

An Investigation into the Bias and Variance of Almost Matching Exactly Methods

by

Marco Morucci

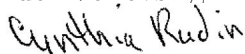
Department of Statistical Science
Duke University

Date: _____

Approved:



Alexander Volfovsky, Advisor



Cynthia D. Rudin



Fan Li

[Committee Member Name]

[Committee Member Name]

A thesis submitted in partial fulfillment of the
requirements for the degree of Master of Science
in the Department of Statistical Science
in the Graduate School of
Duke University

2021

ABSTRACT

An Investigation into the Bias and Variance of Almost Matching Exactly Methods

by

Marco Morucci

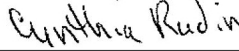
Department of Statistical Science
Duke University

Date: _____

Approved: _____



Alexander Volfovsky, Advisor



Cynthia D. Rudin



Fan Li

[Committee Member Name]

[Committee Member Name]

An abstract of a thesis submitted in partial fulfillment of the
requirements for the degree of Master of Science
in the Department of Statistical Science
in the Graduate School of
Duke University

2021

Copyright © 2021 by Marco Morucci
All rights reserved

Abstract

The development of interpretable causal estimation methods is a fundamental problem for high-stakes decision settings in which results must be explainable. Matching methods are highly explainable, but often lack the accuracy of black-box nonparametric models for causal effects. In this work, we propose to investigate theoretically the statistical bias and variance of Almost Matching Exactly (AME) methods for causal effect estimation. These methods aim to overcome the inaccuracy of matching by learning on a separate training dataset an optimal metric to match units on. While these methods are both powerful and interpretable, we currently lack an understanding of their statistical properties. In this work we present a theoretical characterization of the finite-sample and asymptotic properties of AME. We show that AME with discrete data has bounded bias in finite samples, and is asymptotically normal and consistent at a \sqrt{n} rate. Additionally, we show that AME methods for matching on networked data also have bounded bias and variance in finite-samples, and achieve asymptotic consistency in sparse enough graphs. Our results can be used to motivate the construction of approximate confidence intervals around AME causal estimates, providing a way to quantify their uncertainty.

Contents

Abstract	iv
List of Abbreviations	vi
1 Introduction	1
1.1 Causal Inference	2
1.2 Matching and Almost Matching Exactly	3
1.3 Plan of the Thesis and Summary of Contributions	4
2 Bias and Variance of Almost Matching Exactly Methods with Discrete Variables	5
2.1 A Bound on the Finite-Sample Bias of Discrete AME	6
2.2 Asymptotic Normality of Discrete AME	9
3 Bias of Almost Matching Exactly for Network Data	14
3.1 AME for Subgraph Matching	17
3.2 A Bound on the Finite-Sample Bias of AME for Network Data	18
3.3 Asymptotic Behavior	22
3.4 Heteroskedasticity in The Baseline Effects	23
4 Conclusions	25
Bibliography	26

List of Abbreviations

ADE: Average Direct Effect

AME: Almost Matching Exactly

CATE: Conditional Average Treatment Effect

CI: Conditional Ignorability

CLT: Central Limit Theorem

CRF: Conditional Response Function

MG: Matched Group

SUTVA: Stable Unit Treatment Value Assumption

Chapter 1

Introduction

Causal inference is growing in both usefulness and importance in virtually all areas of high-stakes decision making. Important decisions such as government policies, medical treatments, business strategies and judicial sentences are increasingly being informed by causal estimates obtained on either new or existing data. While these developments have increasingly led to better informed and more accurate decision-making, they also have created an explainability problem for the decision-makers. Complex, black-box algorithms are often used to compute causal estimates that inform high-stakes decisions, and decision-makers are left unable to *explain* how the algorithms themselves reached their conclusions on the given data. This is highly problematic in contexts in which decision-makers wishing to take causal findings into account for their decisions have to be accountable to some external entity: results from uninterpretable, black-box methods would be hard, if not impossible to justify to such an entity.

To remedy this problem, it is possible to employ interpretable methods, instead of black-boxes, to obtain causal estimates, and *matching*, the subject of this work, is one such interpretable method. In matching, similar individuals are compared to each other to determine what would happen if the same person received the treatment, i.e., the policy, medical intervention, business strategy, etc., in question, or did not. This comparison leads to easily interpretable estimates, which can be justified in terms of the cases that were compared to produce them. Unfortunately, the use of an interpretable method such as matching often comes at a cost: the simplicity of matching can lead to less accurate causal estimates than those of a complex black-box model on the same data.

In this work, we study a family of matching methods called Almost Matching Exactly (AME) that tries to marry the accuracy of complex machine-learning black-boxes, with the interpretability of matching. In short, AME tries to use machine learning to learn an accurate similarity metric between cases on the cases themselves. Matching is then performed by comparing cases on this learned, accurate similarity metric. In this way, the AME family of methods can deliver accurate and interpretable causal estimates that can be used to justify any high-stakes decision, while still be explainable “all the way down” to the data used to produce them. While the AME methods have been shown empirically to achieve good performance and to produce interpretable matches on both simulated and real-world datasets, little is still understood about their statistical properties. Specifically, it is important to be able to characterize the bias and variance of any statistical method, ideally in finite samples, but also asymptotically. This work aims to provide such characterization for several AME methods. Using both theoretical and empirical methodologies, we

will characterize the bias and variance of AME for discrete, continuous and network data. We will show that bounds on finite-sample bias and variance can be obtained for some of the AME methods, and that asymptotic behavior can be characterized as well in some cases.

1.1 Causal Inference

The problem of causal inference is that of estimating the effect of a treatment (a policy, business decision, medication, etc.) on an outcome (economic growth, profits, health indicators, etc.). Usually, we will have data on n cases or units, some of which have received the treatment, and some who have not. We will generally use the notation $t = 1$ to denote whether a unit that has received the treatment and $t = 0$ to denote a unit that has not. If we denote the value of the outcome for unit i under the treatment with $Y_i(1) \in \mathbb{R}$ and without the treatment with $Y_i(0) \in \mathbb{R}$, then these two variables will denote the value of the outcome either when the treatment has or has not been received by i . Notably, we will **never** observe both $Y_i(1)$ and $Y_i(0)$ at the same time: a unit either receives the treatment or does not. For all intents and purposes, these outcomes are only *potential*, and never both realized at the same time. Clearly, comparing these two values directly is impossible. This is known as the fundamental problem of causal inference [Hol86], and the aim of statistical causal inference is to provide methodologies to overcome it. In practice, we will model the treatment assignment received by a unit with a binary random variable, T_i , and the value of the outcome we actually observe with $Y_i = Y_i(1)T_i + Y_i(0)(1 - T_i)$. Generally, we will also observe a p -dimensional vector of contextual covariates for each unit. This will be denoted by the random variable $\mathbf{X}_i \in \mathbb{R}^p$. We will also be making the following general assumptions, unless otherwise stated:

- **(Strong) conditional ignorability (CI):** $(Y_i(1), Y_i(0)) \perp\!\!\!\perp T_i | \mathbf{X}_i$
- **Stable Unit Treatment Value Assumption (SUTVA):** Potential outcomes do not depend on treatments assigned to units other than i .
- **Random Sample:** $\{(Y_i(1), Y_i(0), T_i, \mathbf{X}_i)\}_{i=1}^n$ is an i.i.d. random sample from the joint distribution, of these variables, and $(Y(1), Y(0), T, \mathbf{X})$ denotes an arbitrary draw from this distribution, and $y(1), y(0), t, \mathbf{x}$ denote arbitrary values within the domain of this distribution.

Throughout this work, we will specify which of these assumptions holds in the setting studied, or if special cases are being analyzed.

Conditional Ignorability is the key assumption among these that permits us to indirectly compare potential outcomes, and reach a conclusion as to what outcomes a specific unit would attain with or without the treatment. This is the case because, it follows from CI that: $\mathbb{E}[Y(t)|\mathbf{X} = \mathbf{x}] = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}, T = t]$, and the quantity on the

rhs of the equality can be estimated from the observed data using regular statistical tools. This highlights the idea that what makes causal inference possible, are the assumptions that we make on our data, rather than the specific methodologies that we employ. Under the correct set of assumptions, causal results can be produced with interpretable methods as easily as they can be with black-boxes.

Throughout this work, we will be interested in the difference between the value of the potential outcomes of each unit. This quantity is known as the Conditional Average Treatment Effect (CATE), and is denoted as follows:

$$\tau(\mathbf{x}) = \mathbb{E}[Y_i(1) - Y_i(0)|\mathbf{X}_i = \mathbf{x}_i] = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}, T = 1] - \mathbb{E}[Y|\mathbf{X} = \mathbf{x}, T = 0],$$

where the second equality follows from the assumptions introduced before. Note that this quantity is the difference of two components: $\mu_1(\mathbf{x}) = \mathbb{E}[Y(1)|\mathbf{X} = \mathbf{x}]$, and $\mu_0(\mathbf{x}) = \mathbb{E}[Y(0)|\mathbf{X} = \mathbf{x}]$. These two quantities are known as Conditional Response Function (CRF), and their estimation is directly related to the problem of estimating the CATE.

1.2 Matching and Almost Matching Exactly

Once the assumptions needed to estimate the CATE are clarified, the problem of estimator choice arises: which is the best method to estimate $\tau(\mathbf{x})$ on the given data? As outlined at the beginning of this work, we would like our method to not only be accurate, but also interpretable: because of this we choose matching.

In a very general sense, matching works as follows: suppose that a unit, j , has covariates \mathbf{x} , and that we would like to compute $\tau(\mathbf{x})$, i.e., the CATE for unit j . A matching method will select a subset of the n units available, which in this work we call a **matched group**, and which we denote by \mathbf{MG} , such that each unit within \mathbf{MG} is considered matched to j . Given this subset of units, we will then use the following estimators for the CRFs and CATE:

$$\begin{aligned}\hat{\mu}_t(\mathbf{x}) &= \frac{1}{\sum_{i \in \mathbf{MG}} \mathbb{I}(T_i = t)} \sum_{i \in \mathbf{MG}} Y_i \mathbb{I}(T_i = t), \\ \hat{\tau}(\mathbf{x}) &= \hat{\mu}_1(\mathbf{x}) - \hat{\mu}_0(\mathbf{x}).\end{aligned}$$

note that the estimator for $\mu_t(\mathbf{x})$ is just the average of units who were assigned treatment t in the matched group, and the estimator for $\tau(\mathbf{x})$ is the difference between the two estimators for the respective CRFs.

The question now is how should matched groups be constructed: i.e., how should we choose which units to include in \mathbf{MG} ? In practice, most matching methods will compute some distance metric $d(\mathbf{x}, \mathbf{x}')$ between the covariates of candidate units and the unit to be matched, and choose the K -closest units in terms of d . This is simple and interpretable, but often leads to inaccurate estimates using the estimators just

defined. The AME methods aim to solve this problem by informing the selection of units to be included in **MG** with accurate machine learning predictions. The idea behind AME is to choose these units based on the outputs of an accurate but uninterpretable machine learning model applied to a subset of the data: since these models will likely accurately predict $\mu_t(\mathbf{x})$, their outputs can be used to inform which units should be included in **MG** to estimate the CATE accurately, but interpretably.

Characterizing the statistical error and variance of the estimators introduced is fundamental, since virtually all the downstream application of these estimators rely on some form of uncertainty and error quantification. For this purpose, this work aims to offer such quantification for two of the AME methodologies.

1.3 Plan of the Thesis and Summary of Contributions

This project aims to characterize the bias and variance of $\hat{\mu}_t$ and $\hat{\tau}$ when matches are made with the AME methodologies in two settings: first, when covariates are discrete and all three assumptions introduced prior hold. We will see that in this settings the bias of $\hat{\tau}$ for τ can be easily bounded in finite samples with quantities that depend directly on the matching method used, and we will also show that CATE and CRF estimates constructed with AME matches for discrete data are asymptotically normal at a \sqrt{n} rate. We will then move to a setting in which SUTVA does not hold, and the units are connected in an observed network. For an extension of the AME framework to this setting, this work will show that finite-sample bias can indeed be bounded, as a function of certain network features, and that some asymptotic properties of the estimators can indeed be established.

The work in this thesis has been published as part of the papers [WMA⁺17], and [AMO⁺20]. The theoretical results presented in this thesis were developed independently by the author, and included in the papers cited above. The descriptions of the algorithms that the results refer to were developed jointly between the author and his coauthors, and are included here as a way to contextualize and introduce the theoretical results that contribute the main body of this work.

This work will now proceed by introducing AME for discrete data, and by summarizing some theoretical results concerning bias and variance of CRF and CATE estimates. Second, the settings of causal inference with network data will be introduced, and an algorithm that adapts AME to this setting will be described. Following that, theoretical results concerning the finite-sample and asymptotic error of the AME-Networks method will be described.

Chapter 2

Bias and Variance of Almost Matching Exactly Methods with Discrete Variables

We start by introducing AME methods for matching with discrete covariates. In this setting, we have n units denoted by \mathcal{S}^{ma} , and indexed by i , each of which is associated with a set of p covariates \mathbf{x} , taking **binary** value 0 or 1. As previously introduced, we wish to estimate the CATE for each unit i . Note that, since covariates are binary, matches can only be made either exactly, or by completely ignoring certain covariates. When units do not have other units that exactly match with them on all p covariates, we will have to choose which covariate to ignore, and which others to match exactly on. How should we choose which covariates to match on and which to ignore when complete exact matches do not exist in the data?

We will use $\boldsymbol{\theta} \in \{0, 1\}^p$ to denote the variable selection indicator vector for a subset of covariates to match on. For AME we will define the *matched group* for a unit with covariates \mathbf{x} , with respect to covariates selected by $\boldsymbol{\theta}$ as the units that match \mathbf{x} exactly on the covariates $\boldsymbol{\theta}$:

$$\text{MG}(\mathbf{x}, \boldsymbol{\theta}) = \{i \in 1, \dots, n : \mathbf{X}_i \circ \boldsymbol{\theta} = \mathbf{x} \circ \boldsymbol{\theta}\}.$$

Suppose for now that we are given a set of p nonnegative weights, one for each covariate, denoted by \mathbf{w} , and that these weights denote the importance of the covariates for matching: covariates associated with a larger weight should be prioritized for matching. If we were given these weights, then we would select to match \mathbf{x} exactly on the covariates with the highest values of \mathbf{w} to as many units as possible. In practice, we would want to choose $\boldsymbol{\theta}$ that solves the following problem:

$$\boldsymbol{\theta}^* \in \arg \max_{\boldsymbol{\theta} \in \{0, 1\}^p} \mathbf{w}^T \boldsymbol{\theta}, \text{ s.t. : } \exists i \in \mathcal{S}^{ma} : \mathbf{X}_i \circ \boldsymbol{\theta} = \mathbf{x} \circ \boldsymbol{\theta}, \text{ and, } \mathbb{I}(T_i = t) = t,$$

i.e., we wish to find $\boldsymbol{\theta}$ that maximizes the sum of weights of covariates that are considered for matching, but also such that at least a unit with the desired treatment level, t , that matches exactly with \mathbf{x} on all the covariates selected by $\boldsymbol{\theta}$ exists in our dataset. Note that this problem can be solved separately for each unit i , using \mathbf{X}_i as input, to find the matched group $\text{MG}(\mathbf{X}_i, \boldsymbol{\theta})$. Once the problem is solved, the main matched groups can be used to estimate treatment effects, by considering the difference in outcomes between treatment and control units in each group, and possibly smoothing the estimates from the matched groups if desired, to prevent overfitting of treatment effect estimates.

How should we proceed when we are not given a vector of weights \mathbf{w} ? We suggest to use the following criterion to estimate covariate weights from the data: *We would*

like to ensure that each unit is matched using at least a set of covariates that is sufficient to predict outcomes well. Conversely, if a unit is matched using a set of covariates that do not predict outcomes sufficiently well, we would not trust the results from its matched group. Let us formalize the problem of matching each unit on a set of covariates that together predict outcomes well. The value of a set of covariates $\boldsymbol{\theta}$ is determined by how well these covariates can be used together to predict outcomes. In practice, \mathbf{w} can be estimated on a separated training set, by fitting an outcome model to the data, and computing prediction errors. For an in-depth description of how this can be done for discrete covariates, see [WMA⁺17] and [DLR⁺19].

Here, we will focus on studying the statistical bias and variance of the CRF and CATE estimators with either known, or arbitrary importance weights. The asymptotic result that we will give, will hold for any set of covariate importance weights.

2.1 A Bound on the Finite-Sample Bias of Discrete AME

If we do not match on all relevant covariates, a bias is induced on the treatment effect estimates. As shown before, solving the Full-AME problem ensures that this bias is as small as possible, and zero if the covariates excluded are all irrelevant (i.e., have weight 0). Here we present a simple worst-case bound on the in-sample estimation bias when a CATE is estimated with units matched according to a chosen subset of covariates (defined by $\boldsymbol{\theta}$). This bound is worst-case in that it holds for any subset of relevant covariates. This implies that the bias resulting from Full-AME will be much smaller than the bound given here in most cases.

Before proceeding with establishing this bound, we introduce some notation and define some quantities and assumptions: first, we will assume that there is no randomness in the potential outcomes, i.e., $Y_i(t) = \mu_t(\mathbf{x}_i)$. Additionally, define the variables: $n_1(\mathbf{x}, \boldsymbol{\theta}) = \sum_{i \in \text{MG}(\mathbf{x}, \boldsymbol{\theta})} T_i$ and $n_0(\mathbf{x}, \boldsymbol{\theta}) = \sum_{i \in \text{MG}(\mathbf{x}, \boldsymbol{\theta})} (1 - T_i)$ to be the count of units in $\text{MG}(\mathbf{x}, \boldsymbol{\theta})$ with treatment 1 and 0 respectively. Finally, Let $\mathbf{w}^T \mathbf{1}_{(\mathbf{x} \neq \mathbf{x}')}$ be the (positive) weighted Hamming distance between \mathbf{x} and \mathbf{x}' with a vector of weights of length p , denoted by \mathbf{w} , where each entry of \mathbf{w} is a positive real number, and such that $0 < \|\mathbf{w}\|_2 < \infty$, and define $M = \max_{\substack{\mathbf{x}, \mathbf{x}' \in \{0,1\}^p \\ t \in \{0,1\}}} \frac{|\mu_t(\mathbf{x}') - \mu_t(\mathbf{x})|}{\mathbf{w}^T \mathbf{1}_{(\mathbf{x}' \neq \mathbf{x})}}$.

Proposition 1. *Fix a value of the covariates $\mathbf{x} \in \{0,1\}^p$, and a value of $\boldsymbol{\theta} \in \{0,1\}^p$. We have, for any $\boldsymbol{\theta} \in \{0,1\}^p$:*

$$\left| \frac{1}{n_1(\mathbf{x}, \boldsymbol{\theta})} \sum_{i \in \text{MG}(\mathbf{x}, \boldsymbol{\theta})} Y_i T_i - \frac{1}{n_0(\mathbf{x}, \boldsymbol{\theta})} \sum_{i \in \text{MG}(\mathbf{x}, \boldsymbol{\theta})} Y_i (1 - T_i) - \tau(\mathbf{x}) \right| \leq 2M \mathbf{w}^T (\mathbf{1} - \boldsymbol{\theta}), \quad (2.1)$$

where $\mathbf{1}$ is a vector of length p that is 1 at all entries.

Proof. For a set of $i = 1, \dots, n$ units, consider potential outcomes $\mu_t(\mathbf{x}_i)$, with covariates $\mathbf{x} \in \{0, 1\}^p$ and treatment indicator $t \in \{0, 1\}$. Let $\mathbf{1}_{(\mathbf{x} \neq \mathbf{x}')}$ be a vector that is one at position j if $x_j \neq x'_j$ and 0 everywhere else. For all $\mathbf{x}, \mathbf{x}' \in \{0, 1\}^p$ and $t \in \{0, 1\}$, by the definition of M above, we have: $|\mu_t(\mathbf{x}) - \mu_t(\mathbf{x}')| \leq M \mathbf{w}^T \mathbf{1}_{(\mathbf{x} \neq \mathbf{x}')}$ and

$$\mu_t(\mathbf{x}) - M \mathbf{w}^T \mathbf{1}_{(\mathbf{x} \neq \mathbf{x}')} \leq \mu_t(\mathbf{x}') \leq \mu_t(\mathbf{x}) + M \mathbf{w}^T \mathbf{1}_{(\mathbf{x} \neq \mathbf{x}')} \quad (2.2)$$

Now we pick an arbitrary pair of \mathbf{x} and $\boldsymbol{\theta}$ and consider the term $\mathbf{w}^T \mathbf{1}_{(\mathbf{x} \neq \mathbf{x}')}$ for any $i \in \text{MG}(\mathbf{x}, \boldsymbol{\theta})$:

$$\begin{aligned} \mathbf{w}^T \mathbf{1}_{(\mathbf{x} \neq \mathbf{x}')} &= \sum_{j=1}^p w_j \mathbf{1}_{(x_j \neq x'_j)} && (x_j \text{ (resp. } x'_j) \text{ is the } j\text{-th entry of } \mathbf{x} \text{ (resp. } \mathbf{x}')) \\ &= \sum_{j=1}^p w_j (1 - \theta_j) \mathbf{1}_{(x_j \neq x'_j)} && (\text{Since } \mathbf{x}' \in \text{MG}(\mathbf{x}, \boldsymbol{\theta})) \\ &\leq \sum_{j=1}^p w_j (1 - \theta_j) = \mathbf{w}^T (\mathbf{1} - \boldsymbol{\theta}), \end{aligned} \quad (2.3)$$

where $\mathbf{1}$ is a vector of length p that is one in all entries. The second line follows because $\mathbf{x}' \in \text{MG}(\mathbf{x}, \boldsymbol{\theta})$ implies that \mathbf{x}' must match exactly with \mathbf{x} on the covariates selected by $\boldsymbol{\theta}$. We want to use the estimator $\hat{\tau}(\mathbf{x}, \boldsymbol{\theta}) = \frac{1}{n_1(\mathbf{x}, \boldsymbol{\theta})} \sum_{i \in \text{MG}(\mathbf{x}, \boldsymbol{\theta})} Y_i T_i - \frac{1}{n_0(\mathbf{x}, \boldsymbol{\theta})} \sum_{i \in \text{MG}(\mathbf{x}, \boldsymbol{\theta})} Y_i (1 - T_i)$ to estimate $\tau(\mathbf{x}) = \mu_1(\mathbf{x}) - \mu_0(\mathbf{x})$ for some fixed \mathbf{x} and $\boldsymbol{\theta}$. We can construct an upper bound on the estimation error of this estimator as follows:

$$\begin{aligned} &\frac{1}{n_1(\mathbf{x}, \boldsymbol{\theta})} \sum_{i \in \text{MG}(\mathbf{x}, \boldsymbol{\theta})} Y_i T_i - \frac{1}{n_0(\mathbf{x}, \boldsymbol{\theta})} \sum_{i \in \text{MG}(\mathbf{x}, \boldsymbol{\theta})} Y_i (1 - T_i) - \tau(\mathbf{x}) \\ &= \frac{1}{n_1(\mathbf{x}, \boldsymbol{\theta})} \sum_{i \in \text{MG}(\mathbf{x}, \boldsymbol{\theta})} \mu_1(\mathbf{x}_i) T_i - \frac{1}{n_0(\mathbf{x}, \boldsymbol{\theta})} \sum_{\mathbf{x}_i \in \text{MG}(\mathbf{x}, \boldsymbol{\theta})} \mu_0(\mathbf{x}_i) (1 - T_i) - (\mu_1(\mathbf{x}) - \mu_0(\mathbf{x})), \end{aligned} \quad (2.4)$$

where we use $Y_i = T_i \mu_1(\mathbf{x}_i) + (1 - T_i) \mu_0(\mathbf{x}_i)$, and $T_i^2 = T_i$ and $T_i(1 - T_i) = 0$ for any i . Then from (2.2) we get:

$$\begin{aligned} &\frac{1}{n_1(\mathbf{x}, \boldsymbol{\theta})} \sum_{i \in \text{MG}(\mathbf{x}, \boldsymbol{\theta})} \mu_1(\mathbf{x}_i) T_i - \frac{1}{n_0(\mathbf{x}, \boldsymbol{\theta})} \sum_{i \in \text{MG}(\mathbf{x}, \boldsymbol{\theta})} \mu_0(\mathbf{x}_i) (1 - T_i) - (\mu_1(\mathbf{x}) - \mu_0(\mathbf{x})) \\ &\leq \frac{1}{n_1(\mathbf{x}, \boldsymbol{\theta})} \sum_{i \in \text{MG}(\mathbf{x}, \boldsymbol{\theta})} [\mu_1(\mathbf{x}) + M \mathbf{w}^T \mathbf{1}_{(\mathbf{x} \neq \mathbf{x}_i)}] T_i \\ &\quad - \frac{1}{n_0(\mathbf{x}, \boldsymbol{\theta})} \sum_{i \in \text{MG}(\mathbf{x}, \boldsymbol{\theta})} [\mu_0(\mathbf{x}) - M \mathbf{w}^T \mathbf{1}_{(\mathbf{x} \neq \mathbf{x}_i)}] (1 - T_i) \\ &\quad - \mu_1(\mathbf{x}) + \mu_0(\mathbf{x}). \end{aligned} \quad (2.5)$$

Next, we combine (2.4) and (2.5) to get:

$$\begin{aligned}
& \frac{1}{n_1(\mathbf{x}, \boldsymbol{\theta})} \sum_{i \in \text{MG}(\mathbf{x}, \boldsymbol{\theta})} Y_i T_i - \frac{1}{n_0(\mathbf{x}, \boldsymbol{\theta})} \sum_{i \in \text{MG}(\mathbf{x}, \boldsymbol{\theta})} Y_i (1 - T_i) - \tau(\mathbf{x}) \\
& \leq \frac{1}{n_1(\mathbf{x}, \boldsymbol{\theta})} \sum_{i \in \text{MG}(\mathbf{x}, \boldsymbol{\theta})} [\mu_1(\mathbf{x}) + M \mathbf{w}^T \mathbf{1}_{(\mathbf{x} \neq \mathbf{x}_i)}] T_i \\
& \quad - \frac{1}{n_0(\mathbf{x}, \boldsymbol{\theta})} \sum_{i \in \text{MG}(\mathbf{x}, \boldsymbol{\theta})} [\mu_0(\mathbf{x}) - M \mathbf{w}^T \mathbf{1}_{(\mathbf{x} \neq \mathbf{x}_i)}] (1 - T_i) - \mu_1(\mathbf{x}) + \mu_0(\mathbf{x}) \\
& = \mu_1(\mathbf{x}) \frac{1}{n_1(\mathbf{x}, \boldsymbol{\theta})} \sum_{i \in \text{MG}(\mathbf{x}, \boldsymbol{\theta})} T_i - \mu_0(\mathbf{x}) \frac{1}{n_0(\mathbf{x}, \boldsymbol{\theta})} \sum_{i \in \text{MG}(\mathbf{x}, \boldsymbol{\theta})} (1 - T_i) - \mu_1(\mathbf{x}) + \mu_0(\mathbf{x}) \\
& \quad + \sum_{i \in \text{MG}(\mathbf{x}, \boldsymbol{\theta})} M \mathbf{w}^T \mathbf{1}_{(\mathbf{x} \neq \mathbf{x}_i)} \left(\frac{T_i}{n_1(\mathbf{x}, \boldsymbol{\theta})} + \frac{1 - T_i}{n_0(\mathbf{x}, \boldsymbol{\theta})} \right) \\
& = \sum_{i \in \text{MG}(\mathbf{x}, \boldsymbol{\theta})} M \mathbf{w}^T \mathbf{1}_{(\mathbf{x} \neq \mathbf{x}_i)} \left(\frac{T_i}{n_1(\mathbf{x}, \boldsymbol{\theta})} + \frac{1 - T_i}{n_0(\mathbf{x}, \boldsymbol{\theta})} \right) \\
& \leq \sum_{i \in \text{MG}(\mathbf{x}, \boldsymbol{\theta})} M \mathbf{w}^T (1 - \theta) \left(\frac{T_i}{n_1(\mathbf{x}, \boldsymbol{\theta})} + \frac{1 - T_i}{n_0(\mathbf{x}, \boldsymbol{\theta})} \right) \\
& = 2M \mathbf{w}^T (1 - \theta),
\end{aligned}$$

where the inequality in the second to last line follows from Equation (2.3). Using the same set of steps we can construct a lower bound on the estimation error:

$$\begin{aligned}
& \frac{1}{n_1(\mathbf{x}, \boldsymbol{\theta})} \sum_{i \in \text{MG}(\mathbf{x}, \boldsymbol{\theta})} Y_i T_i - \frac{1}{n_0(\mathbf{x}, \boldsymbol{\theta})} \sum_{i \in \text{MG}(\mathbf{x}, \boldsymbol{\theta})} Y_i (1 - T_i) - \tau(\mathbf{x}) \\
& \geq \frac{1}{n_1(\mathbf{x}, \boldsymbol{\theta})} \sum_{i \in \text{MG}(\mathbf{x}, \boldsymbol{\theta})} [\mu_1(\mathbf{x}) - M \mathbf{w}^T \mathbf{1}_{(\mathbf{x} \neq \mathbf{x}_i)}] T_i \\
& \quad - \frac{1}{n_0(\mathbf{x}, \boldsymbol{\theta})} \sum_{i \in \text{MG}(\mathbf{x}, \boldsymbol{\theta})} [\mu_0(\mathbf{x}) + M \mathbf{w}^T \mathbf{1}_{(\mathbf{x} \neq \mathbf{x}_i)}] (1 - T_i) \\
& \quad - \mu_1(\mathbf{x}) + \mu_0(\mathbf{x}) \\
& \geq - \sum_{i \in \text{MG}(\mathbf{x}, \boldsymbol{\theta})} M \mathbf{w}^T (1 - \theta) \left(\frac{T_i}{n_1(\mathbf{x}, \boldsymbol{\theta})} + \frac{1 - T_i}{n_0(\mathbf{x}, \boldsymbol{\theta})} \right) \\
& = - 2M \mathbf{w}^T (1 - \theta).
\end{aligned}$$

Putting together these two bounds we obtain the statement in the proposition. \square

As asserted by Proposition 1, we should select $\boldsymbol{\theta}$ to minimize $\mathbf{w}^T (1 - \boldsymbol{\theta})$ in order to minimize AME's bias. In real problems, we should think about \mathbf{w} as a non-uniform vector that has some small entries. AME would tend to remove those entries so that

the bias is minimized for the remaining covariates that are used for matching. This bound provides guidance in designing the AME procedure as it suggests that the amount of bias, in estimating treatment effects with almost-exact matching, depends heavily on $\boldsymbol{\theta}$: AME will try to match on sets of covariates that minimize such bias.

2.2 Asymptotic Normality of Discrete AME

We now seek to establish the asymptotic behavior of Discrete AME causal estimates. Since we will be dealing with population quantities, we will consider the variable $Y(t)$ as random, with: $\mu_t(\mathbf{x}) = \mathbb{E}[Y(t)|\mathbf{X} = \mathbf{x}]$, for $t \in \{0, 1\}$. The following theorem describes the asymptotic behavior of AME estimates:

Theorem 1. *Let all quantities be defined as prior, assume conditional ignorability, SUTVA, and boundedness of $\mathbb{E}[|Y(t)|^{2+\delta}|\mathbf{X} = \mathbf{x}]$ for all t and \mathbf{x} , and some $\delta > 0$. Let $\boldsymbol{\theta}$ be the solution to the universal AME problem:*

$$\boldsymbol{\theta} \in \arg \min_{\boldsymbol{\theta} \in \{0,1\}^p} \mathbf{1}^T \boldsymbol{\theta}, \text{ s.t. : } \exists i = 1, \dots, n : \mathbf{X}_i \circ \boldsymbol{\theta} = \mathbf{x} \circ \boldsymbol{\theta}, \text{ and, } T_i = t,$$

where $\mathbf{1}$ is a vector of length p that is one in all entries. Then:

$$\sqrt{n}(\hat{\mu}_t(\mathbf{x}) - \mu_t(\mathbf{x})) \xrightarrow{d} \mathcal{N}\left(0, \frac{\sigma_t^2(\mathbf{x})}{\Pr(\mathbf{X} = \mathbf{x}, T = t)}\right),$$

and,

$$\sqrt{n}(\hat{\tau}(\mathbf{x}) - \tau(\mathbf{x})) \xrightarrow{d} \mathcal{N}\left(0, \frac{\sigma_1^2(\mathbf{x})}{\Pr(\mathbf{X} = \mathbf{x}, T = 1)} + \frac{\sigma_0^2(\mathbf{x})}{\Pr(\mathbf{X} = \mathbf{x}, T = 0)}\right).$$

The proof follows. Note that, as we will show, asymptotically $\boldsymbol{\theta}$ that solves the universal AME problem will equal $\mathbf{1}$ almost surely. This implies that any $\boldsymbol{\theta}$ that solves the AME problem with **any** set of weights on the covariates will eventually also converge to $\mathbf{1}$, thus displaying the same set of asymptotic properties as $\boldsymbol{\theta}$. We also note that a similar result in the context of balancing weights is given in Theorem 2 of [LMZ18].

Proof. First we show that, if $\boldsymbol{\theta}$ solves the universal AME problem, then it will almost surely converge to $\mathbf{1}$, we have, for any $\mathbf{x} \in \{0, 1\}^p$:

$$\begin{aligned} \Pr(\boldsymbol{\theta} \neq \mathbf{1}) &= \Pr(\nexists i \in 1, \dots, n : \mathbf{X}_i = \mathbf{x}, T_i = t) \\ &= \prod_{i=1}^n (1 - \Pr(\mathbf{X}_i = \mathbf{x}, T_i = t)) \\ &= (1 - \Pr(\mathbf{X} = \mathbf{x}, T = t))^n, \end{aligned}$$

where the second and third equalities follows by the random sample assumption. Finally, it follows that $\lim_{n \rightarrow \infty} (1 - \Pr(\mathbf{X}_i = \mathbf{x}, T_i = t))^n = 0$. We now move to show asymptotic normality of $\hat{\mu}_t(\mathbf{x})$. Start by noticing that our quantity of interest, $\sqrt{n}(\hat{\mu}_t(\mathbf{x}) - \mu_t(\mathbf{x}))$, can be decomposed as follows:

$$\begin{aligned}
\hat{\mu}_t(\mathbf{x}) &= \frac{\sqrt{n}}{n_t(\mathbf{x}, \boldsymbol{\theta})} \sum_{i \in \text{MG}(\mathbf{x}, \boldsymbol{\theta})} Y_i \mathbb{I}(T_i = t) \\
&= \frac{\sqrt{n}}{n_t(\mathbf{x}, \boldsymbol{\theta})} \sum_{i \in \text{MG}(\mathbf{x}, \boldsymbol{\theta})} (Y_i - \mu_t(\mathbf{X}_i)) \mathbb{I}(T_i = t) + \frac{\sqrt{n}}{n_t(\mathbf{x}, \boldsymbol{\theta})} \sum_{i \in \text{MG}(\mathbf{x}, \boldsymbol{\theta})} (\mu_t(\mathbf{X}_i) - \mu_t(\mathbf{x})) \mathbb{I}(T_i = t) \\
&\leq \frac{\sqrt{n}}{n_t(\mathbf{x}, \boldsymbol{\theta})} \sum_{i \in \text{MG}(\mathbf{x}, \boldsymbol{\theta})} (Y_i - \mu_t(\mathbf{X}_i)) \mathbb{I}(T_i = t) + \sqrt{n} M \mathbf{w}^T (\mathbf{1} - \boldsymbol{\theta}) \\
&= \sum_{i=1}^n D_i + o_p(1),
\end{aligned}$$

where $D_i = \frac{\sqrt{n}}{n_t(\mathbf{x}, \boldsymbol{\theta})} (Y_i - \mu_t(\mathbf{X}_i)) \mathbb{I}(T_i = t)$, and the inequality follows from Proposition 1, and the fact that $M \mathbf{w}^T (\mathbf{1} - \boldsymbol{\theta})$ converges to 0 exponentially fast, as we have just shown.

To prove the theorem, it remains to show that $D_n = \sum_{i \in \text{MG}(\mathbf{x}, \boldsymbol{\theta})} D_i$ is asymptotically normally distributed. We can do so by applying the Lindeberg-Levy CLT to this quantity. The theorem requires three conditions to hold on each D_i . First: $\mathbb{E}[D_i] = 0$, which is easily verified because, by definition $\mathbb{E}[Y_i T_i | \mathbf{X}_i] = \mu_t(\mathbf{X}_i) \mathbb{E}[T_i | \mathbf{X}_i]$ for all i .

Second, the limit as $n \rightarrow \infty$ of $\sum_{i \in \text{MG}(\mathbf{x}, \boldsymbol{\theta})} \mathbb{V}[D_i]$ must be a constant. This can be

seen as follows:

$$\begin{aligned}
\mathbb{V}[D_i | i \in \text{MG}(\mathbf{x}, \boldsymbol{\theta})] &= \mathbb{E}[D_i^2 | i \in \text{MG}(\mathbf{x}, \boldsymbol{\theta})] \\
&= \mathbb{E} \left[\mathbb{E} \left[\left(\frac{\sqrt{n} \mathbb{I}(T_i = t)}{n_t(\mathbf{x}, \boldsymbol{\theta})} \right)^2 \middle| \mathbf{X}_i, i \in \text{MG}(\mathbf{x}, \boldsymbol{\theta}) \right] \sigma^2(\mathbf{X}_i) \right] \\
&= \mathbb{E} \left[\left(\frac{\sqrt{n} \mathbb{I}(T_i = t)}{n_t(\mathbf{x}, \boldsymbol{\theta})} \right)^2 \middle| i \in \text{MG}(\mathbf{x}, \boldsymbol{\theta}) \right] \sigma^2(\mathbf{x}) \\
&\quad + \mathbb{E} \left[\mathbb{E} \left[\left(\frac{\sqrt{n} \mathbb{I}(T_i = t)}{n_t(\mathbf{x}, \boldsymbol{\theta})} \right)^2 \middle| \mathbf{X}_i, i \in \text{MG}(\mathbf{x}, \boldsymbol{\theta}) \right] (\sigma^2(\mathbf{X}_i) - \sigma^2(\mathbf{x})) \right] \\
&\leq \mathbb{E} \left[\left(\frac{\sqrt{n} \mathbb{I}(T_i = t)}{n_t(\mathbf{x}, \boldsymbol{\theta})} \right)^2 \middle| i \in \text{MG}(\mathbf{x}, \boldsymbol{\theta}) \right] \sigma^2(\mathbf{x}) \\
&\quad + \mathbb{E} \left[\left(\frac{\sqrt{n} \mathbb{I}(T_i = t)}{n_t(\mathbf{x}, \boldsymbol{\theta})} \right)^2 \middle| i \in \text{MG}(\mathbf{x}, \boldsymbol{\theta}) \right] M \mathbf{w}^T (\mathbf{1} - \boldsymbol{\theta}) \\
&\leq \mathbb{E} \left[\left(\frac{\sqrt{n} \mathbb{I}(T_i = t)}{n_t(\mathbf{x}, \boldsymbol{\theta})} \right)^2 \middle| i \in \text{MG}(\mathbf{x}, \boldsymbol{\theta}) \right] \sigma^2(\mathbf{x}) + n C \mathbb{E}[\mathbf{w}^T (\mathbf{1} - \boldsymbol{\theta})] \\
&= \mathbb{E} \left[\left(\frac{\sqrt{n} \mathbb{I}(T_i = t)}{n_t(\mathbf{x}, \boldsymbol{\theta})} \right)^2 \middle| i \in \text{MG}(\mathbf{x}, \boldsymbol{\theta}) \right] \sigma^2(\mathbf{x}) + o_p(1).
\end{aligned}$$

The first inequality follows because boundedness of $\mu_t(\mathbf{x})$ for all \mathbf{x} implies boundedness of $\sigma^2(\mathbf{x})$ for all \mathbf{x} and some constant C , which then lets us apply Proposition 1 to bound the difference $\sigma^2(\mathbf{X}_i) - \sigma^2(\mathbf{x})$. The second inequality follows by upper bounding the fraction with 1, and the last line follows because $\boldsymbol{\theta}$ converges to $\mathbf{1}$ exponentially fast. Note now that:

$$\sum_{i \in \text{MG}(\mathbf{x}, \boldsymbol{\theta})} \mathbb{E} \left[\left(\frac{\sqrt{n} \mathbb{I}(T_i = t)}{n_t(\mathbf{x}, \boldsymbol{\theta})} \right)^2 \middle| i \in \text{MG}(\mathbf{x}, \boldsymbol{\theta}) \right] \sigma^2(\mathbf{x}) = \mathbb{E} \left[\frac{n}{n_t(\mathbf{x}, \boldsymbol{\theta})} \right] \sigma^2(\mathbf{x}),$$

by definition of $n_t(\mathbf{x}, \boldsymbol{\theta})$. We now need to establish convergence of the quantity $\frac{n}{n_t(\mathbf{x}, \boldsymbol{\theta})}$. This can be done by noticing that $n_t(\mathbf{x}, \boldsymbol{\theta})$ is a binomial random variable with size n , and probability $\Pr(\mathbf{X} \circ \boldsymbol{\theta} = \mathbf{x} \circ \boldsymbol{\theta}, T = t)$. Therefore:

$$\mathbb{E} \left[\frac{n_t(\mathbf{x}, \boldsymbol{\theta})}{n} \right] = \frac{n \Pr(\mathbf{X} \circ \boldsymbol{\theta} = \mathbf{x} \circ \boldsymbol{\theta}, T = t)}{n} = \Pr(\mathbf{X} \circ \boldsymbol{\theta} = \mathbf{x} \circ \boldsymbol{\theta}, T = t),$$

and since $\boldsymbol{\theta} \xrightarrow{a.s.} \mathbf{1}$, it must be that $\Pr(\mathbf{X} \circ \boldsymbol{\theta} = \mathbf{x} \circ \boldsymbol{\theta}, T = t) \rightarrow \Pr(\mathbf{X} = \mathbf{x}, T = t)$. Therefore: $\mathbb{E} \left[\frac{n}{n_t(\mathbf{x}, \boldsymbol{\theta})} \right] \sigma^2(\mathbf{x}) \rightarrow \frac{\sigma^2(\mathbf{x})}{\Pr(\mathbf{X} = \mathbf{x}, T = t)}$.

Third and final, Lindeberg's regularity condition, which is implied by Lyapunov's condition, a stronger, but easier to verify, criterion. This latter condition states that

$\lim_{n \rightarrow \infty} \mathbb{E}[|D_i|^{2+\delta}] \rightarrow 0$. For simplicity define W_i as $\mathbb{I}(i \in \text{MG}(\mathbf{x}, \boldsymbol{\theta}))$, a random variable denoting whether unit i is a member of $\text{MG}(\mathbf{x}, \boldsymbol{\theta})$. We have, for some $\delta > 0$:

$$\begin{aligned} \mathbb{E}[|D_i|^{2+\delta}] &= \mathbb{E} \left[\left| \underbrace{\frac{\sqrt{n}}{n_t(\mathbf{x}, \boldsymbol{\theta})}}_{\geq 0} \underbrace{W_i \mathbb{I}(T_i = t)}_{\in \{0,1\}} (Y_i - \mu_t(\mathbf{X}_i)) \right|^{2+\delta} \right] \\ &= \mathbb{E} \left[\mathbb{I}(T_i = t) W_i \frac{n^{1+\delta/2}}{n_t(\mathbf{x}, \boldsymbol{\theta})^{2+\delta}} |Y_i - \mu_t(\mathbf{X}_i)|^{2+\delta} \right], \end{aligned}$$

Recall now that $\mathbb{E}[|Y_i|^{2+\delta} | \mathbf{X}_i, \mathbb{I}(T_i = t)] \leq M$ for some constant $M < \infty$ by assumption. Since conditionally on \mathbf{X}_i and $\mathbb{I}(T_i = t)$, both $\mu_t(\mathbf{X}_i)$, $n_t(\mathbf{x}, \boldsymbol{\theta})$, W_i and $\mathbb{I}(T_i = t)$ are constants, it follows that:

$$\begin{aligned} \mathbb{E} \left[W_i \frac{n^{1+\delta/2}}{n_t(\mathbf{x}, \boldsymbol{\theta})^{2+\delta}} |Y_i - \mu_t(\mathbf{X}_i)|^{2+\delta} \right] &= \mathbb{E} \left[W_i \frac{n^{1+\delta/2}}{n_t(\mathbf{x}, \boldsymbol{\theta})^{2+\delta}} \mathbb{E}[|Y_i - \mu_t(\mathbf{X}_i)|^{2+\delta} | \mathbf{X}_i, \mathbb{I}(T_i = t)] \right] \\ &\leq \mathbb{E} \left[W_i \frac{n^{1+\delta/2}}{n_t(\mathbf{x}, \boldsymbol{\theta})^{2+\delta}} M \right]. \end{aligned}$$

Putting the bounds together we have, as $n \rightarrow \infty$:

$$\begin{aligned} \sum_{i=1}^N \mathbb{E}[|D_i|^{2+\delta}] &\leq M \sum_{i=1}^n \mathbb{E} \left[W_i \frac{n^{1+\delta/2}}{n_t(\mathbf{x}, \boldsymbol{\theta})^{2+\delta}} \right] \\ &= M \mathbb{E} \left[\frac{n^{1+\delta/2}}{n_t(\mathbf{x}, \boldsymbol{\theta})^{1+\delta}} \right] = M \frac{1}{n^{\delta/2}} \mathbb{E} \left[\left(\frac{n}{n_t(\mathbf{x}, \boldsymbol{\theta})} \right)^{1+\delta} \right]. \end{aligned}$$

Finally, we can show that the quantity: $\mathbb{E} \left[\left(\frac{n}{n_t(\mathbf{x}, \boldsymbol{\theta})} \right)^{1+\delta} \right]$ is asymptotically bounded for $\delta = 1$ as follows:

$$\begin{aligned} \mathbb{E} \left[\left(\frac{n_t(\mathbf{x}, \boldsymbol{\theta})}{n} \right)^2 \right] &= \frac{n(n-1) \Pr(\mathbf{X} \circ \boldsymbol{\theta} = \mathbf{x} \circ \boldsymbol{\theta}, T = t)^2 + n \Pr(\mathbf{X} \circ \boldsymbol{\theta} = \mathbf{x} \circ \boldsymbol{\theta}, T = t)}{n^2} \\ &= \Pr(\mathbf{X} \circ \boldsymbol{\theta} = \mathbf{x} \circ \boldsymbol{\theta}, T = t)^2 \\ &\quad - \frac{\Pr(\mathbf{X} \circ \boldsymbol{\theta} = \mathbf{x} \circ \boldsymbol{\theta}, T = t)^2}{n} + \frac{\Pr(\mathbf{X} \circ \boldsymbol{\theta} = \mathbf{x} \circ \boldsymbol{\theta}, T = t)}{n} \\ &= \Pr(\mathbf{X} \circ \boldsymbol{\theta} = \mathbf{x} \circ \boldsymbol{\theta}, T = t)^2 + o(1) = O(1), \end{aligned}$$

where the expansion in the first equality follows from properties of binomial random variables. Therefore for $\delta = 1$:

$$= M \frac{1}{n^{r(1/2)}} O(1) \rightarrow 0,$$

and Lyapunov's condition holds, proving the statement in the theorem. The statement for the CATE follows because $\hat{\mu}_1(\mathbf{x})$ and $\hat{\mu}_0(\mathbf{x})$ are independent of each other, since either only treated or control units are used in each estimator, and, therefore, their difference will asymptotically converge to the difference of their limiting distributions. \square

This theorem shows that discrete AME estimates are asymptotically normal at a rate of \sqrt{n} . This result is not unexpected: matches are always exact asymptotically due to the discreteness of the covariate space. However, the result is instructive because it theoretically justifies the construction of approximate confidence intervals around $\hat{\mu}_t(\mathbf{x})$ and $\hat{\tau}(\mathbf{x})$ using the classical standard normal approximation with asymptotic variance given in the theorem. Note that this variance can be estimated with the regular variance estimator applied to $\mathbf{MG}(\mathbf{x}, \boldsymbol{\theta})$, and the probability in the denominator, with any consistent density estimator.

Chapter 3

Bias of Almost Matching Exactly for Network Data

In this section, we present an extension of the AME framework to the problem of interference in networked experiments: experimental units are connected in a network represented as a graph, and treatment is allowed to “spread” from unit to connected unit. This will, in order, lead to biased causal estimates. To remedy this issue, we will propose a set of assumptions together with an algorithm that will allow us to account for this interference. Following the introduction of this methodology, we will study its finite-sample and asymptotic behavior in a certain setting. We begin by stating some additional notation needed for the network causal inference setting.

We have a set of n experimental units indexed by i . These units are connected in a known graph $G = (V, E)$, where $V(G) = \{1, \dots, n\}$ is the set of vertices of G , and $E(G)$ is the set of edges of G . We disallow self-loops in our graph. We say that H is a subgraph of G if $V(H) \subseteq V(G)$ and $E(H) \subseteq E(G)$. Let $t_i \in \{0, 1\}$ represent the treatment indicator for unit i , \mathbf{t} represent the vector of treatment indicators for the entire sample, and \mathbf{t}_{-i} represent the treatment indicators for all units except i . Given a treatment vector \mathbf{t} on the entire sample (i.e., for all vertices in G), we use $G^{\mathbf{t}}$ to denote the *labeled graph*, where each vertex $i \in V(G)$ has been labeled with its treatment indicator t_i . In addition, we use G_P to denote a graph induced by the set of vertices $P \subseteq V(G)$ on G , such that $V(G_P) = P$ and $E(G_P) = \{(e_1, e_2) \in E(G) : e_1 \in P, e_2 \in P\}$. We use the notation $\mathcal{N}_i = \{j : (i, j) \in E(G)\}$ to represent the neighborhood of vertex i . The labeled neighborhood graph of a unit i , $G_{\mathcal{N}_i}^{\mathbf{t}}$, is defined as the graph induced by the neighbors of i , and labeled according to \mathbf{t} . We also define $\mathbf{t}_{\mathcal{N}_i}$ to be the vector of treatment indicators corresponding to unit i ’s neighborhood graph. A unit’s response to the treatment is represented by its random potential outcomes $Y_i(\mathbf{t}) = Y_i(t_i, \mathbf{t}_{-i})$. Unlike other commonly studied causal inference settings, unit i ’s potential outcomes are now a function of both the treatment assigned to i , and of all other units’ treatments. Observed treatments for unit i and the whole sample are represented by the random variables T_i and \mathbf{T} respectively. We assume that the number of treated units is always $n^{(1)}$, i.e., $\sum_{i=1}^n T_i = n^{(1)}$.

Unlike the discrete AME setting, we **do not** have contextual covariates now. Therefore, we assume that ignorability holds unconditionally on the covariates, i.e., $Y_i(1), Y_i(0) \perp\!\!\!\perp T_i$. As stated before, we **do not** make the canonical Stable Unit Treatment Value Assumption (SUTVA) [Rub80], which, among other requirements, states that units are exclusively affected by the treatment assigned to them. We do

not make this assumption because our units are connected in a network: it could be possible for treatments to spread along the edges of the network and to affect connected units' outcomes. We do maintain the assumption of comparable treatments across units, which is commonly included in SUTVA.

Our causal quantity of interest will be the Average Direct Effect (ADE), which is defined as follows:

$$ADE = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[Y_i(1, \mathbf{0}) - Y_i(0, \mathbf{0})], \quad (3.1)$$

where $\mathbf{t}_{-i} = \mathbf{0}$ represents the treatment assignment in which no unit other than i is treated. The summand represents the treatment effect on unit i when no other unit is treated, and, therefore, no interference occurs [HS95].

We outline the requirements of our framework for direct effect estimation under interference. We denote interference effects on a unit i with the function $f_i(\mathbf{t}) : \{0, 1\}^n \mapsto \mathbb{R}$, a function that maps each possible treatment allocation for the n units to the amount of interference on unit i . We will use several assumptions to restrict the domain of f to a much smaller set (and overload the notation f_i accordingly). To characterize f , we rely on the typology of interference assumptions introduced by [SA17]. The first three assumptions (A1-A2) needed in our framework are common in the interference literature [Man13, TK13, EKB16, AEI18]:

A1: Additivity of Main Effects. First, we assume that main treatment effects are additive, i.e., that there is no interaction between units' treatment indicators. This allows us to write:

$$Y_i(t, \mathbf{t}_{-i}) = t\tau_i + f_i(\mathbf{t}_{-i}) + \epsilon_i \quad (3.2)$$

where τ_i is the direct treatment effect on unit i , and ϵ_i is some baseline effect.

A2: Neighborhood Interference. We focus on a specific form of the interference function f_i by assuming that the interference experienced by unit i depends only on treatment of its neighbors. That is, if for two treatment allocations \mathbf{t}, \mathbf{t}' we have $\mathbf{t}_{\mathcal{N}_i} = \mathbf{t}'_{\mathcal{N}_i}$ then $f_i(\mathbf{t}) = f_i(\mathbf{t}')$. To make explicit this dependence on the neighborhood subgraph, we will write $f_i(\mathbf{t}_{\mathcal{N}_i}) \equiv f_i(\mathbf{t})$.

A3: Isomorphic Graph Interference We assume that, if two units i and j have *isomorphic labeled neighborhood graphs*, then they receive the same amount of interference, denoting isomorphism by \simeq , $G_{\mathcal{N}_i}^{\mathbf{t}} \simeq G_{\mathcal{N}_j}^{\mathbf{t}} \implies f_i(\mathbf{t}_{\mathcal{N}_i}) = f_j(\mathbf{t}_{\mathcal{N}_j}) \equiv f(G_{\mathcal{N}_i}^{\mathbf{t}}) = f(G_{\mathcal{N}_j}^{\mathbf{t}})$. While Assumptions A1 and A2 are standard, A3 is new. This assumption allows us to study interference in a setting where units with similar neighborhood subgraphs experience similar amounts of interference.

All our assumptions together induce a specific form for the potential outcomes, namely that they depend on neighborhood structure $G_{\mathcal{N}_i}^{\mathbf{t}}$, but not exactly who the neighbors are (information contained in \mathcal{N}_i) nor treatment assignments for those outside the neighborhood (information contained in $\mathbf{t}_{\mathcal{N}_i}$). Namely:

Proposition 2. *Under ignorability, overlap, and assumptions A1-A3, potential outcomes in (3.2) for all units i can be written as:*

$$Y_i(t, \mathbf{t}_{-i}) = t\tau_i + f(G_{\mathcal{N}_i}^{\mathbf{t}}) + \epsilon_i, \quad (3.3)$$

where τ_i is the direct treatment effect on unit i , and ϵ_i is some baseline response.

In addition, suppose that baseline responses for all units are equal to each other in expectation, i.e., for all i , $\mathbb{E}[\epsilon_i] = \alpha$. Then under assumptions A1-A3, for neighborhood graph structures g_i of unit i and treatment vectors \mathbf{t} , the ADE is identified as:

$$\begin{aligned} ADE = \frac{1}{n(1)} \sum_{i=1}^n \mathbb{E} [T_i \times (\mathbb{E}[Y_i | G_{\mathcal{N}_i}^{\mathbf{T}} \simeq g_i^{\mathbf{t}}, T_i = 1] \\ - \mathbb{E}[Y_i | G_{\mathcal{N}_i}^{\mathbf{T}} \simeq g_i^{\mathbf{t}}, T_i = 0])], \end{aligned}$$

where $G_{\mathcal{N}_i}^{\mathbf{T}}$ is the neighborhood graph of i labelled according to the treatment assignment \mathbf{T} .

The proposition (whose proof follows) states that the interference received by a unit is a function of each unit's neighborhood graph. Further, the outcomes can be decomposed additively into this function and the direct treatment effect on i . The proposition implies that the ADE is identified by matching each treated unit to one or more control units with an isomorphic neighborhood graph, and computing the direct effect on the treated using these matches. This effect is, in expectation over individual treatment assignments, equal to the ADE.

Proof. We start by writing potential outcomes for an arbitrary unit i from (3.2) as $Y_i(t, \mathbf{t}_{-i}) = f_i(t_i, \mathbf{t}_{-i}) + \epsilon_i$, where ϵ_i is some pre-treatment value of the potential outcome, and $f_i(t, \mathbf{t}_{-i})$ is the treatment response function, dependent both on unit i 's treatment and on everyone else's. Using A1-A3 we can write (3.2) as:

$$\begin{aligned} Y_i(t, \mathbf{t}_{-i}) &= t\tau_i + f_i(\mathbf{t}_{-i}) + \epsilon_i && \text{(By A1)} \\ &= t\tau_i + f_i(\mathbf{t}_{\mathcal{N}_i}) + \epsilon_i && \text{(By A2)} \\ &= t\tau_i + f(G_{\mathcal{N}_i}^{\mathbf{t}}) + \epsilon_i && \text{(By A3).} \end{aligned}$$

To prove identification for the ADE, we must show that the individual TE is identified for a treated unit i . We need to show that: $\mathbb{E}[Y_i(1, \mathbf{0}) - Y_i(0, \mathbf{0})] = \mathbb{E}[Y_i | T_i = 1, S(G_{\mathcal{N}_i}^{\mathbf{T}}) = s] - \mathbb{E}[Y_j | T_j = 0, S(G_{\mathcal{N}_j}^{\mathbf{T}}) = s]$. Assuming that $\mathbb{E}[\epsilon_i] = \mathbb{E}[\epsilon_j] = \alpha$ we have first:

$$\mathbb{E}[Y_i(1, \mathbf{0}) - Y_i(0, \mathbf{0})] = \mathbb{E}[\epsilon_i + \tau_i - \epsilon_i] = \tau_i, \quad (3.4)$$

with the first equality following from (3.3). Second we have:

$$\begin{aligned}
& \mathbb{E}[Y_i|G_{\mathcal{N}_i}^{\mathbf{T}} \simeq g^{\mathbf{t}}, T_i = 1] - \mathbb{E}[Y_j|G_{\mathcal{N}_j}^{\mathbf{T}} \simeq g^{\mathbf{t}}, T_j = 0] \\
&= \mathbb{E}[Y_i(1, \mathbf{T}_{-i})T_i + Y_i(0, \mathbf{T}_{-i})(1 - T_i)|G_{\mathcal{N}_i}^{\mathbf{T}} \simeq g^{\mathbf{t}}, T_i = 1] \\
&\quad - \mathbb{E}[Y_j(1, \mathbf{T}_{-j})T_j + Y_j(0, \mathbf{T}_{-j})(1 - T_j)|G_{\mathcal{N}_j}^{\mathbf{T}} \simeq g^{\mathbf{t}}, T_j = 0] \\
&= \mathbb{E}[\tau_i + f(g^{\mathbf{t}}) + \epsilon_i|G_{\mathcal{N}_i}^{\mathbf{T}} \simeq g^{\mathbf{t}}, T_i = 1] - \mathbb{E}[f(g^{\mathbf{t}}) + \epsilon_j|G_{\mathcal{N}_j}^{\mathbf{T}} \simeq g^{\mathbf{t}}, T_j = 0] \\
&= \alpha + f(g^{\mathbf{t}}) + \tau_i - \alpha - f(g^{\mathbf{t}}) \\
&= \tau_i.
\end{aligned} \tag{3.5}$$

The first equality follows from the definition of Y_i and Y_j , the second equality from the result in (3.3) and A3, and the third equality follows from independence of T and Y given by ignorability: $\mathbb{E}[\epsilon_i|T_i] = \mathbb{E}[\epsilon_i] = \alpha$ for all i . Finally, we can use both of the results above to obtain the ADE:

$$\begin{aligned}
& \frac{1}{n^t} \sum_{i=1}^n \mathbb{E}[T_i(\mathbb{E}[Y_i|G_{\mathcal{N}_i}^{\mathbf{T}} \simeq g^{\mathbf{t}}, T_i = 1] - \mathbb{E}[Y_j|G_{\mathcal{N}_j}^{\mathbf{T}} \simeq g^{\mathbf{t}}, T_j = 0])] \\
&= \frac{1}{n^t} \sum_{i=1}^n \mathbb{E}[T_i \tau_i] \tag{By (3.5)} \\
&= \frac{1}{n^t} \sum_{i=1}^n \mathbb{E}[T_i(\mathbb{E}[Y_i(1, \mathbf{0}) - Y_i(0, \mathbf{0})])] \tag{By (3.4)} \\
&= \frac{1}{n^t} \sum_{i=1}^n \Pr(T_i = 1)(\mathbb{E}[Y_i(1, \mathbf{0}) - Y_i(0, \mathbf{0})]) \\
&= \frac{1}{n^t} \sum_{i=1}^n \frac{n^t}{n} (\mathbb{E}[Y_i(1, \mathbf{0}) - Y_i(0, \mathbf{0})]) \\
&= \frac{1}{n} \sum_{i=1}^n \mathbb{E}[Y_i(1, \mathbf{0}) - Y_i(0, \mathbf{0})],
\end{aligned}$$

where $\Pr(T_i = 1) = \frac{n^t}{n}$ by assumption of complete randomization. \square

3.1 AME for Subgraph Matching

Given Proposition 1 and the framework established in the previous section, we would ideally like to match treated and control units that have isomorphic neighborhood graphs. This would allow us to better estimate the ADE without suffering interference bias: for a treated unit i , if a control unit j can be found such that $G_{\mathcal{N}_i}^{\mathbf{t}} \simeq G_{\mathcal{N}_j}^{\mathbf{t}}$, then j 's outcome will be identical in expectation to i 's counterfactual outcome and can be used as a proxy. Unfortunately, the number of non-isomorphic (canonically unique) graphs with a given number of nodes and edges grows incredibly quickly [Har94] and

finding such matches is infeasible for large graphs. We therefore resort to counting all subgraphs that appear in a unit’s neighborhood graph and matching units based on the counts of those subgraphs. However, instead of *exactly* matching on the counts of those subgraphs, we match treated and control units if they have *similar* counts, since matching exactly on all subgraph counts implies isomorphic neighborhoods and is also infeasible. Further, absolutely exact matches may not exist in real networks.

As we have previously shown, AME provides a framework for the above problem that is explicitly geared towards building interpretable, high-quality matches on discrete covariates, which in our setting are the counts of the treated subgraphs in the neighborhood. AME performs inexact matching while *learning* importance weights for each covariate from a training set, prioritizing matches on more important covariates. In this way, it neatly addresses the challenge of inexact matching by learning a metric specific to discrete covariates (namely, a weighted Hamming distance), which in our network interference setting, are vectors of subgraph counts. The vector \mathbf{w} denotes the importance of each subgraph in causing interference. We will leverage both information on outcomes and networks to construct an estimate for it.

The procedure we employ can be summarized as follows: We start by enumerating (up to isomorphism) all the p subgraphs g_1, \dots, g_p that appear in any of the $G_{\mathcal{N}_i}^{\mathbf{t}}, i \in 1, \dots, n$. The covariates for unit i are then given by $S(G_{\mathcal{N}_i}^{\mathbf{t}}) = (S_1(G_{\mathcal{N}_i}^{\mathbf{t}}), \dots, S_p(G_{\mathcal{N}_i}^{\mathbf{t}}))$ where $S_k(G_{\mathcal{N}_i}^{\mathbf{t}})$ denotes the number of times subgraph g_k appears in the subgraphs of $G_{\mathcal{N}_i}^{\mathbf{t}}$. These counts are then converted into binary indicators that are one if the count of subgraph g_k in each unit’s neighborhood is exactly x , for all x observed in the data. Thus, units will be matched exactly if they have identical subgraph counts. We then approximately solve the AME problem to find the optimally important set of subgraphs upon which to exactly match each treated unit, such that there is at least one control unit that matches exactly with the treated unit on the chosen subgraph counts. The key idea behind this approach is that we want to match units exactly on subgraph counts that contribute significantly to the interference function, trading off exactly-matching on these important subgraphs with potential mismatches on subgraphs that contribute less to interference.

3.2 A Bound on the Finite-Sample Bias of AME for Network Data

We study the expected error for one AME match on subgraphs under two assumptions: that the true weights for the AME objective (the weighted Hamming distance) are known, and that the candidate units for matching all have independently generated neighborhoods, and none of the units in these neighborhoods are being matched. Additional information on this setting is available in the proof.

Proposition 3. (*Oracle AME Error With Independent Neighborhoods*) Suppose that there are N independently generated graphs, each with n vertices, and all i.i.d. from

the same distribution over graphs: $\Pr(G_1 = g_1, \dots, G_N = g_N) = \prod_{i=1}^N p(g_i)$. Assume matches are only allowed between units in different graphs. Suppose additionally that $n^{(1)}$ randomly chosen units within each graph are assigned treatment, so that $\Pr(\mathbf{T}_i = \mathbf{t}_i) = \binom{n}{n^{(1)}}^{-1}$. Assume further that interference functions obey the following: $|f(g) - f(h)| \leq K \mathbf{w}^T \mathbb{I}[S(g) \neq S(h)]$, where \mathbf{w} is a vector of positive, real-valued, importance weights for the subgraphs counts, such that $\|\mathbf{w}\|_1 = M_n$ for some constant $0 < M_n < \infty$, and such that the condition above is satisfied for \mathbf{w} , and $\mathbb{I}[S(g) \neq S(h)]$ is the Hamming distance: a vector of the same length as \mathbf{w} that is 0 at position k if graphs g and h have the same count of subgraph k , and 1 otherwise. Assume that baseline responses have variance $\text{Var}(\epsilon_i) = \sigma^2 \forall i$. Then, for a treatment unit i , if j solves the AME problem, i.e., $j \in \arg \min_{\substack{k=1, \dots, n \\ T_k=0}} \mathbf{w}^T \mathbb{I}[S(h_{\mathcal{N}_i}^{\mathbf{t}}) \neq S(G_{\mathcal{N}_k}^{\mathbf{T}})]$, under A1-A3:

$$\begin{aligned} \mathbb{E} [|Y_i - Y_j - \tau_i| | T_i = 1, G_{\mathcal{N}_i}^{\mathbf{t}} = h_{\mathcal{N}_i}^{\mathbf{t}}] &\leq \sqrt{2}\sigma \\ &+ K \binom{n-1}{n^{(1)}}^{-1} \times \sum_{g \in \mathcal{G}_n} \sum_{\substack{\mathbf{t} \in \mathcal{T} \\ t_j=0}} \mathbf{w}^T \mathbb{I}[S(h_{\mathcal{N}_i}^{\mathbf{t}}) \neq S(g_{\mathcal{N}_j}^{\mathbf{t}})] p(g) \\ &\times \left(\frac{n^{(1)}}{n} + \frac{n - n^{(1)}}{n} C(g_{\mathcal{N}_j}^{\mathbf{t}}) \right)^{N-2}, \end{aligned}$$

where \mathcal{G}_n is the set of all graphs with n units, and $C(h_{\mathcal{N}_j}^{\mathbf{t}}) \leq 1$ for all g and \mathbf{t} .

A proof is available in the following section. The first element in the right hand side of the inequality is the standard deviation of the baseline responses. One summation is over all possible graphs with n units, and the other summation is over possible treatment assignments. The expression inside the summation is the product of three terms. First, the weighted Hamming distance between a graph and the target graph we are trying to match. Second, the probability of observing that graph. Third, an upper bound on the probability that unit j is among the minimizers of the weighted Hamming distance. Note that $\mathbf{w}^T \mathbb{I}[S(g_{\mathcal{N}_j}^{\mathbf{t}}) \neq S(h_{\mathcal{N}_i}^{\mathbf{t}})] p(g)$ is bounded for fixed n for all g and \mathbf{t} . This implies that the bound converges to $2\sqrt{\sigma}$ as $N \rightarrow \infty$, as long as the size of neighborhood graphs is held fixed, because perfect matching is possible with large amounts of data in this regime.

Before giving a proof of the statement, we briefly review some notation and assumptions to be used. For the purposes of theory, we study a simplified setting, in which we have to AME-match a unit i to one unit in a set of candidate units of size N such that: a) all the candidate units belong to disconnected graphs, which we refer to as candidate graphs. b) within each candidate graph there is only one pre-determined candidate unit c) candidate units have neighborhood graphs denoted by $G_{\mathcal{N}_j}$. d) all the candidate graphs are drawn independently from the same distribution over graphs: $\Pr(G_{\mathcal{N}_1} = g_1, \dots, G_{\mathcal{N}_N} = g_N) = \prod_{i=1}^N p(g_i)$. The support of p will be \mathcal{G}_n : the set of all graphs with exactly n units. We use $g_{\mathcal{N}_i}$ to denote the subgraph

induced over g by the units in the set of neighbors of unit i , $\mathcal{N}_i \subseteq V(g)$, i.e., $g_{\mathcal{N}_i}$ is the graph consisting only of the vertices that share an edge with i , in g , and of the edges in g that are between these vertices. The ego i is not included in $g_{\mathcal{N}_i}$.

Assigned treatments are denoted by \mathbf{T} , where $\mathbf{T} \in \{0, 1\}^n$, but in this setting treatment assignment is assumed to be independent within the N candidate graphs. Formally, the assumption we make is that $\Pr(\mathbf{T}_1 = \mathbf{t}_1, \dots, \mathbf{T}_N = \mathbf{t}_N) = \prod_{i=1}^N \binom{n}{n^{(1)}}^{-1}$, i.e., $n^{(1)}$ units are always treated uniformly at random within each of the N candidate graphs.

The direct treatment effect for any unit i is given by τ_i . We use $\mathbb{I}[S(g) \neq S(h)]$ to indicate the Hamming distance between subgraph counts of graphs g and h . This means that $\mathbb{I}[S(g) \neq S(h)]$ is a vector of size $|\mathcal{G}_n|$ that will be 1 in the ℓ^{th} entry if g and h have the same amount of occurrences of graph g_ℓ among their subgraphs. Note that this distance is coloring sensitive: two subgraphs that are isomorphic in shape but not labels will belong to different entries in this distance. The matched group of a *treated* unit i , denoted \mathbf{MG}_i is the set of all *control* units that match with i . In our setting $j \in \mathbf{MG}_i$ if it solves the AME problem, that is $j \in \arg \min_{\substack{k=1, \dots, n \\ T_k=0}} \mathbf{w}^T \mathbb{I}[S(G_{\mathcal{N}_k}^{\mathbf{T}_k}) \neq S(g_{\mathcal{N}_i}^{\mathbf{t}_i})]$.

Finally, we assume that both the graph for the unit we want to match and the treatment assignment for that unit's graph are fixed: \mathbf{t}_i is the treatment assignment in the graph of i , and $h_{\mathcal{N}_i}^{\mathbf{t}_i}$ is the neighborhood graph of i , where h denotes unit i 's graph. All other notation is as in the main paper.

Proof. We start by upper-bounding our quantity of interest as follows:

$$\begin{aligned} & \mathbb{E}[|Y_i - Y_j - \tau_i| | j \in \mathbf{MG}_i] \\ &= \mathbb{E}[|Y_i(1, \mathbf{t}_{i-i}) - Y_j(0, \mathbf{T}_{j-j}) - \tau_i| | j \in \mathbf{MG}_i] \\ &= \mathbb{E}[|\tau_i + f(h_{\mathcal{N}_i}^{\mathbf{t}_i}) + \epsilon_i - f(G_{\mathcal{N}_j}^{\mathbf{T}_j}) - \epsilon_j - \tau_i| | j \in \mathbf{MG}_i] \\ &\leq \mathbb{E}[|f(h_{\mathcal{N}_i}^{\mathbf{t}_i}) - f(G_{\mathcal{N}_j}^{\mathbf{T}_j})| | j \in \mathbf{MG}_i] + \mathbb{E}[|\epsilon_i - \epsilon_j| | j \in \mathbf{MG}_i] \\ &\leq K \mathbb{E}[\mathbf{w}^T \mathbb{I}[S(h_{\mathcal{N}_i}^{\mathbf{t}_i}) \neq S(G_{\mathcal{N}_j}^{\mathbf{T}_j})] | j \in \mathbf{MG}_i] + \mathbb{E}[|\epsilon_i - \epsilon_j| | j \in \mathbf{MG}_i], \end{aligned}$$

where the notation \mathbf{T}_{j-j} denotes the treatment indicator for candidate graph j for all units except j . The first equality follows from A1 since the event $j \in \mathbf{MG}_i$ implies that $T_j = 0$, as only control units are allowed in the matched groups. The second equality follows from Proposition 2. The first inequality is an application of the triangle inequality. The last line follows from the condition on the interference functions. Consider the second term. We can use the Cauchy-Schwarz inequality to construct a simple upper bound on it:

$$\begin{aligned} & \mathbb{E}[|\epsilon_i - \epsilon_j| | j \in \mathbf{MG}_i] = \mathbb{E}[|\epsilon_i - \epsilon_j|] \\ &\leq \sqrt{\mathbb{E}[(\epsilon_i - \epsilon_j)^2]} \\ &= \sqrt{\mathbb{E}[\epsilon_i^2] + \mathbb{E}[\epsilon_j^2] - 2\mathbb{E}[\epsilon_i]\mathbb{E}[\epsilon_j]} = \sqrt{2}\sigma \end{aligned}$$

where the last equality follows for the fact that the ϵ_i have mean 0 and are independent, with $Var(\epsilon_i) = \sigma^2$ for all i .

Consider now the term $\mathbb{E}[\mathbf{w}^T \mathbb{I}[S(h_{\mathcal{N}_i}^{\mathbf{t}_i}) \neq S(G_{\mathcal{N}_j}^{\mathbf{T}_j})] | j \in \mathbf{MG}_i]$. To upper-bound this, we write it out as follows using the definition of expectation:

$$\begin{aligned} & \mathbb{E}[\mathbf{w}^T \mathbb{I}[S(h_{\mathcal{N}_i}^{\mathbf{t}_i}) \neq S(G_{\mathcal{N}_j}^{\mathbf{T}_j})] | j \in \mathbf{MG}_i] \\ &= \sum_{g \in \mathcal{G}_n} \sum_{\mathbf{t} \in \mathcal{T}, t_j=0} \mathbf{w}^T \mathbb{I}[S(h_{\mathcal{N}_i}^{\mathbf{t}_i}) \neq S(g_{\mathcal{N}_j}^{\mathbf{t}_j})] \Pr(G_{\mathcal{N}_j}^{\mathbf{T}_j} = g_{\mathcal{N}_j}^{\mathbf{t}_j}, \mathbf{T}_j = \mathbf{t} | j \in \mathbf{MG}_i). \end{aligned}$$

We want to find an upper bound on $\Pr(G_{\mathcal{N}_j}^{\mathbf{T}_j} = g_{\mathcal{N}_j}^{\mathbf{t}_j}, \mathbf{T}_j = \mathbf{t} | j \in \mathbf{MG}_i)$. We start by writing this quantity out as a product of two probabilities:

$$\begin{aligned} & \Pr(G_{\mathcal{N}_j}^{\mathbf{T}_j} = g_{\mathcal{N}_j}^{\mathbf{t}_j}, \mathbf{T}_j = \mathbf{t} | j \in \mathbf{MG}_i) \\ &= \Pr(G_{\mathcal{N}_j}^{\mathbf{T}_j} = g_{\mathcal{N}_j}^{\mathbf{t}_j} | j \in \mathbf{MG}_i, \mathbf{T}_j = \mathbf{t}) \Pr(\mathbf{T}_j = \mathbf{t} | j \in \mathbf{MG}_i) \\ &= \Pr(G_{\mathcal{N}_j}^{\mathbf{T}_j} = g_{\mathcal{N}_j}^{\mathbf{t}_j} | j \in \mathbf{MG}_i, \mathbf{T}_j = \mathbf{t}) \binom{n-1}{n^{(1)}}^{-1}. \end{aligned}$$

Note that $\Pr(\mathbf{T}_j = \mathbf{t} | j \in \mathbf{MG}_i) = \binom{n-1}{n^{(1)}}^{-1}$ because treatment is uniformly randomized with $n^{(1)}$ units always treated in each candidate graph, but $T_j = 0$ conditionally on $j \in \mathbf{MG}_i$.

We use Bayes' rule to write out the first term in the final product as $\Pr(G_{\mathcal{N}_j}^{\mathbf{T}_j} = g_{\mathcal{N}_j}^{\mathbf{t}_j} | j \in \mathbf{MG}_i, \mathbf{T}_j = \mathbf{t}) = \frac{\Pr(j \in \mathbf{MG}_i | \mathbf{T}_j = \mathbf{t}, G_{\mathcal{N}_j}^{\mathbf{T}_j} = g_{\mathcal{N}_j}^{\mathbf{t}_j}) \Pr(G_{\mathcal{N}_j}^{\mathbf{T}_j} = g_{\mathcal{N}_j}^{\mathbf{t}_j} | \mathbf{T}_j = \mathbf{t})}{\Pr(j \in \mathbf{MG}_i | \mathbf{T}_j = \mathbf{t})}$.

By assumption, if all neighborhood graphs are empty, all units are used for all matched groups, and we are restricting ourselves to assignments in which $T_j = 0$, therefore, $\Pr(j \in \mathbf{MG}_i | \mathbf{T}_j = \mathbf{t}) = 1$. Second, by assumption $\Pr(G_{\mathcal{N}_j}^{\mathbf{T}_j} = g_{\mathcal{N}_j}^{\mathbf{t}_j} | \mathbf{T}_j = \mathbf{t}) = p(g)$. This is because treatment assignment is independent of the graph. We are left with having to find an expression for the likelihood, this can be written as:

$$\begin{aligned} & \Pr(j \in \mathbf{MG}_i | \mathbf{T}_j = \mathbf{t}, G_{\mathcal{N}_j}^{\mathbf{T}_j} = g_{\mathcal{N}_j}^{\mathbf{t}_j}) \\ &= \Pr(j \in \arg \min_{\substack{k=1, \dots, N, \\ k \neq i}} \mathbf{w}^T \mathbb{I}[S(h_{\mathcal{N}_i}^{\mathbf{t}_i}) \neq S(G_{\mathcal{N}_k}^{\mathbf{T}_k})] | \mathbf{T}_j = \mathbf{t}, G_{\mathcal{N}_j}^{\mathbf{T}_j} = g_{\mathcal{N}_j}^{\mathbf{t}_j}) \\ &= \prod_{\substack{k=1 \\ k \neq i, j}}^N \left[\Pr(\mathbf{w}^T \mathbb{I}[S(h_{\mathcal{N}_i}^{\mathbf{t}_i}) \neq S(G_{\mathcal{N}_k}^{\mathbf{T}_k})] \geq \mathbf{w}^T \mathbb{I}[S(h_{\mathcal{N}_i}^{\mathbf{t}_i}) \neq S(g_{\mathcal{N}_j}^{\mathbf{t}_j})] | T_k = 0) \Pr(T_k = 0) \right. \\ & \quad \left. + \Pr(T_k = 1) \right] \\ &= \prod_{\substack{k=1 \\ k \neq i, j}}^N \left[\Pr(\mathbf{w}^T \mathbb{I}[S(h_{\mathcal{N}_i}^{\mathbf{t}_i}) \neq S(G_{\mathcal{N}_k}^{\mathbf{T}_k})] \geq \mathbf{w}^T \mathbb{I}[S(h_{\mathcal{N}_i}^{\mathbf{t}_i}) \neq S(g_{\mathcal{N}_j}^{\mathbf{t}_j})] | T_k = 0) \frac{n - n^{(1)}}{n} + \frac{n^{(1)}}{n} \right]. \end{aligned}$$

The second equality follows because k can never be in the matched group of unit i if $T_k = 1$, and, if $T_k = 0$, then k must be one of the minimizers of the weighted Hamming distance between neighborhood subgraph counts. The probability is a product of densities because of independence of candidate subgraphs. For an arbitrary unit, k , we define the following compact notation for the probability that k 's weighted Hamming distance from i is larger than the weighted Hamming distance from j to i :

$$\begin{aligned} \Pr(\mathbf{w}^T \mathbb{I}[S(h_{\mathcal{N}_i}^{\mathbf{t}_i}) \neq S(G_{\mathcal{N}_k}^{\mathbf{T}_k})] \geq \mathbf{w}^T \mathbb{I}[S(h_{\mathcal{N}_i}^{\mathbf{t}_i}) \neq S(g_{\mathcal{N}_j}^{\mathbf{t}})] | T_k = 0) \\ =: C_k(g_{\mathcal{N}_j}^{\mathbf{t}}) \leq 1. \end{aligned}$$

Note that the last inequality follows from the fact that the expression above is a probability. Since graphs and treatment assignments, G_k and \mathbf{T}_k are the only random variables in the probability denoted by $C_k(g_{\mathcal{N}_j}^{\mathbf{t}})$, and since they are all independent, and identically distributed, we can say that $C_1(g_{\mathcal{N}_j}^{\mathbf{t}}) = C_2(g_{\mathcal{N}_j}^{\mathbf{t}}) = \dots = C_N(g_{\mathcal{N}_j}^{\mathbf{t}}) = C(g_{\mathcal{N}_j}^{\mathbf{t}})$. Because of this we have:

$$\Pr(j \in \mathbf{MG}_i | G_{\mathcal{N}_j}^{\mathbf{T}_j} = g_{\mathcal{N}_j}^{\mathbf{t}}, \mathbf{T}_j = \mathbf{t}) = \left(\frac{n^{(1)}}{n} + \frac{n - n^{(1)}}{n} C(g_{\mathcal{N}_j}^{\mathbf{t}}) \right)^{N-2}.$$

Putting all the elements we have together we get the expression for the first term in the bound:

$$\begin{aligned} & \mathbb{E}[\mathbf{w}^T \mathbb{I}[S(h_{\mathcal{N}_i}^{\mathbf{t}_i}) \neq S(G_{\mathcal{N}_j}^{\mathbf{T}_j})] | j \in \mathbf{MG}_i] \\ &= \sum_{g \in \mathcal{G}_n} \sum_{\mathbf{t} \in \mathcal{T}: t_j=0} \mathbf{w}^T \mathbb{I}[S(g_{\mathcal{N}_j}^{\mathbf{t}}) \neq S(h_{\mathcal{N}_i}^{\mathbf{t}_i})] \times \Pr(G_{\mathcal{N}_j}^{\mathbf{T}_j} = g_{\mathcal{N}_j}^{\mathbf{t}}, \mathbf{T}_{\mathcal{N}_j} = \mathbf{t}_{\mathcal{N}_j} | j \in \mathbf{MG}_i) \\ &= \sum_{g \in \mathcal{G}_N} \sum_{\mathbf{t} \in \mathcal{T}: t_j=0} \mathbf{w}^T \mathbb{I}[S(g_{\mathcal{N}_j}^{\mathbf{t}}) \neq S(h_{\mathcal{N}_i}^{\mathbf{t}_i})] \binom{n-1}{n^{(1)}}^{-1} p(g) \left(\frac{n^{(1)}}{n} + \frac{n - n^{(1)}}{n} C(g_{\mathcal{N}_j}^{\mathbf{t}}) \right)^{N-2}. \end{aligned}$$

□

3.3 Asymptotic Behavior

Here we expand on the asymptotic consequences of Proposition 3: note first, that, by assumption $\|\mathbf{w}\|_1 = M_n$, and that, therefore $\mathbf{w}^T \mathbb{I}[S(g) \neq S(h)] \leq M_n$ for any graphs $g, h \in \mathcal{G}_n$. That is to say, the weighted Hamming distance between any two graphs with n units will be upper-bounded by the sum of the weights. Recall also that $C(g_{\mathcal{N}_j}^{\mathbf{t}}) \leq 1$ for all g and \mathbf{t} as this quantity is a probability, and let $C_{max} = \max_{\substack{g \in \mathcal{G}_n \\ \mathbf{t} \in \mathcal{T}}} C(g_{\mathcal{N}_j}^{\mathbf{t}})$.

We can combine all these bounds with the upper bound in Proposition 3 to write:

$$\begin{aligned}
& \mathbb{E}[\mathbf{w}^T \mathbb{I}[S(h_{\mathcal{N}_i}^{\mathbf{t}_i}) \neq S(G_{\mathcal{N}_j}^{\mathbf{T}_j})] | j \in \mathbf{MG}_i] \\
&= \sum_{g \in \mathcal{G}_{\mathcal{N}}} \sum_{\mathbf{t} \in \mathcal{T}, t_j=0} \mathbf{w}^T \mathbb{I}[S(g_{\mathcal{N}_j}^{\mathbf{t}}) \neq S(h_{\mathcal{N}_i}^{\mathbf{t}_i})] \binom{n-1}{n^{(1)}}^{-1} p(g) \left(\frac{n^{(1)}}{n} + \frac{n-n^{(1)}}{n} C(g_{\mathcal{N}_j}^{\mathbf{t}}) \right)^{N-2} \\
&\leq M_n \left(\frac{n^{(1)}}{n} + \frac{n-n^{(1)}}{n} C_{max} \right)^{N-2} \sum_{g \in \mathcal{G}_n} \sum_{\mathbf{t} \in \mathcal{T}, t_j=0} \binom{n-1}{n^{(1)}}^{-1} p(g) \\
&= M_n \left(\frac{n^{(1)}}{n} + \frac{n-n^{(1)}}{n} C_{max} \right)^{N-2}.
\end{aligned}$$

The first equality follows from Proposition 3, the first inequality from the bounds previously discussed, and the second equality follows from the fact that the sum in the second to last line is a sum of probability distributions over their entire domain, and therefore is equal to 1. Under the condition that n , the number of units in each unit's candidate graph, stays fixed, and that $C_{max} < 1$, then, as $N \rightarrow \infty$, we have $M_n \left(\frac{n^{(1)}}{n} + \frac{n-n^{(1)}}{n} C_{max} \right)^{N-2} \rightarrow 0$, because the quantity inside the parentheses is always less than 1. This makes sense, because asymptotically, matches can be made exactly; i.e., units matched in the way described in our theoretical setting have isomorphic neighborhood subgraphs asymptotically. This also has a consequence that the bound in Proposition 3 converges to $\sqrt{2}\sigma$ asymptotically in N . This is the variance of the baseline errors and can be lowered by matching the same unit with multiple others. As noted before, for this argument to apply, candidate graphs must remain of fixed size n as they grow in number, so that the quantity M_n remains constant: this setting is common in cluster-randomized trials where a growing number of units is clustered into fixed-size clusters of size at most n . The asymptotic behavior of our proposed methodology is less clear in settings in which the analyst cannot perform such clustering before randomization and n is allowed to grow with N , and is an avenue for potential future research.

3.4 Heteroskedasticity in The Baseline Effects

In a network setting such as the one we study, it is possible that baseline effects of units do not have equal variance. Here we discuss how this setting affects our result in Proposition 3. Here, we maintain that $\mathbb{E}[\epsilon_i] = \alpha$ for all i , but we assume that $Var(\epsilon_i) = \sigma_i^2$, and that $Cov(\epsilon_i, \epsilon_j) \neq 0$. Starting from the upper bound on the estimation error given in the proof of Proposition 3, we can see that the baseline effects only come in in the term: $\mathbb{E}[|\epsilon_i - \epsilon_j| | j \in \mathbf{MG}_i]$, we therefore focus our attention on this term, as the rest of this bound does not change when the variance of these terms changes. Note first, that $\mathbb{E}[|\epsilon_i - \epsilon_j| | j \in \mathbf{MG}_i] = \mathbb{E}[|\epsilon_i - \epsilon_j|]$ as the event $j \in \mathbf{MG}_i$ is

independent of the baseline effects. We can now apply the Cauchy-Schwarz inequality, in the same way as we do in the proof of Proposition 3, to obtain:

$$\begin{aligned}
\mathbb{E}[|\epsilon_i - \epsilon_j|] &\leq \sqrt{\mathbb{E}[(\epsilon_i - \epsilon_j)^2]} \\
&= \sqrt{\mathbb{E}[\epsilon_i^2] + \mathbb{E}[\epsilon_j^2] - 2\mathbb{E}[\epsilon_i]\mathbb{E}[\epsilon_j]} \\
&= \sqrt{\sigma_i^2 + \alpha^2 + \sigma_j^2 + \alpha^2 - 2\alpha^2} \\
&= \sqrt{\sigma_i^2 + \sigma_j^2}.
\end{aligned}$$

Clearly, this is not too different from the homoskedastic setting we study in the proposition: as long as neither of the unit variances is too large for inference, results in the heteroskedastic setting will suffer from similar bias as they would under independent baseline effects with equal variance.

Chapter 4

Conclusions

Quantification of statistical bias and variance is of fundamental importance in all statistical inference settings, but even more so in causal inference. This is because causal estimates are often used in high-stakes decision making, and statistical uncertainty and error must be known or at least approximated in order to make fully informed decisions. Almost-Matching Exactly methods permit decision-makers to take advantage of causal estimates that are interpretable: similar cases can be pinpointed as the reason why a certain treatment effect was estimated. This allows decision-makers to fully justify the data-based conclusion that they are used to inform their decisions. It follows that being able to quantify statistical uncertainty for AME methods is extremely important: a measure of uncertainty must be reported alongside these interpretable estimates in order to make them really useful and trustworthy to high-stakes decision makers.

In this work we have offered such a quantification for two methods of the AME family: AME for discrete data, and AME-Networks, which applies discrete AME to the problem of matching with network interference. Using theoretical arguments, we have shown that the finite-sample bias of both methodologies can be upper-bounded as a function of the quality of both the data and the matches made. We have additionally shown that these finite-sample bounds have asymptotic consequences: asymptotic normality in the case of discrete AME, and consistency in the case of AME-Networks on disconnected graphs. Our results inform the construction of approximate confidence intervals around AME causal estimates, and can be used to both report statistical uncertainty around such estimates, and to inform empirical investigations into the variance of AME in specific settings.

Our work here suggests several avenues for future research: for example, the extension of our analysis to AME methods for continuous covariates, such as [MOR⁺20], and [PRV18], or the development of algorithms for network settings that do not depend on isolated graphs for consistency. The development of causal inference methods that are interpretable and accurate is of paramount importance for better decision-making, and we hope to further this research agenda with this and future projects.

Bibliography

- [AEI18] Susan Athey, Dean Eckles, and Guido W Imbens. Exact p-values for network interference. *Journal of the American Statistical Association*, 113(521):230–240, 2018.
- [AMO⁺20] M. Usaid Awan, Marco Morucci, Vittorio Orlandi, Sudeepa Roy, Cynthia Rudin, and Alexander Volfovsky. Almost-matching-exactly for treatment effect estimation under network interference. *Proceedings of The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS*, 2020.
- [DLR⁺19] Awa Dieng, Yameng Liu, Sudeepa Roy, Cynthia Rudin, and Alexander Volfovsky. Interpretable Almost-Exact Matching for Causal Inference. In *AISTATS*, pages 2445–2453, 2019.
- [EKB16] Dean Eckles, René F Kizilcec, and Eytan Bakshy. Estimating peer effects in networks with peer encouragement designs. *Proceedings of the National Academy of Sciences*, 113(27):7316–7322, 2016.
- [Har94] Frank Harary. *Graph Theory*. Addison-Wesley, 1994.
- [Hol86] Paul W Holland. Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960, 1986.
- [HS95] M Elizabeth Halloran and Claudio J Struchiner. Causal inference in infectious diseases. *Epidemiology (Cambridge, Mass.)*, 6(2):142–151, 1995.
- [LMZ18] Fan Li, Kari Lock Morgan, and Alan M Zaslavsky. Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, 113(521):390–400, 2018.
- [Man13] Charles F Manski. Identification of treatment response with social interactions. *The Econometrics Journal*, 16(1):S1–S23, 2013.
- [MOR⁺20] Marco Morucci, Vittorio Orlandi, Sudeepa Roy, Cynthia Rudin, and Alexander Volfovsky. Adaptive hyper-box matching for interpretable individualized treatment effect estimation. *Proceedings of the Thirty-Sixth Conference on Uncertainty in Artificial Intelligence, UAI*, 2020.
- [PRV18] Harsh Parikh, Cynthia Rudin, and Alexander Volfovsky. MALTS: Matching After Learning to Stretch. *arXiv preprint arXiv:1811.07415*, 2018.

- [Rub80] Donald B Rubin. Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American Statistical Association*, 75(371):591–593, 1980.
- [SA17] Daniel L Sussman and Edoardo M Airolidi. Elements of estimation theory for causal effects in the presence of network interference. *arXiv preprint arXiv:1702.03578*, 2017.
- [TK13] Panos Toulis and Edward Kao. Estimation of causal peer influence effects. In *International Conference on Machine Learning (ICML)*, pages 1489–1497, 2013.
- [WMA⁺17] Tianyu Wang, Marco Morucci, M Awan, Yameng Liu, Sudeepa Roy, Cynthia Rudin, and Alexander Volfovsky. FLAME: A Fast Large-Scale Almost Matching Exactly Approach to Causal Inference. *arXiv preprint arXiv:1707.06315*, 2017.