbr}$  and  $n_{cbr}$  be the number of clusters and units respectively in treatment arm  $cbr$  ( $W_i = 0$ ). Recall that  $M = m_{cr} + m_{cbr}$  and  $N = n_{cr} + n_{cbr}$ . We will assume that the clustering of the graph is balanced, such that we have

$$\frac{M}{N} = \frac{m_{cr}}{n_{cr}} = \frac{m_{cbr}}{n_{cbr}}$$

Let  $S_{tc}^+ := \sigma^2(\mathbf{Y}_j^+(1) - \mathbf{Y}_j^+(0))$  be the variance of the difference of the aggregated potential outcomes.

**Theorem 1.** *If SUTVA holds, and each cluster in  $\mathcal{C}$  has the same size, then the expectation and variance of the difference-in-difference-in-means estimator are given by:*

$$\mathbb{E}_{\mathbf{W}, \mathbf{Z}}[\Delta] = 0 \tag{A.15}$$

$$\text{Var}_{\mathbf{W}, \mathbf{Z}}[\Delta] = \frac{n_{cr}}{n_{cr} - 1} \frac{M}{M - 1} \sigma_{cr}^2 + \left(1 - \frac{m_{cbr}}{N(n_{cr} - 1)}\right) \sigma_{cbr}^2 + \frac{M}{n_{cr} n_{cbr}} S_{tc}^+ \tag{A.16}$$

See Equations A.4 and A.11 for definitions of  $\sigma_{cr}^2$  and  $\sigma_{cbr}^2$ . If SUTVA holds,  $Y_i(\mathbf{Z}) = Y_i(Z_i)$ , such that

$$\mathbb{E}_{\mathbf{Z}}[\Delta | \mathbf{W}] = \frac{1}{n_{cr}} \sum_{i=1}^N W_i (Y_i(1) - Y_i(0)) - \frac{1}{n_{cbr}} \sum_{i=1}^N (1 - W_i) (Y_i(1) - Y_i(0)).$$

As a result, under SUTVA,

$$\mathbb{E}_{\mathbf{W}, \mathbf{Z}}[\Delta] = \frac{1}{N} \sum_{i=1}^N Y_i(1) - Y_i(0) - \frac{1}{N} \sum_{i=1}^N Y_i(1) - Y_i(0) = 0$$

We now compute the theoretical variance of the  $\Delta$  estimator. By Eve's law, we have:

$$\text{Var}_{\mathbf{W}, \mathbf{Z}}[\Delta] = \mathbb{E}_{\mathbf{W}}[\text{Var}_{\mathbf{Z}}[\hat{\tau}_{cr} | \mathbf{W}]] + \mathbb{E}_{\mathbf{W}}[\text{Var}_{\mathbf{Z}}[\hat{\tau}_{cbr} | \mathbf{W}]] + \text{Var}_{\mathbf{W}}[\mathbb{E}_{\mathbf{Z}}[\hat{\tau}_{cr} - \hat{\tau}_{cbr} | \mathbf{W}]]$$

We compute each term separately. Recall that  $n_{cr,t}$  and  $n_{cr,c}$  are the number of treated and control units in treatment arm  $cr$  and  $n_{cbr,t}$  and  $n_{cbr,c}$  are the number of treated and control units in treatment arm  $cbr$ .

Furthermore, we introduce the following variances of potential outcomes, restricted to the treatment arm  $cr$  and  $cbr$ . These variances are expressed conditionally on the assignment  $\mathbf{W}$  of clusters to each treatment arm. Let  $S_{cr,t} := \sigma^2(Y_i(1) : W_i = 1)$ ,  $S_{cr,c} := \sigma^2(Y_i(0) : W_i = 1)$  and  $S_{cr,tc} := \sigma^2(Y_i(1) - Y_i(0) : W_i = 1)$ . Let  $\omega \in \{0, 1\}^M$  be the cluster indicator vector for whether a cluster has been assigned to treatment arm  $cr$  ( $\omega_j = 1$ ) or assigned to treatment arm  $cbr$  ( $\omega_j = 0$ ):  $\omega$  is merely another representation, at the cluster-level, of  $\mathbf{W}$ . Let  $S_{cbr,t}^+ := \sigma^2(Y_j^+(1) : \omega_j = 0)$ ,  $S_{cbr,c}^+ := \sigma^2(Y_j^+(0) : \omega_j = 0)$  and  $S_{cbr,tc} := \sigma^2(Y_j^+(1) - Y_j^+(0) : \omega_j = 0)$ .

Conditioned on the assignment of units to treatment arms, it is a classical result of causal inference that:

$$\begin{aligned} \text{Var}_{\mathbf{Z} \sim cr}[\hat{\tau}_{cr} | \mathbf{W}] &= \frac{1}{n_{cr,t}} S_{cr,t} + \frac{1}{n_{cr,c}} S_{cr,c} - \frac{1}{n_{cr}} S_{cr,tc} \\ \text{Var}_{\mathbf{Z} \sim cbr}[\hat{\tau}_{cbr} | \mathbf{W}] &= \frac{m_{cbr}}{n_{cbr}} \left( \frac{1}{m_{cbr,t}} S_{cbr,t} + \frac{1}{m_{cbr,c}} S_{cbr,c} - \frac{1}{m_{cbr}} S_{cbr,tc} \right). \end{aligned}$$

By linearity of expectation, we now compute the expectation of each term with respect to  $\mathbf{W}$ .

We begin with the treatment arm  $cbr$ :

$$\begin{aligned} \mathbb{E}_{\mathbf{W}}[S_{cbr,t}] &= \frac{1}{m_{cbr} - 1} \sum_{j=1}^M \mathbb{E}[\omega_j] (Y_j^+(1))^2 - \frac{m_{cbr}}{m_{cbr} - 1} \mathbb{E} \left[ \overline{\mathbf{Y}_{cbr}^+(1)} \right]^2 \\ &= \frac{m_{cbr}}{m_{cbr} - 1} \overline{\mathbf{Y}^+(1)^2} - \frac{m_{cbr}}{m_{cbr} - 1} \mathbb{E} \left[ \overline{\mathbf{Y}_{cbr}^+(1)^2} \right] \end{aligned}$$

We now compute the second term, introducing the variable  $d_j := \omega_j - \frac{m_{cbr}}{M}$ .

$$\begin{aligned}
\mathbb{E}_{\mathbf{W}} \left[ \overline{\mathbf{Y}_{cbr,t}^+}^2 \right] &= \frac{1}{m_{cbr}^2} \sum_{j=1}^M \sum_{k=1}^M \mathbb{E}[\omega_j \omega_k] Y_j^+(1) Y_k^+(1) \\
&= \frac{1}{m_{cbr}^2} \sum_{j=1}^M \mathbb{E} \left[ d_j^2 + \frac{m_{cbr}^2}{M^2} \right] (Y_j^+(1))^2 + \frac{1}{m_{cbr}^2} \sum_{j=1}^M \sum_{k \neq j}^M \mathbb{E} \left[ d_j d_k + \frac{m_{cbr}^2}{M^2} \right] Y_j^+(1) Y_k^+(1) \\
&= \frac{1}{m_{cbr}^2} \left( \frac{m_{cbr}(M - m_{cbr})}{M^2} + \frac{m_{cbr}^2}{M^2} \right) \cdot M \cdot \overline{(Y^+(1))^2} \\
&\quad + \frac{1}{m_{cbr}^2} \left( -\frac{m_{cbr}(M - m_{cbr})}{m^2(M - 1)} + \frac{m_{cbr}^2}{M^2} \right) \sum_{j=1}^M \sum_{k \neq j}^M Y_j(1) Y_k(1) \\
&= \frac{1}{m_{cbr}} \cdot \overline{(Y^+(1))^2} + \frac{1}{m_{cbr}^2} \frac{m_{cbr}}{m^2} \frac{M(m_{cbr} - 1)}{M - 1} \sum_{j=1}^M \sum_{k \neq j}^M Y_j(1) Y_k(1) \\
&= \left( \frac{1}{m_{cbr}} - M \cdot \frac{m_{cbr} - 1}{m_{cbr} M (M - 1)} \right) \overline{(Y^+(1))^2} + \frac{m_{cbr} - 1}{m_{cbr} M (M - 1)} \sum_{j=1}^M \sum_{k=1}^M Y_j^+(1) Y_k^+(1) \\
&= \frac{m_{cr}}{m_{cbr}(M - 1)} \overline{(Y^+(1))^2} + \frac{M(m_{cbr} - 1)}{m_{cbr}(M - 1)} \left( \overline{Y^+(1)} \right)^2
\end{aligned}$$

Putting both terms together, we obtain:

$$\begin{aligned}
\mathbb{E}[S_{cbr,t}] &= \left( \frac{m_{cbr}}{m_{cbr} - 1} - \frac{m_{cr}}{(m_{cbr} - 1)(M - 1)} \right) \overline{(Y^+(1))^2} - \frac{M}{M - 1} \overline{Y^+(1)}^2 \\
&= \left( \frac{m_{cbr}}{m_{cbr} - 1} - \frac{m_{cr}}{(m_{cbr} - 1)(M - 1)} \right) \overline{(Y^+(1))^2} - \frac{M}{M - 1} \overline{Y^+(1)}^2 \\
&= \frac{M}{M - 1} \overline{(Y^+(1))^2} - \frac{M}{M - 1} \overline{Y^+(1)}^2 \\
&= S_t^+ := \sigma^2(\mathbf{Y}^+(1))
\end{aligned}$$

Similarly,  $\mathbb{E}[S_{cbr,c}] = S_c^+$  and  $\mathbb{E}[S_{cbr,tc}] = S_{tc}^+$ . We therefore have that:

$$\mathbb{E}_{\mathbf{W}}[\text{Var}[\hat{t}_{cbr} | \mathbf{W}]] = \frac{m_{cbr}^2}{n_{cbr}^2} \left( \frac{S_t^+}{m_{cbr,t}} + \frac{S_c^+}{m_{cbr,c}} - \frac{S_{tc}^+}{m_{cbr}} \right)$$

We now repeat the analysis for the treatment arm  $cr$ :

$$\begin{aligned}
\mathbb{E}_{\mathbf{W}}[S_{cr,t}] &= \frac{1}{n_{cr} - 1} \sum_{i=1}^N \mathbb{E}_{\mathbf{W}} \left[ W_i \left( Y_i(1) - \overline{\mathbf{Y}_{cr}(1)} \right)^2 \right] \\
&= \frac{1}{n_{cr} - 1} \sum_{i=1}^N \mathbb{E}_{\mathbf{W}}[W_i] Y_i^2(1) - \frac{n_{cr}}{n_{cr} - 1} \mathbb{E}_{\mathbf{W}} \left[ \overline{\mathbf{Y}_{cr}(1)}^2 \right].
\end{aligned}$$

The expectation of the first term is given by

$$\frac{1}{n_{cr}-1} \sum_{i=1}^N \mathbb{E}_{\mathbf{W}}[W_i] Y_i^2(1) = \frac{1}{n_{cr}-1} \sum_{i=1}^N \frac{m_{cr}}{M} Y_i^2(1) = \frac{m_{cr}}{M} \frac{N}{n_{cr}-1} \overline{\mathbf{Y}^2(1)} = \frac{n_{cr}}{n_{cr}-1} \overline{\mathbf{Y}^2(1)}.$$

We now compute the second term. Note that  $\overline{\mathbf{Y}_{cr}(1)} = \frac{m_{cr}}{n_{cr}} \overline{\mathbf{Y}_{cr}^+(1)}$  and  $\overline{\mathbf{Y}(1)} = \frac{m_{cr}}{n_{cr}} \overline{\mathbf{Y}^+(1)}$ . By analogy from the computation for the treatment arm *cbr*, we have that:

$$\begin{aligned} \mathbb{E}_{\mathbf{W}} \left[ \overline{\mathbf{Y}_{cr}(1)}^2 \right] &= \frac{m_{cr}^2}{n_{cr}^2} \left( \frac{m_{cbr}}{m_{cr}(M-1)} \overline{(\mathbf{Y}^+)^2(1)} + \frac{M(m_{cr}-1)}{m_{cr}(M-1)} \overline{\mathbf{Y}^+(1)}^2 \right) \\ &= \frac{m_{cr}^2}{n_{cr}^2} \left( \frac{m_{cbr}}{m_{cr}(M-1)} \overline{(\mathbf{Y}^+)^2(1)} + \frac{M(m_{cr}-1)}{m_{cr}(M-1)} \frac{n_{cr}^2}{m_{cr}^2} \overline{\mathbf{Y}(1)}^2 \right) \\ &= \frac{1}{n_{cr}^2} \frac{m_{cr} m_{cbr}}{M-1} \overline{(\mathbf{Y}^+)^2(1)} + \frac{N}{n_{cr}} \frac{m_{cr}-1}{M-1} \overline{\mathbf{Y}(1)}^2 \end{aligned}$$

Putting the two terms together, we obtain:

$$\begin{aligned} \mathbb{E}_{\mathbf{W}}[S_{cr,t}] &= \frac{n_{cr}}{n_{cr}-1} \overline{\mathbf{Y}^2(1)} - \frac{n_{cr}}{n_{cr}-1} \left( \frac{1}{n_{cr}^2} \frac{m_{cr} m_{cbr}}{M-1} \overline{(\mathbf{Y}^+)^2(1)} + \frac{N}{n_{cr}} \frac{m_{cr}-1}{M-1} \overline{\mathbf{Y}(1)}^2 \right) \\ &= \frac{n_{cr}}{n_{cr}-1} \overline{\mathbf{Y}^2(1)} - \frac{N}{n_{cr}-1} \frac{m_{cr}-1}{M-1} \overline{\mathbf{Y}(1)}^2 - \frac{1}{n_{cr}(n_{cr}-1)} \frac{m_{cr} m_{cbr}}{M-1} \overline{(\mathbf{Y}^+)^2(1)} \\ &= \frac{1}{n_{cr}-1} \frac{1}{M-1} \left( n_{cr}(M-1) \overline{\mathbf{Y}^2(1)} - N(m_{cr}-1) \overline{\mathbf{Y}(1)}^2 - \frac{m_{cr} m_{cbr}}{n_{cr}} \overline{(\mathbf{Y}^+)^2(1)} \right) \\ &= \frac{1}{n_{cr}-1} \frac{1}{M-1} \left( n_{cr}(M-1) \left( \overline{\mathbf{Y}^2(1)} - \overline{\mathbf{Y}(1)}^2 \right) + n_{cbr} \overline{\mathbf{Y}(1)}^2 - \frac{m_{cr} m_{cbr}}{n_{cr}} \overline{(\mathbf{Y}^+)^2(1)} \right) \\ &= \frac{1}{n_{cr}-1} \frac{1}{M-1} \left( n_{cr}(M-1) \left( \overline{\mathbf{Y}^2(1)} - \overline{\mathbf{Y}(1)}^2 \right) + \frac{1}{n_{cr}} \left( \left( n_{cbr} n_{cr} \overline{\mathbf{Y}(1)}^2 - m_{cr} m_{cbr} \overline{(\mathbf{Y}^+)^2(1)} \right) \right) \right) \end{aligned}$$

Using again the fact that  $\overline{\mathbf{Y}(1)}^2 = \frac{m_{cr}}{n_{cr}} \frac{m_{cbr}}{n_{cbr}} \overline{\mathbf{Y}^+(1)}^2$ , we have:

$$n_{cr} n_{cbr} \overline{\mathbf{Y}(1)}^2 - m_{cr} m_{cbr} \overline{(\mathbf{Y}^+)^2(1)} = -m_{cr} m_{cbr} \left( \overline{(\mathbf{Y}^+)^2(1)} - \overline{\mathbf{Y}^+(1)}^2 \right)$$

Recalling the notation  $S_t := \sigma^2(\mathbf{Y}(1))$  and  $S_t^+ := \sigma^2(\mathbf{Y}^+(1))$ , we have:

$$\begin{aligned} \mathbb{E}_{\mathbf{W}}[S_{cr,t}] &= \frac{n_{cr}}{n_{cr}-1} \frac{M-1}{M} S_t - \frac{m_{cr} m_{cbr}}{n_{cr}(n_{cr}-1)M} S_t^+ \\ &= \frac{n_{cr}}{n_{cr}-1} \frac{M-1}{M} S_t - \frac{m_{cbr}}{N(n_{cr}-1)} S_t^+ \end{aligned}$$

We obtain an identical expression for  $\mathbb{E}_{\mathbf{W}}[S_{cr,c}]$ :

$$\mathbb{E}_{\mathbf{W}}[S_{cr,c}] = \frac{n_{cr}}{n_{cr}-1} \frac{M-1}{M} S_c - \frac{m_{cbr}}{N(n_{cr}-1)} S_c^+$$

where  $S_c := \sigma^2(\mathbf{Y}(0))$ . Furthermore, letting  $S_{tc} := \sigma^2(\mathbf{Y}(1) - \mathbf{Y}(0))$ :

$$\mathbb{E}_{\mathbf{W}}[S_{cr,tc}] = \frac{n_{cr}}{n_{cr} - 1} \frac{M - 1}{M} S_{tc} - \frac{m_{cbr}}{N(n_{cr} - 1)} S_{tc}^+$$

Finally, we compute  $\text{Var}_{\mathbf{W}}[\mathbb{E}_{\mathbf{Z}}[\hat{\tau}_{cr} - \hat{\tau}_{cbr} | \mathbf{W}]]$ .

$$\begin{aligned} \text{Var}_{\mathbf{W}}[\mathbb{E}_{\mathbf{Z}}[\hat{\tau}_{cr} - \hat{\tau}_{cbr} | \mathbf{W}]] &= \text{Var}_{\mathbf{W}} \left[ \frac{1}{n_{cr}} \sum_{i \in G} W_i (Y_i(1) - Y_i(0)) - \frac{1}{n_{cbr}} \sum_{i \in G} (1 - W_i) (Y_i(1) - Y_i(0)) \right] \\ &= \text{Var}_{\mathbf{W}} \left[ \frac{1}{N} \sum_{i=1}^N \left( W_i - \frac{m_{cr}}{M} \right) \left( \frac{M}{m_{cr}} + \frac{M}{m_{cbr}} \right) (Y_i(1) - Y_i(0)) \right] \\ &= \text{Var}_{\mathbf{W}} \left[ \frac{1}{N} \sum_{i \in G} D_i Y'_i \right] \quad \text{where } Y'_i = \frac{M^2}{m_{cr} m_{cbr}} (Y_i(1) - Y_i(0)) \text{ and } D_i = W_i - \frac{m_{cr}}{M} \\ &= \frac{M^2}{N^2} \text{Var}_{\mathbf{W}} \left[ \frac{1}{M} \sum_{j=1}^M d_j Y_j'^+ \right] \\ &= \frac{M^2}{N^2} \frac{m_{cr} m_{cbr}}{M^3 (M - 1)} \sum_{j=1}^M \left( Y_j'^+(1) - Y_j'^+(0) - \left( \overline{\mathbf{Y}'^+(1)} - \overline{\mathbf{Y}'^+(0)} \right) \right)^2 \\ &= \frac{M^2}{N^2} \frac{m_{cr} m_{cbr}}{M^3 (M - 1)} \frac{M^4}{m_{cr}^2 m_{cbr}^2} \sum_{j=1}^M \left( Y_j^+(1) - Y_j^+(0) - \left( \overline{\mathbf{Y}^+(1)} - \overline{\mathbf{Y}^+(0)} \right) \right)^2 \\ &= \frac{M}{n_{cr} n_{cbr}} S_{tc}^+ \quad \text{where } S_{tc}^+ = \sigma^2(\mathbf{Y}^+(1) - \mathbf{Y}^+(0)) \end{aligned}$$

This concludes our proof of Theorem 1.

## A.5 Proof of Theorem 2

Recall the notation introduced in the proof of Theorem 1. We let  $S_{cr,t} := \sigma^2(Y_i(1) : W_i = 1)$ ,  $S_{cr,c}(0) := \sigma^2(Y_i(0) : W_i = 1)$  and  $S_{cr,tc} := \sigma^2(Y_i(1) - Y_i(0) : W_i = 1)$ . Furthermore, we let  $S_{cbr,t}^+ := \sigma^2(Y_j^+(1) : \omega_j = 0)$ ,  $S_{cbr,c}^+ := \sigma^2(Y_j^+(0) : \omega_j = 0)$  and  $S_{cbr,tc} := \sigma^2(Y_j(1) - Y_j(0) : \omega_j = 0)$ . Let us rewrite the formula for the theoretical variance using this notation:

$$\begin{aligned} \text{Var}_{\mathbf{W}, \mathbf{Z}}[\Delta] &= \frac{1}{n_{cr,t}} \mathbb{E}_{\mathbf{W}}[S_{cr,t}] + \frac{1}{n_{cr,c}} \mathbb{E}_{\mathbf{W}}[S_{cr,c}] - \frac{1}{n_{cr}} \mathbb{E}_{\mathbf{W}}[S_{cr,tc}] \\ &\quad + \frac{m_{cbr}^2}{n_{cbr}^2} \left( \frac{1}{m_{cbr,t}} \mathbb{E}_{\mathbf{W}}[S_{cbr,t}] + \frac{1}{m_{cbr,c}} \mathbb{E}_{\mathbf{W}}[S_{cbr,c}^+] - \frac{1}{m_{cbr}} \mathbb{E}_{\mathbf{W}}[S_{cbr,tc}^+] \right) + \text{Var}_{\mathbf{W}}[\mathbb{E}_{\mathbf{Z}}[\hat{\tau}_{cr} - \hat{\tau}_{cbr} | \mathbf{W}]] \end{aligned}$$

We begin by noting that  $\mathbb{E}_{\mathbf{W}}[S_{cr,tc}] \geq 0$ . Furthermore, we have the following equalities from the previous proof:

$$\begin{aligned}\text{Var}_{\mathbf{W}}[\mathbb{E}_{\mathbf{Z}}[\hat{\tau}_{cr} - \hat{\tau}_{cbr}|\mathbf{W}]] &= \frac{m}{n_{cr}n_{cbr}}S_{tc}^+ \\ \mathbb{E}_{\mathbf{W}}[S_{cbr,tc}^+] &= S_{tc}^+\end{aligned}$$

We observe that:

$$\left(-\frac{m_{cbr}^2}{n_{cbr}^2} \frac{1}{m_{cbr}} + \frac{M}{n_{cr}n_{cbr}}\right) S_{tc}^+ < 0$$

Thus, it follows that:

$$\text{Var}_{\mathbf{W},\mathbf{Z}}[\Delta] \leq \frac{1}{n_{cr,t}}\mathbb{E}_{\mathbf{W}}[S_{cr,t}] + \frac{1}{n_{cr,c}}\mathbb{E}_{\mathbf{W}}[S_{cr}(0)] + \frac{m_{cbr}^2}{n_{cbr}^2} \left( \frac{1}{m_{cbr,t}}\mathbb{E}_{\mathbf{W}}[S_{cbr,c}^+] + \frac{1}{m_{cbr,c}}\mathbb{E}_{\mathbf{W}}[S_{cbr}^+(0)] \right)$$

We introduce the following notation:  $\hat{S}_{cr,t} := \sigma^2(Y_i : W_i = 1 \wedge Z_i = 1)$ ,  $\hat{S}_{cr,c} := \sigma^2(Y_i : W_i = 1 \wedge Z_i = 0)$ ,  $\hat{S}_{cbr,t} := \sigma^2(Y_i : W_i = 0 \wedge Z_i = 1)$ ,  $\hat{S}_{cbr,c} := \sigma^2(Y_i : W_i = 0 \wedge Z_i = 0)$ . Let us consider the following empirical quantity:

$$\hat{\sigma}^2 := \frac{\hat{S}_{cr,t}}{n_{cr,t}} + \frac{\hat{S}_{cr,c}}{n_{cr,c}} + \frac{m_{cbr}^2}{n_{cbr}^2} \left( \frac{\hat{S}_{cbr,t}^+}{m_{cbr,t}} + \frac{\hat{S}_{cbr,c}^+}{m_{cbr,c}} \right)$$

Note that  $\mathbb{E}_{\mathbf{Z}}[\hat{S}_{cr,t}|\mathbf{W}] = S_{cr,c}$ ,  $\mathbb{E}_{\mathbf{Z}}[\hat{S}_{cr,c}|\mathbf{W}] = S_{cr}(0)$ ,  $\mathbb{E}_{\mathbf{Z}}[\hat{S}_{cbr,t}^+|\mathbf{W}] = S_{cbr,c}^+$ ,  $\mathbb{E}_{\mathbf{Z}}[\hat{S}_{cbr,c}^+|\mathbf{W}] = S_{cbr}^+(0)$ . As a result,  $\mathbb{E}_{\mathbf{W},\mathbf{Z}}[\hat{\sigma}^2] \geq \text{Var}_{\mathbf{W},\mathbf{Z}}[\Delta]$ .

## A.6 Proof of Theorem 3

Under Fisher's null,  $\forall i, Y_i(1) = Y_i(0)$ . From Theorem 1, we obtain by simple substitution:

$$\text{Var}_{\mathbf{W},\mathbf{Z}}[\Delta] = \frac{n_{cr}}{n_{cr}-1} \frac{M}{M-1} \frac{n_{cr}}{n_{cr,t}n_{cr,c}} S + \left(1 - \frac{m_{cbr}}{N(n_{cr}-1)}\right) \frac{m_{cbr}}{m_{cbr,t}m_{cbr,c}} S^+$$

where  $S := \sigma^2(\mathbf{Y})$  is the variance of all observed potential outcomes,  $S^+ := \sigma^2(\mathbf{Y}^+)$  is the variance of all observed *aggregated* outcomes.

## A.7 Proof of Theorem 4

We now compute the expectation of both  $\hat{\tau}_{cr}$  and  $\hat{\tau}_{cbr}$  under their respective completely randomized and cluster-based randomized assignment assuming the following model of interference:

$$\forall i, Y_i(\mathbf{Z}) = \alpha + \beta Z_i + \gamma \rho_i + \epsilon_i$$

We first decompose  $\Delta$  as a sum of three differences:

$$\Delta = (\hat{\tau}_{cr,\alpha} - \hat{\tau}_{cbr,\alpha}) + (\hat{\tau}_{cr,\beta} - \hat{\tau}_{cbr,\beta}) + (\hat{\tau}_{cr,\gamma} - \hat{\tau}_{cbr,\gamma})$$

which are defined as:

$$\begin{aligned} \hat{\tau}_{cr,\alpha} &:= \sum_{i=1}^N W_i (-1)^{1-Z_i} \frac{1}{n_{cr,t}^{Z_i}} \frac{1}{n_{cr,c}^{1-Z_i}} (\alpha + \epsilon_i) \\ \hat{\tau}_{cr,\beta} &:= \beta \sum_{i=1}^N W_i (-1)^{1-Z_i} \frac{1}{n_{cr,t}^{Z_i}} \frac{1}{n_{cr,c}^{1-Z_i}} Z_i \\ \hat{\tau}_{cr,\gamma} &:= \gamma \sum_{i=1}^N W_i (-1)^{1-Z_i} \frac{1}{n_{cr,t}^{Z_i}} \frac{1}{n_{cr,c}^{1-Z_i}} \frac{1}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} Z_j \\ \hat{\tau}_{cbr,\alpha} &:= \frac{m_{cbr}}{n_{cbr}} \sum_{i=1}^N (1 - W_i) (-1)^{1-Z_i} \frac{1}{m_{cbr,t}^{Z_i}} \frac{1}{m_{cbr,c}^{1-Z_i}} (\alpha + \epsilon_i) \\ \hat{\tau}_{cbr,\beta} &:= \beta \frac{m_{cbr}}{n_{cbr}} \sum_{i=1}^N (1 - W_i) (-1)^{1-Z_i} \frac{1}{m_{cbr,t}^{Z_i}} \frac{1}{m_{cbr,c}^{1-Z_i}} Z_i \\ \hat{\tau}_{cbr,\gamma} &:= \gamma \frac{m_{cbr}}{n_{cbr}} \sum_{i=1}^N (1 - W_i) (-1)^{1-Z_i} \frac{1}{m_{cbr,t}^{Z_i}} \frac{1}{m_{cbr,c}^{1-Z_i}} \frac{1}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} Z_j \end{aligned}$$

We now compute the expectation of each term with respect to  $\{\mathbf{W}, \mathbf{Z}\}$ . For the first difference, we first compute the expectation w.r.t to  $\epsilon$ , which cancels out.

$$\mathbb{E}_{\mathbf{Z}}[\hat{\tau}_{cr,\alpha} | \mathbf{W}] = \alpha \sum_{i=1}^N W_i \left( \frac{1}{n_{cr,t}} \cdot \frac{n_{cr,t}}{n_{cr}} - \frac{1}{n_{cr,c}} \cdot \frac{n_{cr,c}}{n_{cr}} \right) = 0$$

The same goes for  $\mathbb{E}_{\mathbf{Z}}[\hat{\tau}_{cbr,\alpha}|\mathbf{W}]$  and therefore  $\mathbb{E}_{\mathbf{W},\mathbf{Z}}[\hat{\tau}_{cr,\alpha} - \hat{\tau}_{cbr,\alpha}] = 0$ . We now compute the second difference.

$$\begin{aligned}\mathbb{E}_{\mathbf{W}} [\mathbb{E}_{\mathbf{Z}}[\hat{\tau}_{cr,\beta}|\mathbf{W}]] &= \beta \sum_{i=1}^N \mathbb{E}[W_i] \left( \frac{1}{n_{cr,t}} \frac{n_{cr,t}}{n_{cr}} \right) = \beta \sum_{i=1}^N \frac{m_{cr}}{m} \frac{1}{n_{cr}} = \beta \\ \mathbb{E}_{\mathbf{W}} [\mathbb{E}_{\mathbf{Z}}[\hat{\tau}_{cbr,\beta}|\mathbf{W}]] &= \beta \frac{m_{cbr}}{n_{cbr}} \sum_{i \in G} \mathbb{E}_{\mathbf{W}} [(1 - W_i)] \left( \frac{1}{m_{cbr,t}} \cdot \frac{m_{cbr,t}}{m_{cbr}} \right) = \beta\end{aligned}$$

Therefore,  $\mathbb{E}_{\mathbf{W},\mathbf{Z}}[\hat{\tau}_{cr,\beta} - \hat{\tau}_{cbr,\beta}] = 0$ . We compute the expectation of the last difference. In order to simplify the calculus, we will suppose that  $n_{cr} \gg 1$  and  $m_{cbr}$  such that  $Z_i$  and  $Z_j$  can be considered independent if both  $i$  and  $j$  are in the same treatment arm (but not in the same cluster in the case of treatment arm  $cbr$ ).

$$\begin{aligned}\mathbb{E}_{\mathbf{Z}}[\hat{\tau}_{cr,\gamma}|\mathbf{W}] &= \gamma \sum_{i=1}^N W_i \frac{1}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} \mathbb{E}_{\mathbf{Z}} \left[ Z_j (-1)^{1-Z_i} \frac{1}{n_{cr,t}^{Z_i}} \frac{1}{n_{cr,c}^{1-Z_i}} \middle| \mathbf{W} \right] \\ &= \gamma \sum_{i=1}^N W_i \frac{1}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} W_j \cdot 0 + (1 - W_j) \cdot 0 \\ &= 0\end{aligned}$$

We now compute the expectation under the second arm:

$$\begin{aligned}\mathbb{E}_{\mathbf{Z}}[\hat{\tau}_{cbr,\gamma}|\mathbf{W}] &= \gamma \frac{m_{cbr}}{n_{cbr}} \sum_{i \in G} (1 - W_i) \frac{1}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} \mathbb{E}_{\mathbf{Z}} \left[ Z_j (-1)^{1-Z_i} \frac{1}{m_{cbr,t}^{Z_i}} \frac{1}{m_{cbr,c}^{1-Z_i}} \middle| \mathbf{W} \right] \\ &= \gamma \frac{m_{cbr}}{n_{cbr}} \sum_{i \in G} (1 - W_i) \frac{1}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i \cap C(i)} \frac{1 - W_j}{m_{cbr}}\end{aligned}$$

Taking the expectation over  $\mathbf{W}$ ,

$$\begin{aligned}\mathbb{E}_{\mathbf{W},\mathbf{Z}}[\hat{\tau}_{cbr,\gamma}] &= \gamma \frac{1}{n_{cbr}} \sum_{i=1}^N \frac{1}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i \cap C(i)} \mathbb{E}_{\mathbf{W}}[(1 - W_i)^2] \\ &= \gamma \frac{1}{n_{cbr}} \sum_{i=1}^N \frac{1}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i \cap C(i)} \frac{n_{cbr}}{N} \\ &= \gamma \frac{1}{N} \sum_{i=1}^N \frac{|\mathcal{N}_i \cap C(i)|}{|\mathcal{N}_i|}\end{aligned}$$

If, for any clustering  $C$ , we let  $\rho_C := \frac{1}{n} \sum_{i \in G} \frac{|\mathcal{N}_i \cap C(i)|}{|\mathcal{N}_i|}$  be the average fraction of a node's neighbors contained in the cluster, then

$$\mathbb{E}_{\mathbf{W}, \mathbf{Z}} [\hat{\tau}_{cr} - \hat{\tau}_{cbr}] \approx \rho_C \cdot \gamma,$$

where the approximation made is that  $n_{cr} \gg 1$ ,  $m_{cbr} \gg 1$ , and  $m_{cr} \gg 1$ .

## A.8 Proof of Theorem 5

We compare the variance of the difference-in-means estimator under the Completely Randomized (CR) assignment to its variance under a Bernoulli Randomized (BR) assignment. In order to avoid situations where no units are assigned to treatment or control, we condition on the event that at least one unit is assigned to treatment and control. In other words, BR denotes the *re-randomized* Bernoulli assignment strategy, which rejects the edge cases where all units are assigned to treatment or control.

Let  $\eta_t$  (resp.  $\eta_c$ ) be the *realized* number of units assigned to treatment (resp. control) under the BR assignment, and  $N_t$  be the desired number of units assigned to treatment under the CR assignment. Naturally,  $\mathbb{E}[\eta_t] = n_t$ . Let  $p := \frac{n_t}{N}$ , where  $N$  is the total number of units. Under the BR assignment, using Eve's law, we have the following:

$$\text{Var}_{\mathbf{Z}}[\hat{\tau}] = \text{Var}_{\eta_t} [\mathbb{E}_{\mathbf{Z}} [\hat{\tau} | \eta_t]] + \mathbb{E}_{\eta_t} [\text{Var}_{\mathbf{Z}} [\hat{\tau} | \eta_t]] \quad (\text{A.17})$$

We now examine the first term of Equation A.17:

$$\mathbb{E}_{\mathbf{Z}}[\hat{\tau} | \eta_t] = \frac{1}{\eta_t} \sum_i \mathbb{E}_{\mathbf{Z}}[Z_i | \eta_t] Y_i(1) - \frac{1}{\eta_c} \sum_i \mathbb{E}_{\mathbf{Z}}[1 - Z_i | \eta_t] Y_i(0)$$

Note that:

$$\mathbb{E}_{\mathbf{Z}}[Z_i | \eta_t] = \frac{\mathbb{P}(Z_i = 1, \eta_t)}{\mathbb{P}(\eta_t)} = \frac{\binom{N-1}{\eta_t-1} p^{\eta_t} (1-p)^{(N-1)-(\eta_t-1)}}{\binom{N}{\eta_t} p^{\eta_t} (1-p)^{N-\eta_t}} = \frac{\eta_t}{N}$$

Similarly,  $\mathbb{E}_{\mathbf{Z}}[1 - Z_i | \eta_t] = \frac{\eta_c}{N}$ . As a result,  $\mathbb{E}_{\mathbf{Z}}[\hat{\tau} | \eta_t] = \frac{1}{N} \sum_i Y_i(1) - Y_i(0)$ . And thus,

$$\text{Var}_{\eta_t} [\mathbb{E}_{\mathbf{Z}} [\hat{\tau} | \eta_t]] = 0$$

We now examine the second term of Eq. A.17:

$$\mathbb{E}_{\eta_t} \left[ \frac{S_t}{\eta_t} + \frac{S_c}{\eta_c} - \frac{S_{tc}}{N} \right] = S_t \mathbb{E}_{\eta_t} \left[ \frac{1}{\eta_t} \right] + S_c \mathbb{E}_{\eta_c} \left[ \frac{1}{\eta_c} \right] - \frac{S_{tc}}{N}$$

We need to compute  $\mathbb{E} \left[ \frac{1}{\eta_t} \right]$ . We rely on the following crude-upper bound given by Lemma 3:

**Lemma 3.** *If  $p^N + (1-p)^N \leq \frac{1}{N^2} \leq \frac{1}{4}$ , we have the crude upper-bound:*

$$\left| \mathbb{E}_{\eta_t} \left[ \frac{1}{\eta_t} \right] - \frac{1}{n_t} \right| \leq \frac{5}{n_t^2}$$

It follows therefore that:

$$\forall N \geq 2, \forall p \in (0,1), |\text{Var}_{\mathbf{Z} \sim BR}[\hat{\tau}] - \text{Var}_{\mathbf{Z} \sim CR}[\hat{\tau}]| \leq 5 \left( \frac{S_t}{n_t^2} + \frac{S_c}{n_{cbr}^2} \right)$$

### Proof of Lemma 3

If  $X$  is a binomial of parameters  $B(n, p)$ , let  $p_k := \mathbb{P}(X = k)$  and let  $\alpha := p^n + (1-p)^n$ . It is easy to show that:

$$\begin{aligned} \forall k \in [1, n-1], \mathbb{P}(n_t = k) &= \sum_{i=0}^{+\infty} \mathbb{P}(i^{\text{th}} \text{ throw} = k | \text{first } i-1 \text{ throws} = 0 \text{ or } 1) \\ &= \sum_{i=0}^{+\infty} p_k \prod_{j=0}^{i-1} (p^n + (1-p)^n) \\ &= p_k \sum_{i=0}^{+\infty} (p^n + (1-p)^n)^i \\ &= \frac{p_k}{1 - p^n - (1-p)^n} \\ &= \frac{p_k}{1 - \alpha} \end{aligned}$$

As expected  $n_t$  behaves *almost* like a binomial distribution when  $n \rightarrow +\infty$ . There is a known closed form formula for the first negative moment of a binomial distribution from Chao and Strawderman (1972). For a binomial  $X$  of parameters  $(n, p)$ ,

$$\mathbb{E}_X \left[ \frac{1}{1+X} \right] = \frac{1}{p(n+1)} \left( 1 - (1-p)^{n+1} \right)$$

Let  $X \sim B(n, p)$ :

$$\left| \mathbb{E}_{n_t} \left[ \frac{1}{n_t} \right] - \frac{1}{n_1} \right| = \left| \mathbb{E}_{n_t} \left[ \frac{1}{n_t} \right] - \frac{1}{1-\alpha} \mathbb{E} \left[ \frac{1}{1+X} \right] \right| + \left| \frac{1}{1-\alpha} \mathbb{E} \left[ \frac{1}{1+X} \right] - \frac{1}{n_1} \right|$$

We study the second term:

$$\left| \frac{1}{1-\alpha} \mathbb{E} \left[ \frac{1}{1+X} \right] - \frac{1}{n_1} \right| = \frac{\alpha}{1-\alpha} \frac{1}{n_1+p} - \frac{1}{1-\alpha} \frac{(1-p)^{n+1}}{n_1+p} \leq \frac{p^n}{(1-\alpha)n}$$

We now study the first term:

$$\begin{aligned} \left| \mathbb{E} \left[ \frac{1}{X} \right] - \frac{1}{1-\alpha} \mathbb{E} \left[ \frac{1}{1+X} \right] \right| &= \left| \mathbb{E} \left[ \frac{1}{X} \right] - \frac{1}{1-\alpha} \mathbb{E}_X \left[ \frac{1}{1+X} \right] \right| \\ &= \frac{1}{1-\alpha} \left| \sum_{k=1}^{n-1} \frac{p_k}{k} - \sum_{k=0}^n \frac{p_k}{k+1} \right| \\ &= \frac{1}{1-\alpha} \left( \sum_{k=1}^{n-1} \frac{p_k}{k(k+1)} + \frac{p^n}{n+1} + (1-p)^n \right) \end{aligned}$$

We find a crude  $O(\frac{1}{n^2})$ -upper-bound of the summation term:

$$\begin{aligned} \sum_{k=1}^{n-1} \frac{p_k}{k(k+1)} &= \sum_{k=1}^{n-1} \frac{1}{k(k+1)} \binom{n}{k} p^k (1-p)^{n-k} \\ &\leq \sum_{k=1}^{n-1} \frac{3}{(k+1)(k+2)} \binom{n}{k} p^k (1-p)^{n-k} \\ &= \frac{3}{(n+1)(n+2)} \sum_{k=1}^{n-1} \binom{n+2}{k+2} p^k (1-p)^{n-k} \\ &= \frac{3}{p^2(n+1)(n+2)} \sum_{k=1}^{n-1} \binom{n+2}{k+2} p^{k+2} (1-p)^{n-k} \\ &= \frac{3}{p^2(n+1)(n+2)} \sum_{k=3}^{n+1} \binom{n+2}{k} p^k (1-p)^{n+2-k} \\ &\leq \frac{3}{p^2(n+1)(n+2)} \leq \frac{3}{(np)^2} \end{aligned}$$

As a result,

$$\left| \mathbb{E} \left[ \frac{1}{X} \right] - \frac{1}{np} \right| \leq \frac{1}{1-\alpha} \left( \frac{3}{(np)^2} + \frac{p^n}{n+1} + (1-p)^n + \frac{p^n}{n} \right)$$

If  $\alpha \leq \frac{1}{n^2} \leq \frac{1}{4}$ , we can find a crude upper-bound of the RHS term:

$$\left| \mathbb{E} \left[ \frac{1}{X} \right] - \frac{1}{np} \right| \leq \frac{5}{(np)^2}$$

## Appendix B

# Appendix to Chapter 2

### B.1 Proof of Proposition 5 and 6

Assume the following interference model:

$$\forall \mathbf{Z}, Y_i(\mathbf{Z}) = \alpha_i + \beta_i \cdot Z_i + \gamma_i \frac{1}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} Z_j + \epsilon_i,$$

where  $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ . Recall the definition of the estimand:  $TTE = \frac{1}{N} \sum_i Y_i(\mathbf{1}) - Y_i(\mathbf{0})$ . Plugging in the expression for  $Y_i(\mathbf{Z})$ , we obtain:  $TTE = \frac{1}{N} \sum_i \beta_i + \frac{1}{N} \sum_i \gamma_i$ . The estimator is given by:

$$\hat{\tau} = \frac{M}{N} \sum_{i=1}^N \frac{(-1)^{1-Z_i}}{m_t^{Z_i} m_c^{(1-Z_i)}} Y_i(\mathbf{Z}),$$

where  $m_t$  (resp.  $m_c$ ) is the number of clusters in treatment (resp. control). Plugging in the expression for  $Y_i(\vec{Z})$ , we obtain:

$$\mathbb{E}_{\mathbf{Z} \sim \mathcal{C}}[\hat{\tau}] = \frac{1}{N} \sum_{i=1}^N \beta_i + \frac{1}{N} \sum_{i=1}^N \gamma_i \left( \frac{|\mathcal{N}_i \cap \mathcal{C}(i)|}{|\mathcal{N}_i|} - \frac{1}{M-1} \frac{|\mathcal{N}_i \setminus \mathcal{C}(i)|}{|\mathcal{N}_i|} \right)$$

We obtain the desired result by taking the difference between these quantities. Prop. 5 follows by substituting  $\gamma_i = \gamma$ .

## B.2 Proof of Proposition 7

The proposition can be established by rewriting the definition of  $\mathcal{P}$ -increasing interference mechanisms,

$$TTE - \mathbb{E}_{\mathbf{Z} \sim \mathcal{C}}[\hat{\tau}] = \frac{1}{N} \sum_{i=1}^N \left( Y_i(\mathbf{1}) - \mathbb{E}_{\mathbf{Z} \sim \mathcal{C}}[Y_i(\mathbf{Z}) | z_{C(i)} = 1] \right) + \left( \mathbb{E}_{\mathbf{Z} \sim \mathcal{C}}[Y_i(\mathbf{Z}) | z_{C(i)} = 0] - Y_i(\mathbf{0}) \right)$$

It follows that a sufficient condition of the model to be  $\mathcal{P}$ -increasing is for the two following inequalities to hold:

$$Y_i(\mathbf{1}) > \mathbb{E}_{\mathbf{Z} \sim \mathcal{C}}[Y_i(\mathbf{Z}) | z_{C(i)} = 1]$$

$$Y_i(\mathbf{0}) < \mathbb{E}_{\mathbf{Z} \sim \mathcal{C}}[Y_i(\mathbf{Z}) | z_{C(i)} = 0]$$

If increasing the number of treated units in that unit's neighborhood increases that unit's outcome — holding that unit's treatment assignment constant — then the two previous inequalities hold.

## B.3 Proof of Proposition 8

Recall that for  $k \in \{1, 2\}$ , our estimator can be written as:

$$\hat{\tau}_k^{\mathbf{W}} = \frac{M_k}{N_k} \sum_{i=1}^N W_i Y_i(\mathbf{Z}) \frac{(-1)^{1-Z_i}}{m_{k,t}^{Z_i} m_{k,c}^{1-Z_i}},$$

where  $m_{k,t}$  (resp.  $m_{k,c}$ ) is the number of treated (resp. control) clusters in design arm  $k$  and  $N_k$  is the number of units in design arm  $k$ . We begin by first considering the no-interference case. We have that:

$$\mathbb{E}_{\mathbf{Z} \sim \mathcal{C}_k^{\mathbf{W}}}[\hat{\tau}_k | \mathbf{W}] = \frac{1}{N_k} \sum_{i=1}^N W_i (Y_i(1) - Y_i(0)).$$

By the law of iterated expectations, we have  $\mathbb{E}_{\mathbf{W}, \mathbf{Z} \sim \mathcal{C}_k^{\mathbf{W}}}[\hat{\tau}_k^{\mathbf{W}}] = TTE$ . We now consider the linear model suggested in Eq. 2.7, where we assume heterogeneous network effects  $(\gamma_i)_{i \in [1, N]}$ . From the proof of Proposition 6, we have that

$$\mathbb{E}_{\mathbf{Z} \sim \mathcal{C}_k^{\mathbf{W}}}[\hat{\tau}_k^{\mathbf{W}} | \mathbf{W}] = \bar{\beta} + \frac{M_k}{M_k - 1} \frac{1}{N_k} \sum_i W_i \gamma_i (\rho_{C_k^{\mathbf{W}}, i} - 1)$$

Note that we have:

$$\mathbb{E}_{\mathbf{W}}[W_i \rho_{C_k^{\mathbf{W}}, i}] = \frac{N_k(N_k - 1)}{N(N - 1)} \rho_{C_k, i}.$$

It follows that, if  $M_1 \gg 1$ ,  $M_2 \gg 1$ , and  $N_1 = N_2 = \frac{N}{2}$ ,

$$\begin{aligned} \mathbb{E}_{\mathbf{W}, \mathbf{Z} \sim \mathcal{C}_1^{\mathbf{W}}}[\hat{\tau}_1^{\mathbf{W}}] - \mathbb{E}_{\mathbf{W}, \mathbf{Z} \sim \mathcal{C}_2^{\mathbf{W}}}[\hat{\tau}_2^{\mathbf{W}}] &\approx \frac{1}{2N} \sum_i \gamma_i \rho_i \\ &\approx \mathbb{E}_{\mathbf{Z} \sim \mathcal{C}_1}[\hat{\tau}] - \mathbb{E}_{\mathbf{Z} \sim \mathcal{C}_2}[\hat{\tau}] \end{aligned}$$

We conclude that the linear model of interference is transitive.

## B.4 Proof of Proposition 9

Without specifying a parametric model of interference, theoretical bounds on the power of even the simplest randomized experiment are hard to come by. While the joint assumption of monotonicity and transitivity allow us to design a sensible test for detecting the better of two partitions, they are not sufficient to bound its power without stronger assumptions. We thus rely on simulations, like the ones run in Section 2.3, or theoretical approximations, like the ones suggested in Prop. 9. It approximates  $\mathbb{E}_{\mathbf{W}, \mathbf{Z}}[\hat{\tau}_k^{\mathbf{W}}]$ , for  $k \in \{1, 2\}$  by two independently-distributed Gaussian variables of mean  $\hat{\tau}_k^{\mathbf{W}}$  and variance  $\hat{\sigma}_k^{\mathbf{W}}$ , given in Eq. 2.14. Their difference has the distribution  $\mathcal{N}(\hat{\tau}_1^{\mathbf{W}} - \hat{\tau}_2^{\mathbf{W}}, \hat{\sigma}_1^{\mathbf{W}} + \hat{\sigma}_2^{\mathbf{W}})$ . Recall that Neyman's variance estimator is an upper-bound of the true variance, under SUTVA, in expectation over the assignment  $\mathbf{Z}$  (cf. Imbens and Rubin (2015)) if SUTVA holds. We prove in the lemma below that this still holds true for an experiment-of-experiments assignment.

**Lemma 4.** *Under SUTVA, Neyman's variance estimator is an upper-bound in expectation of the true variance of the HT estimator:*

$$\mathbb{E}_{\mathbf{W}, \mathbf{Z}}[\hat{\sigma}_k^{\mathbf{W}}] \geq \text{Var}_{\mathbf{W}, \mathbf{Z}}[\hat{\tau}_k^{\mathbf{W}}]$$

*Proof.* By Eve's law,

$$\text{Var}_{\mathbf{W}, \mathbf{Z}}[\hat{\tau}_k^{\mathbf{W}}] = \mathbb{E}_{\mathbf{W}}[\text{Var}_{\mathbf{Z} \sim \mathcal{C}_k^{\mathbf{W}}}[\hat{\tau}_k^{\mathbf{W}} | \mathbf{W}]] + \text{Var}_{\mathbf{W}}[\mathbb{E}_{\mathbf{Z} \sim \mathcal{C}_k^{\mathbf{W}}}[\hat{\tau}_k^{\mathbf{W}}]].$$

From Imbens and Rubin (2015), the first term can be equal to:

$$\frac{M_k}{N_k} \left( \frac{\text{Var}[\mathbf{Y}^+(1)]}{M_{k,t}} + \frac{\text{Var}[\mathbf{Y}^+(0)]}{M_{k,c}} - \frac{\text{Var}[\mathbf{Y}^+(1) - \mathbf{Y}^+(0)]}{M_k} \right),$$

where  $Y_j^+(Z) = \sum_{i \in C_k^w(j)} Y_i(Z)$ , the cluster-level outcomes. The second term can be shown to be equal to  $\frac{\text{Var}[\mathbf{Y}(1) - \mathbf{Y}(0)]}{N}$ . Since we have that:

$$\mathbb{E}_{\mathbf{w}, \mathbf{z}}[\hat{\sigma}_k^2] = \frac{M_k}{N_k} \left( \frac{\text{Var}[\mathbf{Y}^+(1)]}{M_{k,t}} + \frac{\text{Var}[\mathbf{Y}^+(0)]}{M_{k,c}} \right),$$

we must prove:

$$\frac{\text{Var}[\mathbf{Y}^+(1) - \mathbf{Y}^+(0)]}{N_k} \geq \frac{\text{Var}[\mathbf{Y}(1) - \mathbf{Y}(0)]}{N}.$$

This follows from an application of the Cauchy-Schwarz inequality for balanced clusters:  $\sum_j (\sum_i Y_i)^2 \leq \sum_j |C_j| \sum_i Y_i^2$ , where  $C_j$  are the cluster sizes, equal to  $\frac{N}{N_k}$  in the balanced case.  $\square$

In order to determine the greater of two clusterings, we can perform two one-sided t-tests. The Bayesian approach is to compute the posterior distribution of the difference of the two estimates, using a conjugate Gaussian prior. In order to assess the impact of assuming that the two estimates are independent Gaussians, we suggest running a sensitivity analysis, which varies the value of the correlation coefficient.

## Appendix C

# Appendix to Chapter 3

### C.1 Proof of Lemma 2

We show that the variance of the treatment proportions vector  $\boldsymbol{\pi}$  is maximized, constrained to verify  $\bar{\pi} = \frac{n_t}{N}$ , only for vectors  $\boldsymbol{\pi}^* \in \{0, 1\}^M$  assigning either all of a cluster to treatment or none, assuming that a solution in  $\{0, 1\}^M$  verifying the equality constraint exists. One direction is easy. Let  $\boldsymbol{\pi}^*$  be any assignment placing all of a cluster to treatment or none, and verifying the inequality constraint.

$$\begin{aligned}\text{Var}[\boldsymbol{\pi}] &= \frac{1}{M} \sum_{j=1}^M \pi_j^2 - \left(\frac{n_t}{N}\right)^2 \\ &= \frac{1}{M} \sum_{j=1}^M \pi_j - \left(\frac{n_t}{N}\right)^2 \\ &= \frac{n_t n_c}{n^2}\end{aligned}$$

We prove the other direction. Consider  $\pi_j < \pi_i$ . Consider increasing  $\pi_i$  and decreasing  $\pi_j$  by  $\epsilon$  such that the total number of treated units is constant:  $\pi'_i = \pi_i + \epsilon$ ,  $\pi'_j = \pi_j - \epsilon$ , and  $\forall k \notin \{i, j\}, \pi'_k = \pi_k$ , such that:

$$\text{Var}[\boldsymbol{\pi}'] = \text{Var}[\boldsymbol{\pi}] + (\pi_i + \epsilon)^2 + (\pi_j - \epsilon)^2 - \pi_i^2 - \pi_j^2 = 2\epsilon^2 + 2\epsilon(\pi_i - \pi_j)$$

Since  $\pi_i > \pi_j$ ,  $\text{Var}[\boldsymbol{\pi}'] \geq \text{Var}[\boldsymbol{\pi}]$ , which concludes the proof.

## C.2 Proof of Proposition 10

The expectation of the difference-in-means estimator conditioned on the proportion of units assigned to treatment is given by:

$$\begin{aligned}
\mathbb{E}_{\mathbf{Z}} [\hat{\tau} | \boldsymbol{\pi}] &= \mathbb{E}_{\mathbf{Z}} \left[ \sum_{i=1}^n (Z_i Y_i(1) + (1 - Z_i) Y_i(0)) \frac{(-1)^{1-Z_i}}{n_t^{Z_i} n_c^{1-Z_i}} \right] \\
&= \mathbb{E}_{\mathbf{Z}} \left[ \sum_{j=1}^M \sum_{i \in \mathcal{C}_j} (Z_i Y_i(1) + (1 - Z_i) Y_i(0)) \frac{(-1)^{1-Z_i}}{n_t^{Z_i} n_c^{1-Z_i}} \right] \\
&= \frac{1}{n_t} \sum_{j=1}^M \pi_j \sum_{i \in \mathcal{C}_j} Y_i(1) - \frac{1}{n_c} \sum_{j=1}^M (1 - \pi_j) \sum_{i \in \mathcal{C}_j} Y_i(0)
\end{aligned}$$

We now introduce the permutation matrix  $P$ .

$$\begin{aligned}
\mathbb{E}_{\mathbf{Z}, \boldsymbol{\pi}} [\hat{\tau}] &= \mathbb{E}_P \left[ \frac{1}{n_t} \sum_{j,k=1}^M P_{jk} \pi_k \sum_{i \in \mathcal{C}_j} Y_i(1) - \frac{1}{n_c} \sum_{j,k=1}^M P_{jk} (1 - \pi_k) \sum_{i \in \mathcal{C}_j} Y_i(0) \right] \\
&= \frac{1}{n_t} \sum_{j,k=1}^M \frac{\pi_k}{M} \sum_{i \in \mathcal{C}_j} Y_i(1) - \frac{1}{n_c} \sum_{j,k=1}^M \frac{1 - \pi_k}{M} \sum_{i \in \mathcal{C}_j} Y_i(0) \\
&= \frac{1}{n_t} \left( \sum_{k=1}^M \frac{\pi_k}{M} \right) \left( \sum_j \sum_{i \in \mathcal{C}_j} Y_i(1) \right) - \frac{1}{n_c} \left( \sum_{k=1}^M \frac{1 - \pi_k}{M} \right) \left( \sum_j \sum_{i \in \mathcal{C}_j} Y_i(0) \right) \\
&= \frac{1}{N} \sum_i Y_i(1) - \frac{1}{N} \sum_i Y_i(0)
\end{aligned}$$

This last quantity corresponds to the total treatment effect, hence the proof that the difference-in-means estimators is unbiased under the stable unit treatment value assumption for a randomized saturation design.

## C.3 Proof of Proposition 11

Using Eve's law, we have that

$$\text{Var}_{\mathbf{Z}, \boldsymbol{\pi}} [\hat{\tau}] = \text{Var}_{\boldsymbol{\pi}} [\mathbb{E}_{\mathbf{Z}} [\hat{\tau}]] + \mathbb{E}_{\boldsymbol{\pi}} [\text{Var}_{\mathbf{Z}} [\hat{\tau}]]$$

We first compute the variance of the estimator conditional on an assignment of the treatment proportions vector  $\boldsymbol{\pi}$ . Let  $D_i = Z_i - \frac{n_j}{N_j}$  and  $Y'_i = \left(\frac{N_j}{n_t} Y_i(1) + \frac{N_j}{n_c} Y_i(0)\right)$ .

$$\begin{aligned}
\text{Var}_{\mathbf{Z}} [\hat{\tau} | \boldsymbol{\pi}] &= \text{Var}_{\mathbf{Z}} \left[ \frac{1}{n_t} \sum_{j=1}^M \sum_{i \in \mathcal{C}_j} Z_i Y_i(1) - \frac{1}{n_c} \sum_{j=1}^M \sum_{i \in \mathcal{C}_j} (1 - Z_i) Y_i(0) \middle| \boldsymbol{\pi} \right] \\
&= \sum_{j=1}^M \text{Var}_{\mathbf{Z}} \left[ \frac{1}{N_j} \sum_{i \in \mathcal{C}_j} \frac{N_j}{n_t} Z_i Y_i(1) - \frac{N_j}{n_c} (1 - Z_i) Y_i(0) \middle| \boldsymbol{\pi} \right] \\
&= \sum_{j=1}^M \text{Var}_{\mathbf{Z}} \left[ \frac{1}{N_j} \sum_{i \in \mathcal{C}_j} D_i \left( \frac{N_j}{n_t} Y_i(1) + \frac{N_j}{n_c} Y_i(0) \right) \middle| \boldsymbol{\pi} \right] + 0 \\
&= \sum_{j=1}^M \text{Var}_{\mathbf{Z}} \left[ \frac{1}{N_j} \sum_{i \in \mathcal{C}_j} D_i Y'_i \middle| \boldsymbol{\pi} \right]
\end{aligned}$$

Note that:

$$\begin{aligned}
\mathbb{E}_{\mathbf{Z}}[D_i] &= 0 \\
\mathbb{E}_{\mathbf{Z}}[D_i^2] &= \frac{n_j(N_j - n_j)}{N_j^2} \\
\mathbb{E}_{\mathbf{Z}}[D_i D_j] &= -\frac{n_j(N_j - n_j)}{N_j^2(N_j - 1)} \quad \text{for } i \neq j
\end{aligned}$$

We introduce the following quantities:

$$\begin{aligned}
S_{tj} &= \frac{1}{N_j - 1} \sum_{i \in \mathcal{C}_j} \left( Y_i(1) - \bar{Y}_i(1) \right)^2 \\
S_{cj} &= \frac{1}{N_j - 1} \sum_{i \in \mathcal{C}_j} \left( Y_i(0) - \bar{Y}_i(0) \right)^2 \\
S_{tcj} &= \frac{2}{N_j - 1} \sum_{i \in \mathcal{C}_j} (Y_i(1) - Y_i(0)) (\bar{Y}(1) - \bar{Y}(0)).
\end{aligned}$$

Thus, we have

$$\begin{aligned}
\text{Var}_{\mathbf{Z}} [\hat{\tau} | \boldsymbol{\pi}] &= \sum_{j=1}^M \frac{1}{N_j^2} \mathbb{E}_{\mathbf{Z}} \left[ \left( \sum_{i=1}^N D_i Y'_i \right)^2 \right] \\
&= \sum_{j=1}^M \frac{n_j(N_j - n_j)}{N_j^3(N_j - 1)^2} \sum_{i \in \mathcal{C}_j} (Y'_i - \bar{Y}')^2
\end{aligned}$$

$$\begin{aligned}
&= \sum_{j=1}^M \frac{n_j(N_j - n_j)}{N_j^3(N_j - 1)} \sum_{i \in \mathcal{C}_j} \left( \frac{N_j}{n_t} Y_i(1) + \frac{N_j}{n_c} Y_i(0) - \left( \frac{N_j}{n_t} \bar{Y}(1) + \frac{N_j}{n_c} \bar{Y}(0) \right) \right)^2 \\
&= \sum_{j=1}^M \frac{n_j(N_j - n_j)}{N_j^3(N_j - 1)} \frac{N_j^2}{n_t^2} \sum_{i \in \mathcal{C}_j} (Y_i(1) - \bar{Y}(1))^2 + \frac{n_j(N_j - n_j)}{N_j^3(N_j - 1)} \frac{N_j^2}{n_c^2} \sum_{i \in \mathcal{C}_j} (Y_i(0) - \bar{Y}(0))^2 \\
&\quad + 2 \frac{n_j(N_j - n_j)}{N_j^3(N_j - 1)} \frac{N_j^2}{n_t n_c} \sum_{i \in \mathcal{C}_j} (Y_i(1) - Y_i(0)) (\bar{Y}(1) - \bar{Y}(0)) \\
&= \sum_{j=1}^M \frac{n_j(N_j - n_j)}{N_j n_t^2} \frac{1}{N_j - 1} \sum_{i=1}^M (Y_i(1) - \bar{Y}_i(1))^2 + \frac{n_j(N_j - n_j)}{N_j n_c^2} \frac{1}{N_j - 1} \sum_{i=1}^M (Y_i(0) - \bar{Y}(0))^2 \\
&\quad + 2 \frac{n_j(N_j - n_j)}{N_j n_t n_c} \frac{1}{N_j - 1} \sum_{i \in \mathcal{C}_j} (Y_i(1) - Y_i(0)) (\bar{Y}(1) - \bar{Y}(0)) \\
&= \sum_{j=1}^M \frac{n_j(N_j - n_j)}{N_j n_t^2} S_{tj} + \frac{n_j(N_j - n_j)}{N_j n_c^2} S_{cj} + \frac{n_j(N_j - n_j)}{N_j n_t n_c} S_{tcj}
\end{aligned}$$

Since  $S_{tcj} = S_{tj} + S_{cj} - S_{tcj}$ ,

$$\begin{aligned}
\text{Var}_{\mathbf{Z}} [\hat{\tau} | \boldsymbol{\pi}] &= \frac{N}{n_t n_c} \sum_{j=1}^M \frac{n_j(N_j - n_j)}{N_j n_t} S_{tj} + \frac{n_j(N_j - n_j)}{N_j n_c} S_{cj} - \sum_{j=1}^M \frac{n_j(N_j - n_j)}{N_j n_t n_c} S_{tcj} \\
&= \sum_{j=1}^M \frac{N}{n_t^2 n_c} N_j \pi_j (1 - \pi_j) S_{tj} + \frac{N}{n_t n_c^2} N_j \pi_j (1 - \pi_j) S_{cj} - \frac{1}{n_t n_c} N_j \pi_j (1 - \pi_j) S_{tcj}
\end{aligned}$$

If we take the expectation with respect to  $\boldsymbol{\pi}$ ,

$$\mathbb{E}_{\boldsymbol{\pi}} [\text{Var}_{\mathbf{Z}} [\hat{\tau} | \boldsymbol{\pi}]] = \left( \frac{n_t}{N} - \text{Var}[\boldsymbol{\pi}] - \frac{n_t^2}{N^2} \right) \left( \sum_{j=1}^M \frac{N}{n_t^2 n_c} N_j S_{tj} + \frac{N}{n_t n_c^2} N_j S_{cj} - \frac{1}{n_t n_c} N_j S_{tcj} \right)$$

An alternative formulation is given by:

$$\mathbb{E}_{\boldsymbol{\pi}} \text{Var}_{\mathbf{Z}} [\hat{\tau} | \boldsymbol{\pi}] = \left( \frac{n_t n_c}{N^2} - \text{Var}[\boldsymbol{\pi}] \right) \left( \sum_{j=1}^M \frac{N_j}{n_t^2} S_{tj} + \frac{N_j}{n_c^2} S_{cj} + \frac{N_j}{n_t n_c} S_{tcj} \right)$$

This concludes the first part of the Eve equation. We now compute the second part of the Eve equation. Let  $Y_j^+(1) = \sum_{i \in \mathcal{C}_j} Y_i(1)$  and  $Y_j^+(0) = \sum_{i \in \mathcal{C}_j} Y_i(0)$ . Finally, we let  $\tilde{Y}_j = \frac{Y_j^+(1)}{n_t} + \frac{Y_j^+(0)}{n_c}$ . Furthermore, let  $C_j = \pi_j - \frac{n_j}{N}$ .

$$\begin{aligned}
\text{Var}_{\boldsymbol{\pi}} [\mathbb{E}_{\mathbf{Z}} [\hat{\tau} | \boldsymbol{\pi}]] &= \text{Var}_{\boldsymbol{\pi}} \left[ \frac{1}{n_t} \sum_{j=1}^M \frac{n_j}{N_j} Y_j^+(1) - \frac{1}{n_c} \sum_{j=1}^M \left( 1 - \frac{n_j}{N_j} \right) Y_j^+(0) \right] \\
&= \text{Var}_{\boldsymbol{\pi}} \left[ \sum_{j=1}^M \pi_j \frac{Y_j^+(1)}{n_t} - (1 - \pi_j) \frac{Y_j^+(0)}{n_c} \right]
\end{aligned}$$

$$\begin{aligned}
&= \text{Var}_{\boldsymbol{\pi}} \left[ \sum_{j=1}^M \pi_j \tilde{Y}_j \right] + 0 \\
&= \text{Var}_{\boldsymbol{\pi}} \left[ \sum_{j=1}^M C_j \tilde{Y}_j \right] + 0 \\
&= \mathbb{E}_{\boldsymbol{\pi}} \left[ \left( \sum_{j=1}^M C_j \tilde{Y}_j \right)^2 \right]
\end{aligned}$$

Note that,

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\pi}} [C_j^2] &= \text{Var}[\boldsymbol{\pi}] \\
\mathbb{E}_{\boldsymbol{\pi}} [C_p C_q] &= \mathbb{E}[\pi_p \pi_q] - \frac{n_t^2}{N^2} \\
&= \frac{1}{M(M-1)} \sum_{p \neq q} \pi_p \pi_q + \frac{1}{M(M-1)} \sum_p \pi_p^2 - \frac{1}{M(M-1)} \sum_p \pi_p^2 - \frac{n_t^2}{N^2} \\
&= \frac{1}{M(M-1)} \left( \frac{Mn_t}{N} \right)^2 - \frac{1}{M-1} \left( \text{Var}[\boldsymbol{\pi}] + \frac{n_t^2}{N^2} \right) - \frac{n_t^2}{N^2} \\
&= -\frac{\text{Var}[\boldsymbol{\pi}]}{M-1}
\end{aligned}$$

Finally, we introduce the following quantities:

$$\begin{aligned}
\text{Var}(Y_t^+) &= \frac{1}{M} \sum_{j=1}^M \left( Y_j^+(1) - \bar{Y}_j^+(1) \right)^2 \\
\text{Var}(Y_c^+) &= \frac{1}{M} \sum_{j=1}^M \left( Y_j^+(0) - \bar{Y}_j^+(0) \right)^2 \\
\text{Var}(Y_{tc}^+) &= \frac{1}{M} \sum_{j=1}^M \left( Y_j^+(1) - \bar{Y}_j^+(1) \right) \left( Y_j^+(0) - \bar{Y}_j^+(0) \right).
\end{aligned}$$

It follows that:

$$\begin{aligned}
\text{Var}_{\boldsymbol{\pi}} [\mathbb{E}_{\mathbf{Z}} [\hat{\boldsymbol{\tau}} | \boldsymbol{\pi}]] &= \text{Var}[\boldsymbol{\pi}] \left( \sum_{j=1}^M \tilde{Y}_j^2 - \frac{1}{M-1} \sum_j \sum_{k \neq j} \tilde{Y}_j \tilde{Y}_k \right) \\
&= \text{Var}[\boldsymbol{\pi}] \left( \frac{M}{M-1} \sum_{j=1}^M \tilde{Y}_j^2 - \frac{1}{M-1} \sum_{k,j} \tilde{Y}_j \tilde{Y}_k \right) \\
&= \text{Var}[\boldsymbol{\pi}] \frac{M^2}{M-1} \left( \frac{1}{M} \sum_{j=1}^M \tilde{Y}_j^2 - \left( \frac{1}{M} \sum_j \tilde{Y}_j \right)^2 \right)
\end{aligned}$$

$$\begin{aligned}
&= \text{Var}[\boldsymbol{\pi}] \frac{M}{M-1} \sum_{j=1}^M (\check{Y}_j - \bar{Y})^2 \\
&= \text{Var}[\boldsymbol{\pi}] \frac{M^2}{M-1} \left( \frac{\text{Var}(Y_t^+)}{n_t^2} + \frac{\text{Var}(Y_c^+)}{n_c^2} + \frac{\text{Var}(Y_{tc}^+)}{n_c n_t} \right)
\end{aligned}$$

In conclusion,

$$\begin{aligned}
\text{Var}_{\mathbf{z}, \boldsymbol{\pi}} [\hat{t}] &= \frac{\text{Var}[\boldsymbol{\pi}] M^2}{M-1} \left( \frac{\text{Var}(Y_t^+)}{n_t^2} + \frac{\text{Var}(Y_c^+)}{n_c^2} + \frac{\text{Var}(Y_{tc}^+)}{n_c n_t} \right) \\
&\quad + \left( \frac{n_t n_c}{N^2} - \text{Var}[\boldsymbol{\pi}] \right) \left( \sum_{j=1}^M N_j \frac{S_{tj}^2}{n_t^2} + N_j \frac{S_{cj}}{n_c^2} + N_j \frac{S_{tcj}}{n_t n_c} \right) \\
&= \frac{n_t n_c}{N^2} \sum_{j=1}^M N_j \left( \frac{S_{tj}}{n_t^2} + \frac{S_{cj}}{n_c^2} + \frac{S_{tcj}}{n_t n_c} \right) \\
&\quad + \text{Var}[\boldsymbol{\pi}] \left( \frac{M^2}{M-1} \left( \frac{\text{Var}(Y_t^+)}{n_t^2} + \frac{\text{Var}(Y_c^+)}{n_c^2} + \frac{\text{Var}(Y_{tc}^+)}{n_c n_t} \right) - \sum_{j=1}^M N_j \left( \frac{S_{tj}}{n_t^2} + \frac{S_{cj}}{n_c^2} + \frac{S_{tcj}}{n_t n_c} \right) \right)
\end{aligned}$$

The coefficient in front of  $\text{Var}[\boldsymbol{\pi}]$  is not always positive. Consider a case where  $\forall j, k, Y_j^+ = Y_k^+$ , such that  $\text{Var}(Y_t^+) = 0$ ,  $\text{Var}(Y_c^+) = 0$ ,  $\text{Var}(Y_{tc}^+) = 0$ . It is still possible however for  $Y_i(1)$ ,  $Y_i(0)$  to have some variance within each cluster, such that  $S_{tj} > 0$  and  $S_{cj} > 0$ :

$$\frac{M^2}{M-1} \left( \frac{\text{Var}(Y_t^+)}{n_t^2} + \frac{\text{Var}(Y_c^+)}{n_c^2} + \frac{\text{Var}(Y_{tc}^+)}{n_c n_t} \right) - \sum_{j=1}^M N_j \left( \frac{S_{tj}}{n_t^2} + \frac{S_{cj}}{n_c^2} + \frac{S_{tcj}}{n_t n_c} \right) \leq - \sum_{j=1}^M N_j \left( \frac{S_{tj}}{n_t^2} + \frac{S_{cj}}{n_c^2} \right) < 0$$

## C.4 Proof of Corollary 1

The simplified equations are obtained by noticing that  $\text{Var}[\boldsymbol{\pi}] = 0$  if the vector  $\boldsymbol{\pi}$  is constant, and  $\text{Var}[\boldsymbol{\pi}] = \frac{n_t n_c}{N^2}$  in the case of a cluster-based randomized assignment. We also use the fact that:  $S_{tcj} = S_{jt} + S_{cj} - S_{tcj}$  and idem for  $\text{Var}(Y_{tc}^+)$ .

## C.5 Proof of Proposition 13

The stratified estimator is given by:

$$\begin{aligned}\hat{\tau}' &= \sum_{j=1}^M \lambda_j \hat{\tau}(j) \\ &= \sum_{j=1}^M \lambda_j \sum_{i=1}^{N_j} (Z_i Y_i(1) + (1 - Z_i) Y_i(0)) \frac{(-1)^{1-Z_i}}{n_j^{Z_i} (N_j - n_j)^{1-Z_i}}\end{aligned}$$

The expectation of the stratified difference-in-means estimator conditioned on the proportion of units assigned to treatment is given by:

$$\mathbb{E}_{\mathbf{Z}} [\hat{\tau}^s | \boldsymbol{\pi}] = \sum_{j=1}^M \frac{\lambda_j}{N_j} \sum_{i \in \mathcal{C}_j} Y_i(1) - Y_i(0)$$

If  $\lambda_j = \frac{N_j}{N}$ , then the stratified estimator is unbiased for the total treatment effect conditioned on the assignment of treatment proportions to clusters. Same in expectation over that assignment.

## C.6 Proof of Proposition 14

According to Eve's law, we must compute two terms. The first term is equal to 0.

$$\text{Var}_{\boldsymbol{\pi}} [\mathbb{E}_{\mathbf{Z}} [\hat{\tau} | \boldsymbol{\pi}]] = \text{Var}_{\boldsymbol{\pi}} \left[ \sum_{j=1}^M \frac{\lambda_j}{N_j} \sum_{i \in \mathcal{C}_j} Y_i(1) - Y_i(0) \right] = 0$$

such that  $\text{Var}_{\mathbf{Z}} [\hat{\tau}^s] = \mathbb{E}_{\boldsymbol{\pi}} [\text{Var}_{\mathbf{Z}} [\hat{\tau}^s | \boldsymbol{\pi}]]$ . We compute this remaining term:

$$\begin{aligned}\text{Var}_{\mathbf{Z}} [\hat{\tau}^s | \boldsymbol{\pi}] &= \sum_{j=1}^M \lambda_j^2 \text{Var}_{\mathbf{Z}} [\hat{\tau}(j) | \boldsymbol{\pi}_j] \\ &= \sum_{j=1}^M \lambda_j^2 \left( \frac{S_{tj}}{n_j} + \frac{S_{cj}}{N_j - n_j} - \frac{S_{tcj}}{N_j} \right) \\ \mathbb{E}_{\boldsymbol{\pi}} [\text{Var}_{\mathbf{Z}} [\hat{\tau}^s | \boldsymbol{\pi}]] &= \sum_{j=1}^M \lambda_j^2 \left( S_{tj} \mathbb{E}_{\boldsymbol{\pi}} \left[ \frac{1}{n_j} \right] + S_{cj} \mathbb{E}_{\boldsymbol{\pi}} \left[ \frac{1}{N_j - n_j} \right] - \frac{S_{tcj}}{N_j} \right)\end{aligned}$$

$$= \sum_{j=1}^M \frac{\lambda_j^2}{N_j} \left( S_{tj} \mathbb{E} \pi \left[ \frac{1}{\pi_j} \right] + S_{cj} \mathbb{E} \pi \left[ \frac{1}{1 - \pi_j} \right] - S_{tcj} \right)$$

Let  $\pi^\dagger = \left( \frac{1}{M} \sum_{j=1}^M \frac{1}{\pi_j} \right)^{-1}$  be the harmonic mean of  $\pi$ . If we use  $\lambda_j = \frac{N_j}{N}$ , then the above formula becomes:

$$\text{Var}_{\mathbf{Z}} [\hat{\tau}^s] = \sum_{j=1}^M \frac{N_j}{N} \left( \frac{S_{tj}}{N \pi^\dagger} + \frac{S_{cj}}{N(1 - \pi)^\dagger} - \frac{S_{tcj}}{N} \right)$$

Since any mean-preserving spread (Mitchell, 2004) of  $\pi$  will decrease the harmonic mean, the optimal randomized saturation design is one with the lowest variance for  $\pi$ , i.e.  $\pi = \left( \frac{n_t}{N} \right)_M$ .

## C.7 Proof of Theorem 8

Recall that  $\mathcal{C}(i) \in [1, M]$  is the cluster that unit  $i$  belongs to. We assume block-fixed interference effects and compute the expectation of the difference-in-means estimator, conditioned on the assignment of clusters to treatment-proportions  $\pi$ .

$$\begin{aligned} \mathbb{E}_{\mathbf{Z}} [\hat{\tau} | \pi] &= \mathbb{E}_{\mathbf{Z}} \left[ \sum_{i=1}^n \left( \alpha_i + \beta_i Z_i + \frac{\gamma_i}{|\mathcal{N}_i|} \sum_{k \in \mathcal{N}_i} Z_k \right) \frac{(-1)^{1-Z_i}}{n_t^{Z_i} n_c^{1-Z_i}} \right] \\ &= \mathbb{E}_{\mathbf{Z}} \left[ \sum_{j=1}^M \sum_{i \in \mathcal{C}_j} \left( \alpha_i + \beta_i Z_i + \frac{\gamma_i}{|\mathcal{N}_i|} \sum_{k \in \mathcal{N}_i} Z_k \right) \frac{(-1)^{1-Z_i}}{n_t^{Z_i} n_c^{1-Z_i}} \right] \\ &= \sum_{j=1}^M \sum_{i \in \mathcal{C}_j} \alpha_i \left( \frac{\pi_j}{n_t} - \frac{1 - \pi_j}{n_c} \right) + \sum_{j=1}^M \sum_{i \in \mathcal{C}_j} \beta_i \frac{\pi_j}{n_t} \\ &\quad + \frac{\gamma_i}{|\mathcal{N}_i|} \sum_{k \in \mathcal{N}_i \cap \mathcal{C}_j} \frac{n_j}{N_j} \frac{n_j - 1}{N_j - 1} \frac{1}{n_t} + \frac{n_j}{N_j} \frac{N_j - n_j - 1}{N_j - 1} \frac{1}{n_c} \\ &\quad + \frac{\gamma_i}{|\mathcal{N}_i|} \sum_{k \in \mathcal{N}_i \setminus \mathcal{C}_j} \frac{n_j}{N_j} \left( 1 - \frac{n_{\mathcal{C}(k)}}{N_{\mathcal{C}(k)}} \right) \frac{-1}{n_c} + \frac{n_j}{N_j} \frac{n_{\mathcal{C}(k)}}{N_{\mathcal{C}(k)}} \frac{1}{n_t} \\ &= \alpha^* + \tilde{\beta} + \sum_{j=1}^M \frac{n_j}{N_j} \left( \frac{n_j - 1}{N_j - 1} \frac{1}{n_t} - \frac{N_j - n_j}{N_j - 1} \frac{1}{n_c} \right) \gamma^{(j)} \sum_{i \in \mathcal{C}_j} \frac{|\mathcal{N}_i \cap \mathcal{C}_j|}{|\mathcal{N}_i|} \\ &\quad + \sum_{j=1}^M \sum_{l \neq j} \frac{n_j}{N_j} \left( \frac{n_l}{N_l} \frac{1}{n_t} - \left( 1 - \frac{n_l}{N_l} \right) \frac{1}{n_c} \right) \gamma^{(j)} \sum_{i \in \mathcal{C}_j} \frac{|\mathcal{N}_i \cap \mathcal{C}_l|}{|\mathcal{N}_i|} \end{aligned}$$

$$\begin{aligned}
&= \alpha^* + \tilde{\beta} + \sum_{j,l} \frac{n_j}{N_j} \left( \frac{n_l}{N_l} \frac{1}{n_t} - \left(1 - \frac{n_l}{N_l}\right) \frac{1}{n_c} \right) \gamma^{(j)} \sum_{i \in \mathcal{C}_j} \frac{|\mathcal{N}_i \cap \mathcal{C}_l|}{|\mathcal{N}_i|} + o(\text{XXX}) \\
&= \alpha^* + \tilde{\beta} + \sum_{j,l} \pi_j \left( \frac{\pi_l}{n_t} - \frac{1 - \pi_l}{n_c} \right) \gamma^{(j)} \sum_{i \in \mathcal{C}_j} \frac{|\mathcal{N}_i \cap \mathcal{C}_l|}{|\mathcal{N}_i|} + o(\text{XXX})
\end{aligned}$$

We want to take the expectation of this with respect to a permutation of  $\pi$ , where the last equality is true if the clusters are of equal size.

$$\begin{aligned}
\mathbb{E}_\pi[\tilde{\beta}] &= \sum_{j=1}^M \sum_{i \in \mathcal{C}_j} \beta_i \sum_k \frac{\pi_k}{n_t} \mathbb{E}_\pi[P_{jk}] \\
&= \sum_{j=1}^M \sum_{i \in \mathcal{C}_j} \beta_i \frac{1}{n_t} \left( \frac{1}{M} \sum_k \pi_k \right) \\
&= \tilde{\beta}
\end{aligned}$$

Similarly with  $\alpha^*$ , where the last inequality holds if the clusters are of equal size.

$$\begin{aligned}
\mathbb{E}_\pi[\tilde{\alpha}] &= \sum_{j=1}^M \sum_{i \in \mathcal{C}_j} \alpha_i \sum_k \left( \frac{\pi_k}{n_t} - \frac{1 - \pi_k}{n_t} \right) \mathbb{E}_\pi[P_{jk}] \\
&= \sum_{j=1}^M \sum_{i \in \mathcal{C}_j} \alpha_i \left( \frac{1}{M} \sum_k \left( \frac{\pi_k}{n_t} - \frac{1 - \pi_k}{n_c} \right) \right) \\
&= 0
\end{aligned}$$

As for the  $\gamma$  term, if the clusters are of equal size, and for a permutation  $P$ , this simplifies to:

$$\begin{aligned}
\mathbb{E}_{\mathbf{Z}, \pi}[\tilde{\gamma}] &= \mathbb{E}_{\mathbf{Z}, \pi} \left[ \sum_{j,l} \pi_j \left( \frac{\pi_l}{n_t} - \frac{1 - \pi_l}{n_c} \right) \gamma^{(j)} \sum_{i \in \mathcal{C}_j} \frac{|\mathcal{N}_i \cap \mathcal{C}_l|}{|\mathcal{N}_i|} \right] \\
&= \mathbb{E}_{\mathbf{Z}, \pi} \left[ \frac{N}{n_t n_c} \sum_{j,l} \gamma^{(j)} \pi_j \pi_l \sum_{i \in \mathcal{C}_j} \frac{|\mathcal{N}_i \cap \mathcal{C}_l|}{|\mathcal{N}_i|} - \frac{1}{n_c} \sum_j \gamma^{(j)} \pi_j \right] \\
&= \frac{N}{n_t n_c} \sum_{j,l,p,q} \gamma^{(j)} \sum_{i \in \mathcal{C}_j} \frac{|\mathcal{N}_i \cap \mathcal{C}_l|}{|\mathcal{N}_i|} \pi_j \pi_q \mathbb{E}_{\mathbf{Z}, \pi} [P_{jp} P_{lq}] - \frac{n_t}{n_c} \tilde{\gamma}
\end{aligned}$$

We introduce the quantity

$$\gamma' := \frac{1}{M} \sum_j \frac{\gamma^{(j)}}{N_j} \sum_{i \in \mathcal{C}_j} \frac{|\mathcal{N}_i \cap \mathcal{C}_j|}{|\mathcal{N}_i|}$$

Continuing to explore the first term,

$$\begin{aligned}\mathbb{E}_{\mathbf{Z}, \boldsymbol{\pi}}[\tilde{\gamma}] &= \frac{N}{n_t n_c} \left( \sum_{\substack{j \neq l \\ p \neq q}} \gamma^{(j)} \sum_{i \in \mathcal{C}_j} \frac{|\mathcal{N}_i \cap \mathcal{C}_l|}{|\mathcal{N}_i|} \pi_p \pi_q \frac{1}{M(M-1)} + \sum_{j,p} \gamma^{(j)} \sum_{i \in \mathcal{C}_j} \frac{|\mathcal{N}_i \cap \mathcal{C}_j|}{|\mathcal{N}_i|} \pi_p^2 \frac{1}{M} \right) \\ &= \frac{N}{n_t n_c} \left[ \frac{1}{M(M-1)} \left( \sum_{p \neq q} \pi_p \pi_q \right) \left( \sum_{j \neq l} \gamma^{(j)} \sum_{i \in \mathcal{C}_j} \frac{|\mathcal{N}_i \cap \mathcal{C}_l|}{|\mathcal{N}_i|} \right) + \left( \sum_p \pi_p^2 \right) \left( \frac{1}{M} \sum_j \gamma^{(j)} \sum_{i \in \mathcal{C}_j} \frac{|\mathcal{N}_i \cap \mathcal{C}_j|}{|\mathcal{N}_i|} \right) \right]\end{aligned}$$

Assuming that the clusters are of equal size such that  $\bar{\pi} = \frac{n_t}{N}$ , and noting that

$$\begin{aligned}\sum_{p \neq q} \pi_p \pi_q &= \frac{n_t^2 M^2}{N^2} - \sum_p \pi_p^2 \\ \frac{1}{M} \sum_j \gamma^{(j)} \sum_{i \in \mathcal{C}_j} \frac{|\mathcal{N}_i \cap \mathcal{C}_j|}{|\mathcal{N}_i|} &= \frac{N}{M} \gamma' \\ \sum_{j \neq l} \gamma^{(j)} \sum_{i \in \mathcal{C}_j} \frac{|\mathcal{N}_i \cap \mathcal{C}_l|}{|\mathcal{N}_i|} &= N(\bar{\gamma} - \gamma') \\ \sum_p \pi_p^2 &= M \left( \text{Var}[\boldsymbol{\pi}] + \frac{n_t^2}{N^2} \right)\end{aligned}$$

It follows

$$\begin{aligned}\mathbb{E}_{\mathbf{Z}, \boldsymbol{\pi}}[\tilde{\gamma}] &= \frac{N}{n_t n_c} \left[ \frac{1}{M(M-1)} \left( \frac{M^2 n_t^2}{N^2} - M \text{Var}[\boldsymbol{\pi}] - \frac{M n_t^2}{N^2} \right) N(\bar{\gamma} - \gamma') + M \left( \text{Var}[\boldsymbol{\pi}] + \frac{n_t^2}{N^2} \right) \frac{N}{M} \gamma' \right] \\ &= \frac{N}{n_t n_c} \left( N \gamma' - \frac{N}{M-1} (\bar{\gamma} - \gamma') \right) \text{Var}[\boldsymbol{\pi}] + \frac{N}{n_t n_c} \frac{n_t^2}{N} (\gamma' + (\bar{\gamma} - \gamma')) \\ &= \frac{N^2}{n_t n_c} \left( \gamma' - \frac{\bar{\gamma} - \gamma'}{M-1} \right) \text{Var}[\boldsymbol{\pi}] + \frac{N}{n_t n_c} \frac{n_t^2}{N} \bar{\gamma} \\ &= \frac{N^2}{n_t n_c} \left( \gamma' - \frac{\bar{\gamma} - \gamma'}{M-1} \right) \text{Var}[\boldsymbol{\pi}] + \frac{n_t}{n_c} \bar{\gamma}\end{aligned}$$

In conclusion,

$$\mathbb{E}_{\mathbf{Z}, \boldsymbol{\pi}}[\hat{\tau}] = \bar{\beta} + \frac{N^2}{n_t n_c} \left( \gamma' - \frac{\bar{\gamma} - \gamma'}{M-1} \right) \text{Var}[\boldsymbol{\pi}]$$

## C.8 Proof of Proposition 15

Recall that  $\hat{\tau}^s$  is the stratified estimator. We seek to understand its expectation under the linear interference model in Equation 3.13.

$$\begin{aligned}
\mathbb{E}_{\mathbf{Z}}[\hat{\tau}^s | \boldsymbol{\pi}] &= \mathbb{E}_{\mathbf{Z}} \left[ \sum_{j=1}^M \lambda_j \sum_{i \in \mathcal{C}_j} \left( \alpha_i + \beta_i Z_i + \frac{\gamma_i}{|\mathcal{N}_i|} \sum_{k \in \mathcal{N}_i} Z_k \right) \frac{(-1)^{1-Z_i}}{n_j^{Z_i} (N_j - n_j)^{1-Z_i}} \right] \\
&= \sum_{j=1}^M \lambda_j \sum_{i \in \mathcal{C}_j} \beta_i \frac{n_j}{N_j} \frac{1}{n_j} + \frac{\gamma_i}{|\mathcal{N}_i|} \sum_{l \neq j} \sum_{k \in \mathcal{C}_l \cap \mathcal{N}_i} \frac{n_l}{N_l} \left( \frac{n_j}{N_j} \frac{1}{n_j} - \frac{N_j - n_j}{N_j} \frac{1}{N_j - n_j} \right) \\
&\quad + \frac{\gamma_i}{|\mathcal{N}_i|} \sum_{k \in \mathcal{C}_j \cap \mathcal{N}_i} \frac{n_j}{N_j} \frac{1}{n_j} \\
&= \sum_{j=1}^M \lambda_j \bar{\beta}_j + \frac{\lambda_j}{N_j} \sum_{i \in \mathcal{C}_j} \gamma_i \frac{|\mathcal{N}_i \cap \mathcal{C}_j|}{|\mathcal{N}_i|}
\end{aligned}$$

If  $\lambda_j = \frac{N_j}{N}$ , the previous formula simplifies to:

$$\mathbb{E}_{\mathbf{Z}}[\hat{\tau}^s | \boldsymbol{\pi}] = \bar{\beta} + \frac{1}{N} \sum_{j=1}^M \sum_{i \in \mathcal{C}_j} \gamma_i \frac{|\mathcal{N}_i \cap \mathcal{C}_j|}{|\mathcal{N}_i|}$$

If the interference effects are constant, then the formula becomes:

$$\mathbb{E}_{\mathbf{Z}}[\hat{\tau}^s | \boldsymbol{\pi}] = \bar{\beta} + \gamma \rho_c$$

## C.9 Proof of Corollary 2

Assume block-fixed interference effects and that the graph is perfectly clustered. It follows then that  $\gamma' = \bar{\gamma}$ . The expectation of the difference-in-means estimator is

$$\mathbb{E}_{\mathbf{Z}}[\hat{\tau}] = \bar{\beta} + \frac{N^2}{n_t n_c} \bar{\gamma} \text{Var}[\boldsymbol{\pi}] + o(X)$$

Furthermore, the total treatment effect is  $\bar{\beta} + \bar{\gamma}$ , thus concluding the first part of our proof. For constant direct and interference effects,  $\bar{\beta} = \tilde{\beta} = \beta$  and  $\bar{\gamma} = \gamma$ . The total treatment effect is given by  $\beta + \gamma$ . If the graph is perfectly clustered into clusters of equal size, the conditional expectation of the difference-in-means estimator  $\hat{\tau}$  is

$$\begin{aligned}
\mathbb{E}_{\mathbf{Z}}[\hat{\tau} | \boldsymbol{\pi}] &= \beta + \gamma \left( \sum_{j=1}^M N_j \pi_j^2 \right) \left( \frac{1}{n_t} + \frac{1}{n_c} \right) - \gamma \frac{n_t}{n_c} \\
&= \beta + \gamma \left( \frac{N^2}{n_t n_c M} \sum_j \pi_j^2 - \frac{n_t}{n_c} \right)
\end{aligned}$$

$$\begin{aligned}
&= \beta + \gamma \frac{N^2}{n_t n_c} \left( \frac{1}{M} \sum_j \pi_j^2 - \frac{n_t^2}{N^2} \right) \\
&= \beta + \frac{\gamma N^2}{n_t n_c} \text{Var}[\boldsymbol{\pi}]
\end{aligned}$$

Thus, we have showed that when the graph is perfectly clustered into clusters of equal size, the bias is linear in the variance of the treatment proportions vector. The other direction follows from Lemma 2.

### C.10 Proof for Corollary 3

Assume that the graph is randomly clustered. Asymptotically,  $\frac{|\mathcal{N}_i \cap \mathcal{C}_i|}{|\mathcal{N}_i|} \approx \frac{1}{M}$ , such that  $\gamma' \approx \frac{\bar{\gamma}}{M}$ .

From the expression of  $\mathbb{E}_{\mathbf{Z}}[\hat{\tau}]$  in Theorem 8, we have:

$$\mathbb{E}_{\mathbf{Z}}[\hat{\tau}] \approx \bar{\beta} + \frac{N^2}{n_t n_c} \text{Var}[\boldsymbol{\pi}] \left( \frac{\bar{\gamma}}{M} - \frac{\bar{\gamma} - \frac{\bar{\gamma}}{M}}{M-1} \right) \approx \bar{\beta}$$

The rest follows immediately.

### C.11 Proof of Theorem 9

The bias of the difference-in-means estimator for a randomized saturation design under the linear model of interference in Equation 3.13 is

$$\left| \frac{N^2}{n_t n_c} \left( \gamma' - \frac{\bar{\gamma} - \gamma'}{M-1} \right) \text{Var}[\boldsymbol{\pi}] - \bar{\gamma} \right|$$

Since  $\text{Var}[\boldsymbol{\pi}] \in [0, \frac{n_t n_c}{N^2}]$ , we define  $\lambda := \frac{N^2}{n_t n_c} \text{Var}[\boldsymbol{\pi}]$  such that  $\lambda \in [0, 1]$ . Furthermore,  $\gamma' \in [0, \bar{\gamma}]$ , such that we define  $\mu = \frac{\gamma'}{\bar{\gamma}}$ , with  $\mu \in [0, 1]$ . The bias becomes

$$\begin{aligned}
|TTE - \mathbb{E}_{\mathbf{Z}}[\hat{\tau}]| &= \bar{\gamma} \left| \left( \mu - \frac{1-\mu}{M-1} \right) \lambda - 1 \right| \\
&= \bar{\gamma} \left| \frac{M\mu - 1}{M-1} \lambda - 1 \right|
\end{aligned}$$

Since  $\lambda(M\mu - 1) \leq M - 1$ , the bias becomes

$$|TTE - \mathbb{E}_{\mathbf{Z}}[\hat{\tau}]| = \bar{\gamma} \left( 1 - \frac{M\mu - 1}{M - 1} \lambda \right)$$

We distinguish two cases. If  $\mu \geq \frac{1}{M}$ , then  $f$  reaches a minimum at  $\lambda = 1$ , equal to  $\frac{M}{M-1}(\bar{\gamma} - \gamma')$ . If  $\mu \leq \frac{1}{M}$ , then  $f$  reaches a minimum at  $\lambda = 0$ , equal to  $\bar{\gamma}$ .

## C.12 Proof of Example 1

Let  $Y_j^+ := \sum_{i \in \mathcal{C}_j} Y_i$  be the cluster-level outcomes. Recall the definition of  $f$  and the difference-in-means estimator  $\hat{\tau}$  under the stable unit treatment value assumption:

$$\begin{aligned} f(\boldsymbol{\pi}, \mathcal{C}, \Theta) &= |TTE - \mathbb{E}_{\mathbf{Z}}[\hat{\tau}|\boldsymbol{\pi}]| \\ &= \left| \frac{1}{N} \sum_{i=1}^N (Y_i(1) - Y_i(0)) - \left( \frac{1}{n_t} \sum_{j=1}^M \pi_j \sum_{i \in \mathcal{C}_j} Y_i(1) - \frac{1}{n_c} \sum_{j=1}^M (1 - \pi_j) \sum_{i \in \mathcal{C}_j} Y_i(0) \right) \right| \\ &= \left| \sum_{j=1}^M \left( \frac{\pi_j}{n_t} - \frac{1}{N} \right) Y_j^+(1) - \left( \frac{1 - \pi_j}{n_c} - \frac{1}{N} \right) Y_j^+(0) \right| \end{aligned}$$

It is easy to see that for  $\boldsymbol{\pi}^* = \left(\frac{n_t}{N}\right)_{M'}$ , we have  $f(\boldsymbol{\pi}^*, \mathcal{C}, \Theta) = 0$ , such that

$$\left(\frac{n_t}{N}\right)_M \in \arg \min_{\boldsymbol{\pi} \in \mathcal{S}} f(\boldsymbol{\pi}, \mathcal{C}, \Theta)$$

## C.13 Proof of Example 2

Let  $f$  be the mean-squared error of the difference-in-means estimator  $\hat{\tau}$  under the stable unit treatment value assumption:

$$f : (\boldsymbol{\pi}, \mathcal{C}, \{\mathbf{Y}(1), \mathbf{Y}(0)\}) \mapsto (TTE - \mathbb{E}_{\mathbf{Z}}[\hat{\tau}|\boldsymbol{\pi}])^2 + \text{Var}_{\mathbf{Z}}[\hat{\tau}|\boldsymbol{\pi}]$$

Assuming that the cluster-level outcomes are identical,

$$\forall \mathcal{C}_k, \mathcal{C}_j, Y_k^+(1) = Y_j^+(1) = y_t^+ \text{ and } Y_k^+(0) = Y_j^+(0) = y_c^+$$

then, for any  $\boldsymbol{\pi} \in \mathcal{S}$ , the bias squared is equal to

$$\begin{aligned}
(TTE - \mathbb{E}_{\mathbf{Z}}[\hat{\boldsymbol{\pi}}|\boldsymbol{\pi}])^2 &= \left( \sum_{j=1}^M \left( \frac{\pi_j}{n_t} - \frac{1}{N} \right) y_t^+ - \left( \frac{1 - \pi_j}{n_c} - \frac{1}{N} \right) y_c^+ \right)^2 \\
&= \left( y_t^+ \frac{M}{n_t} \left( \bar{\boldsymbol{\pi}} - \frac{n_t}{N} \right) - y_c^+ \frac{M}{n_c} \left( \bar{\boldsymbol{\pi}} - \frac{n_t}{N} \right) \right)^2 \\
&= \left( y_t^+ \frac{M}{n_t} - y_c^+ \frac{M}{n_c} \right)^2 \left( \bar{\boldsymbol{\pi}} - \frac{n_t}{N} \right)^2 \\
&= 0
\end{aligned}$$

Hence, the objective boils down to minimizing the conditional variance. From the proof of Proposition 11,

$$\text{Var}_{\mathbf{Z}}[\hat{\boldsymbol{\pi}}|\boldsymbol{\pi}] = \frac{N^2}{n_t n_c} \sum_{j=1}^M \frac{N_j}{N} \pi_j (1 - \pi_j) \left( \frac{S_{tj}}{n_t} + \frac{S_{cj}}{n_c} - \frac{S_{tcj}}{N} \right)$$

Since the coefficient  $\psi_j := \left( \frac{S_{tj}}{n_t} + \frac{S_{cj}}{n_c} - \frac{S_{tcj}}{N} \right)$  is always positive, the above is a sum of concave functions of  $\pi_j$ , which is itself a concave function of  $\boldsymbol{\pi}$ . Minimizing a concave function under the constraint  $\boldsymbol{\pi} \in [0, 1]^M$  is easy. Without loss of generality, we will assume that the coefficients  $\boldsymbol{\psi}$  are sorted in increasing order. We need to maintain the constraint that  $\sum_{j=1}^M \pi_j^* = \frac{Mn_t}{N}$ . While  $\sum_{j=1}^k \pi_j^* < \frac{Mn_t}{N}$ , set  $\pi_{k+1}^* = \min \left( 1, \frac{Mn_t}{N} - \sum_{j=1}^k \pi_j^* \right)$ . Set the remaining values of  $\boldsymbol{\pi}^*$  to 0.