European Journal of Operational Research xxx (xxxx) xxx

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Discrete Optimization

Network flow methods for the minimum covariate imbalance problem

Dorit S. Hochbaum [a,1], Xu Rao [a], Jason Sauppe [b,2,*]

[a] Department of Industrial Engineering and Operations Research, University of California, Berkeley, United States
[b] Department of Computer Science, University of Wisconsin, La Crosse, United States

ARTICLE INFO

Article history:
Received 7 February 2021
Accepted 21 October 2021
Available online xxx

Keywords:
Networks
Causal inference
Covariate imbalance

ABSTRACT

In an observational study, one is given disjoint samples of treatment units and control (untreated) units, and the goal is to compare outcomes between the two samples in order to estimate a treatment effect. A complication is that the treatment and control units often differ on important pre-treatment attributes, and these differences, referred to as covariate imbalance, can bias the estimate. One method to correct for covariate imbalance is to select a subset of the control sample that has minimum imbalance with respect to the treatment sample, and then use this control subset for estimating the treatment effect. While this optimization problem is NP-hard in general, certain special cases can be solved efficiently. Specifically, the variant of this optimization problem with one covariate is easy to solve, the variant with three or more covariates is NP-hard, and the variant with two covariates is solvable in polynomial time. We present several network flow formulations for the problem of minimizing imbalance on two nominal covariates. First, we present a minimum cost network flow formulation for solving the problem with the constraint that the control subset must have the same size as the treatment sample. We then derive an improved maximum flow formulation. For alternate size restrictions on the control subset, we use a proportional imbalance objective which leads to non-integral supplies and demands in the preceding network flow formulations. We then derive an alternate minimum cost network flow formulation that ensures integrality and solves the proportional imbalance problem in polynomial time.

1. Introduction

Researchers in many fields are confronted with the problem of identifying causal connections between actions (treatments) and outcomes. In an observational study, the researcher observes a sample of units that received treatment (called the *treatment sample*) along with another (typically larger) sample of units that did not (called the *control sample*). The researcher generally does not have direct control over the treatment allocation process, which makes it difficult to attribute differences in outcomes between the two samples to the treatment itself. Despite this difficulty, the ease of access to ever-increasing quantities of observational data make such studies popular across a wide range of disciplines, including social sciences and medicine (Rosenbaum, Ross, & Silber, 2007; Yang, Small, Silber, & Rosenbaum, 2012; Zubizarreta, 2012). Additionally, observational studies can often be performed more quickly than rigorous experimental studies, which require significant time and effort to set up and conduct. This allows researchers to address important time-sensitive questions where it is desirable to have a good answer now instead of waiting for a perfect answer later. The COVID-19 pandemic is one such situation; see Kim & Eisen (2020) for additional discussion.

Many methods have been developed for estimating treatment effects in observational studies. A common theme across these methods is that they attempt to adjust for differences in important pre-treatment attributes called covariates that may confound the treatment effect estimate. Examples of covariates in medical data include age, height, weight, blood pressure, disease history, and/or genetic information. Differences on these covariates between the treatment and control samples are referred to as covariate imbalance. There are many ways to measure imbalance based on the type of the covariates. For example, for a continuous covariate, the difference in mean value between treatment and control samples is referred to as mean imbalance. For a nominal covariate with a discrete set of values, or levels, the difference in number or proportion of treatment and control units at each level, summed across all levels, provides another imbalance measure. We focus here on a particular goal for adjustment with nominal covariates called the *min-imbalance* problem. The min-imbalance problem is trivial for a single covariate and NP-hard for three or more covariates, as discussed next. For the case of two covariates, we introduce and an-

* Corresponding author.
E-mail address: jsauppe@uwlax.edu (J. Sauppe).
[1] Research supported in part by NSF award no. CMMI-1760102.
[2] Research supported in part by DoD (32 CFR 168a) and NSF (DGE-1144245).

https://doi.org/10.1016/j.ejor.2021.10.041
0377-2217/© 2021 Elsevier B.V. All rights reserved.

alyze several network flow algorithms that solve the problem in polynomial time.

Let the sizes of the treatment and control samples be n and n', respectively, and let P be the number of covariates to be balanced. Each covariate p partitions the treatment and control samples into k_p levels each. Let the levels of the treatment sample under covariate $p = 1, \ldots, P$ be $L_{p,1}, L_{p,2}, \ldots, L_{p,k_p}$ of sizes $\ell_{p,1}, \ell_{p,2}, \ldots, \ell_{p,k_p}$, and let the levels of the control sample under covariate p be $L'_{p,1}, L'_{p,2}, \ldots, L'_{p,k_p}$ of sizes $\ell'_{p,1}, \ell'_{p,2}, \ldots, \ell'_{p,k_p}$. The min-imbalance problem is to find a subset of the control sample of size n, called the *selection* and denoted by S, such that the numbers of units at each level in the treatment sample and the selection are as close as possible:

$$\text{(min-imbalance)} \quad \min \sum_{p=1}^{P} \sum_{i=1}^{k_p} \left| \left| S \cap L'_{p,i} \right| - \ell_{p,i} \right| \quad \text{s.t. } |S| = n.$$

In the case of a single covariate, $P = 1$, the solution to the min-imbalance problem is trivial: for any level i, if $\ell_{1,i} \le \ell'_{1,i}$, then the optimal selection takes any $\ell_{1,i}$ control units from level i; otherwise, the optimal selection takes all $\ell'_{1,i}$ control units from level i. After this, take enough remaining control units from any level to reach a selection size of n. Let $\ell''_{1,i}$ denote the number of control units in the selection at level i. Then the value of the objective function corresponding to this selection is $\sum_{i=1}^{k_1} \left| \ell''_{1,i} - \ell_{1,i} \right|$, which is the optimal value for the single covariate min-imbalance problem.

For the min-imbalance problem with multiple covariates, Bennett, Vielma, & Zubizarreta (2020) presented a mixed integer programming (MIP) formulation and showed that the corresponding linear programming (LP) relaxation yields an integer solution when $P \le 2$. For $P = 3$ they presented an example where the LP relaxation's solution is not integral and noted that the problem is NP-hard (proofs can be found in Hochbaum & Rao, 2020; Sauppe, 2015, and Appendix A).

1.1. Additional imbalance problems

The min-imbalance problem with one covariate occurs in the context of the *near-fine balance* matching procedure of Yang et al. (2012). In near-fine balance, the goal is to find an optimal matching of treatment units to control units subject to the constraint that imbalance on a nominal covariate is as small as possible (see Section 1.3 for more details). This problem can be solved using a two-stage process that first determines the minimum imbalance and then seeks an optimal matching (based on a distance measure between treatment and control units) that meets the imbalance requirement.

We consider a general formulation of the near-fine balance problem involving multiple covariates where each treatment unit needs to be matched to κ control units, where κ is an integer that satisfies $1 \le \kappa \le \frac{n'}{n}$. We refer to this problem here as *κ-Matching-Balance* (MB). In the first stage of MB, the goal is to find a selection S of size κn, that solves the *min-κ-imbalance* problem defined as:

$$\text{(min-}\kappa\text{-imbalance)} \quad \min \sum_{p=1}^{P} \sum_{i=1}^{k_p} \left| \left| S \cap L'_{p,i} \right| - \kappa \ell_{p,i} \right| \quad \text{s.t. } |S| = \kappa n.$$

In the second stage of MB, among all selections of control units that attain the minimum κ-imbalance, one chooses the selection that minimizes the distances from matching each treatment unit to exactly κ selected control units. Yang et al. (2012) studied the MB problem for a single covariate and proposed two network flow algorithms. There is no prior work for the MB problem with two or more covariates, even in the first stage.

The min-imbalance problem can also be extended to situations in which the selection size q need not equal the size of the treatment sample. This gives rise to the *min-proportional imbalance* problem defined as:

$$\text{(min-proportional imbalance)}$$
$$\min \sum_{p=1}^{P} \sum_{i=1}^{k_p} \left| \frac{\left| S \cap L'_{p,i} \right|}{q} - \frac{\ell_{p,i}}{n} \right| \quad \text{s.t. } |S| = q.$$

With $q = n$, min-proportional imbalance is equivalent to min-imbalance with objective scaled by $1/n$; with $q = \kappa n$ for some integer κ, min-proportional imbalance is equivalent to min-κ-imbalance with objective scaled by $1/(\kappa n)$.

1.2. Contributions of paper

Our main results here are efficient algorithms for the min-imbalance, min-κ-imbalance, and min-proportional imbalance problems with two covariates.

For the min-imbalance problem with two covariates, we present an integer programming formulation related to that in Bennett et al. (2020) and show that the constraint matrix is totally unimodular. This implies that the linear programming relaxation to the problem has integer extreme points, and in particular that its optimal solution is integral. We then show that the min-imbalance problem with two covariates can be solved with specialized graph algorithms for network flow problems, which is more efficient than solving via linear programming. Specifically, we show how to formulate this problem as a minimum cost network flow problem and solve it in $O(n \cdot (n' + n \log n))$ steps. We also provide a more efficient maximum flow formulation that can be solved in $O(n'^{3/2} \log^2 n)$ steps. These two methods can also be applied to the min-imbalance problem with a selection size q with $q \ne n$ (see Appendix B).

For the min-proportional imbalance problem with two covariates and a selection size q with $q \ne n$, we show a minimum cost network flow formulation that can be solved in $O(q \cdot (n' + n \log n))$ steps (an alternate minimum cost network flow formulation was provided in the unpublished thesis of Sauppe (2015) in the context of the BOSS framework discussed in Section 1.3).

We also show (in Appendix C) that the min-κ-imbalance problem is equivalent to the min-imbalance problem (for any number of covariates), and therefore an optimal solution to a corresponding min-imbalance problem provides an optimal solution to the min-κ-imbalance problem. As such, the first stage of the MB problem with two covariates can be solved with any of the algorithms shown here. We also observe that, for any number of covariates, if the optimal solutions to the first stage problem of minimizing the κ-imbalance have a particular form (see Appendix C for details), then an optimal solution to the second stage, and therefore to the MB problem itself, can be obtained by solving a network flow problem. This implies that, under certain conditions, the MB problem with two covariates can be solved efficiently with network flow techniques.

Lastly, we provide a proof (in Appendix A), similar to the proof in Sauppe (2015) and presented independently in the arXiv paper of Hochbaum & Rao (2020), that the min-imbalance problem is NP-hard for three or more covariates.

Demonstrating how to solve these problems efficiently with two covariates sheds important light on their structure. Additionally, knowing that the problems are NP-hard with 3 or more covariates justifies the use of implicit enumeration techniques and/or heuristics for solving them in practice. It also opens up directions for future work in solving these NP-hard problems by using two-covariate subproblems in various ways (e.g., for bounds in branch-and-bound).

1.3. Related work

Matching methods (Stuart, 2010) have been widely used to estimate treatment effects within observational studies. These methods attempt to reduce covariate imbalance by pairing each unit in the treatment sample with a similar unit in the control sample, where similarity is defined using a function of the units' covariate values. The treatment effect is then estimated as the average difference in outcomes across all matched pairs, where unmatched units from the control sample are ignored. If the matched pairs have identical values for all covariates that impact the outcomes (i.e., they are exactly matched), then the associated treatment effect estimate will be unbiased.

In practice, exactly matched pairs often do not exist in the available data. Rosenbaum (1989) and Rosenbaum et al. (2007) developed a concept known as *fine balance*. In fine balance, the distances between units are computed using all but one nominal covariate. The goal is to find a minimum-distance matching subject to the constraint that the number of matched control units equals the number (or proportion) of treatment units at each level of the nominal covariate. Rosenbaum (1989) gave a minimum cost network flow formulation for the fine balance problem; we show in Appendix C that fine balance is a special case of κ-matching balance so it can also be solved with the method presented there.

It is not always feasible to satisfy the fine balance requirement. Several papers considered the goal of minimizing the violation of this requirement, which we refer to as *(covariate) imbalance*. Yang et al. (2012) proposed *near-fine balance*, which considers finding an optimal matching between the treatment sample and a subset of the control sample where the subset minimizes an imbalance measure on a single nominal covariate. Zubizarreta (2012) extended *near-fine balance* to multiple covariates which could be nominal or quantitative.

Nikolaev, Jacobson, Cho, Sauppe, & Sewell (2013) sought to address the difficulties of finding exactly matched pairs by eliminating matching entirely and instead focusing on covariate balance as the measure of quality. In the Balance Optimization Subset Selection (BOSS) framework, the goal is to identify a subset of the control sample that minimizes a covariate imbalance measure with respect to the treatment sample. The choice of imbalance measure used by BOSS is flexible and can accommodate nominal and quantitative covariates as well as covariate interactions. The min-imbalance and min-proportional imbalance problems are special cases of BOSS; Bennett et al. (2020) examined the min-imbalance problem using integer and linear programming. Computational experiments by Nikolaev et al. (2013), Sauppe, Jacobson, & Sewell (2014), and Sauppe & Jacobson (2017) show the value of using covariate balance as a primary objective in observational studies.

Sauppe et al. (2014) combined covariate balance and matching into the *balanced matching problem* which seeks a matching that pairs each treatment unit with one or more control units and minimizes an objective function consisting of a linear combination of the matched pair distances and a covariate imbalance measure. For a particular imbalance measure defined on nominal covariates, Sauppe et al. provided a mixed integer programming (MIP) model and proved that the problem can be solved in polynomial time through network flow techniques if balance is sought on a single covariate. In addition, they proved that the problem is NP-hard if balance is sought on two or more covariates. The special case of balanced matching where the objective uses only the imbalance component and omits the matching distances reduces to BOSS; Sauppe et al. noted that the resulting problem is solvable in polynomial time for two covariates but NP-hard for three or more covariates. These remarks applied only for selection sizes that are integer multiple of n. The unpublished thesis of Sauppe (2015) provides proofs of these results for any selection size.

1.4. Overview of paper

Section 2 provides the notation used here. In Section 3 we analyze an integer programming formulation of the min-imbalance problem with two covariates. We then present and analyze a minimum cost network flow formulation for this problem in Section 4 and a more efficient maximum-flow formulation in Section 5. In Section 6 we show how to solve the min-imbalance and min-proportional imbalance problems with two covariates and a selection size q with $q \neq n$. Concluding remarks are presented in Section 7, and supplemental results are provided in the appendices.

2. Preliminaries and notation

The goal of the min-imbalance problem is to identify a subset of the control sample, called the selection, such that the numbers of treatment units and selected control units in each level of each covariate are as close as possible. For a selection S of control units we define the *discrepancy* at level i under covariate p as $dis(S, p, i) = \left|S \cap L'_{p,i}\right| - \ell_{p,i}$. The discrepancy of a level can be positive or negative. If the discrepancy is positive we refer to it as *excess* which is defined as $e_{p,i}(S) = \max\{0, dis(S, p, i)\}$, and if negative, we refer to it as *deficit* $d_{p,i}(S) = \max\{0, -dis(S, p, i)\}$. With this notation the imbalance of a selection S is $IM(S) = \sum_{p=1}^{P}\sum_{i=1}^{k_p}\left(e_{p,i}(S) + d_{p,i}(S)\right)$, which is identical to $\sum_{p=1}^{P}\sum_{i=1}^{k_p}||S \cap L'_{p,i}| - \ell_{p,i}|$. For this imbalance measure, if any covariate p has two levels i and j with $i \neq j$ and $\ell_{p,i} = \ell_{p,j} = 0$, then we can merge levels i and j without impacting imbalance. As such, we assume that the number of levels k_p for any covariate p is at most $n+1$.

We now present the integer programming formulation that was given by Bennett et al. (2020) for the min-imbalance problem. That integer program involves two sets of decision variables: for each $j = 1, \ldots, n'$, the binary variable z_j is equal to 1 if control unit j is in the selection S, and 0 otherwise; and for each $p = 1, \ldots, P$, and $i = 1, \ldots, k_p$, the variable $y_{p,i} = |dis(S, p, i)| = ||S \cap L'_{p,i}| - \ell_{p,i}|$ represents the absolute value of the discrepancy at level i under covariate p. With these variables the formulation is:

$$\min \quad \sum_{p=1}^{P}\sum_{i=1}^{k_p} y_{p,i} \tag{1a}$$

$$\text{s.t.} \quad \sum_{j \in L'_{p,i}} z_j - \ell_{p,i} \leq y_{p,i} \qquad p = 1, \ldots, P, \quad i = 1, \ldots, k_p \tag{1b}$$

$$\ell_{p,i} - \sum_{j \in L'_{p,i}} z_j \leq y_{p,i} \qquad p = 1, \ldots, P, \quad i = 1, \ldots, k_p \tag{1c}$$

$$\sum_{j=1}^{n'} z_j = n \tag{1d}$$

$$z_j \in \{0, 1\} \qquad j = 1, \ldots, n'. \tag{1e}$$

For each pair p, i with $p = 1, \ldots, P$ and $i = 1, \ldots, k_p$, constraints (1b) and (1c) ensure that $y_{p,i}$ assumes the absolute value of the difference between the number of selected level i control units and $\ell_{p,i}$ at an optimal solution. These constraints also ensure that any feasible $y_{p,i}$ is non-negative and therefore a non-negativity constraint is not required for variable $y_{p,i}$. Constraint (1d) specifies that the size of the selected subset equals the size of the treatment sample.

Bennett et al. (2020) proved that any basic solution of the linear programming relaxation of (1) is integral for $P = 2$. We provide in Section 3 a stronger result showing that a slightly modified form of formulation (1) has a constraint matrix which is totally unimodular for $P = 2$. This implies that every basic solution is integer, and furthermore the constraint matrix is that of a minimum cost network flow problem. As such, the problem can be solved more efficiently with specialized graph algorithms instead of through linear programming.

An optimal solution to formulation (1) specifies for each control unit whether or not it is in the selection. We observe however that the output to the min-imbalance problem, for any number of covariates, can be presented more compactly in terms of *level-intersections*. Let $K = \prod_{p=1}^{P}\{1, 2, \ldots, k_p\}$. For any $(i_1, i_2, \ldots, i_P) \in K$, define the *level-intersection* $L'_{i_1,i_2,\ldots,i_P} = L'_{1,i_1} \cap L'_{2,i_2} \cap \ldots \cap L'_{P,i_P}$. The collection of level-intersections forms a partition of the control sample. The number of non-empty level-intersections is at most $\min\left\{n', \prod_{p=1}^{P} k_p\right\}$. So instead of specifying which control units belong to the selection, it is sufficient to determine the number of selected control units in each level-intersection because the identity of the particular selected control units has no effect on the imbalance. This allows for reformulating (1) with variables $x_{i_1,i_2,\ldots,i_P}$ representing the level-intersection sizes for each $(i_1, i_2, \ldots, i_P) \in K$. To derive a selection given the level-intersection sizes, one selects any $x_{i_1,i_2,\ldots,i_P}$ control units from each level-intersection $L'_{i_1,i_2,\ldots,i_P}$, for all $(i_1, i_2, \ldots, i_P) \in K$. This idea will be revisited in Section 4 and also in Appendix C.

3. A modified formulation with a totally unimodular constraint matrix for $P = 2$

In this section we present an alternate integer programming formulation for the min-imbalance problem with two covariates and show that its constraint matrix is totally unimodular. In this formulation, instead of using variables $y_{p,i}$, we use variables for excess and deficit. As discussed in Section 2, $\left|\left|S \cap L'_{p,i}\right| - \ell_{p,i}\right| = e_{p,i}(S) + d_{p,i}(S)$ for each p and i. We let the variable for excess for p and i be $e_{p,i}$ and the variable for deficit be $d_{p,i}$. Note that $y_{p,i} = e_{p,i} + d_{p,i}$ where both $e_{p,i}$ and $d_{p,i}$ are non-negative variables. Additionally, for each p and i, $\left|S \cap L'_{p,i}\right| - \ell_{p,i} = e_{p,i} - d_{p,i}$ if and only if $\left|S \cap L'_{p,i}\right| + d_{p,i} - e_{p,i} = \ell_{p,i}$.

In the modified formulation shown below, the constraints (2b) and (2c) for the two covariates are separated to facilitate the identification of the total unimodularity property. Because $L'_{1,1}, \ldots, L'_{1,k_1}$ is a partition of the control sample, $\sum_{i=1}^{k_1} |S \cap L'_{1,i}| = |S|$. Also, because $\ell_{1,1}, \ldots, \ell_{1,k_1}$ are the sizes of the levels of the treatment sample for the first covariate, it follows that $\sum_{i=1}^{k_1} \ell_{1,i} = n$. Therefore, $\sum_{i=1}^{k_1}\left(e_{1,i} - d_{1,i}\right) = \sum_{i=1}^{k_1}\left(|S \cap L'_{1,i}| - \ell_{1,i}\right) = |S| - n$. So specifying $|S| = n$ is equivalent to constraint (2d) in formulation (2) given below:

$$\min \quad \sum_{p=1}^{2}\sum_{i=1}^{k_p}\left(e_{p,i} + d_{p,i}\right) \tag{2a}$$

$$\text{s.t.} \quad \sum_{j \in L'_{1,i}} z_j + d_{1,i} - e_{1,i} = \ell_{1,i} \qquad i = 1, \ldots, k_1 \tag{2b}$$

$$\sum_{j \in L'_{2,i}} z_j + d_{2,i} - e_{2,i} = \ell_{2,i} \qquad i = 1, \ldots, k_2 \tag{2c}$$

$$-\sum_{i=1}^{k_1} d_{1,i} + \sum_{i=1}^{k_1} e_{1,i} = 0 \tag{2d}$$

$$e_{p,i}, d_{p,i} \geq 0 \qquad p = 1, 2, \quad i = 1, \ldots, k_p \tag{2e}$$

$$z_j \in \{0, 1\} \qquad j = 1, \ldots, n'. \tag{2f}$$

Similar observations can be used to show that $\sum_{i=1}^{k_p} e_i = \sum_{i=1}^{k_p} d_i$ for any covariate p, so the objective (2a) can also be written with only the excess variables.

Lemma 1. *The constraint matrix of the LP relaxation of formulation (2) is totally unimodular.*

Proof. In the constraint matrix of (2) each entry is 0, 1 or −1. Consider the matrix resulting by multiplying the rows of constraint (2c) by −1. Each column in this new matrix has at most one 1 and at most one −1:

1. Both $\{L'_{1,1}, \ldots, L'_{1,k_1}\}$ and $\{L'_{2,1}, \ldots, L'_{2,k_2}\}$ are partitions of the control sample, so $L'_{1,1}, \ldots, L'_{1,k_1}$ are mutually disjoint as are $L'_{2,1}, \ldots, L'_{2,k_2}$. The column of each z_j has exactly one 1 in rows corresponding to (2b), and one −1 (after multiplication) in rows corresponding to (2c).
2. For each i, the column of $d_{1,i}$ has exactly one 1 in rows corresponding to (2b) and exactly one −1 in rows corresponding to (2d); the column of $e_{1,i}$ has exactly one −1 in rows corresponding to (2b) and exactly one 1 in rows corresponding to (2d).
3. For each i, the column of $d_{2,i}$ has exactly one non-zero, 1 or −1, entry in rows corresponding to (2c); the column of $e_{2,i}$ has exactly one non-zero, 1 or −1, entry in rows corresponding to (2c).

Hence, by a well-known theorem (Theorem 7 in Appendix D) this new matrix is totally unimodular. Multiplying some rows of a totally unimodular matrix by −1 preserves total unimodularity. Therefore, the constraint matrix of the LP relaxation of (2) is also totally unimodular. □

Formulation (2) is also a minimum cost network flow (MCNF) formulation (see Appendix D for a generic formulation of MCNF). A generic MCNF formulation has exactly one 1 and one −1 in each column of the constraint matrix. To make formulation (2) have this structure, we multiply all coefficients in constraints (2c) by −1 and add a redundant constraint $\sum_{i=1}^{k_2} d_{2,i} - \sum_{i=1}^{k_2} e_{2,i} = 0$. In the next section, we streamline this network flow formulation.

4. Network flow formulation for $P = 2$

Here we use the level-intersection sizes as variables, x_{i_1,i_2} for $i_1 = 1, \ldots, k_1$, $i_2 = 1, \ldots, k_2$. These variables can also be written as $x_{i_1,i_2} = \sum_{j \in L'_{1,i_1} \cap L'_{2,i_2}} z_j$, with upper bounds given by $u_{i_1,i_2} = \left|L'_{1,i_1} \cap L'_{2,i_2}\right|$. With these decision variables we get the following network flow formulation:

$$\min \quad \sum_{p=1}^{2}\sum_{i=1}^{k_p}\left(e_{p,i} + d_{p,i}\right) \tag{3a}$$

$$\text{s.t.} \quad \sum_{i_2=1}^{k_2} x_{i_1,i_2} + d_{1,i_1} - e_{1,i_1} = \ell_{1,i_1} \qquad i_1 = 1, \ldots, k_1 \tag{3b}$$

$$-\sum_{i_1=1}^{k_1} x_{i_1,i_2} - d_{2,i_2} + e_{2,i_2} = -\ell_{2,i_2} \qquad i_2 = 1, \ldots, k_2 \tag{3c}$$

$$-\sum_{i_1=1}^{k_1} d_{1,i_1} + \sum_{i_1=1}^{k_1} e_{1,i_1} = 0 \tag{3d}$$

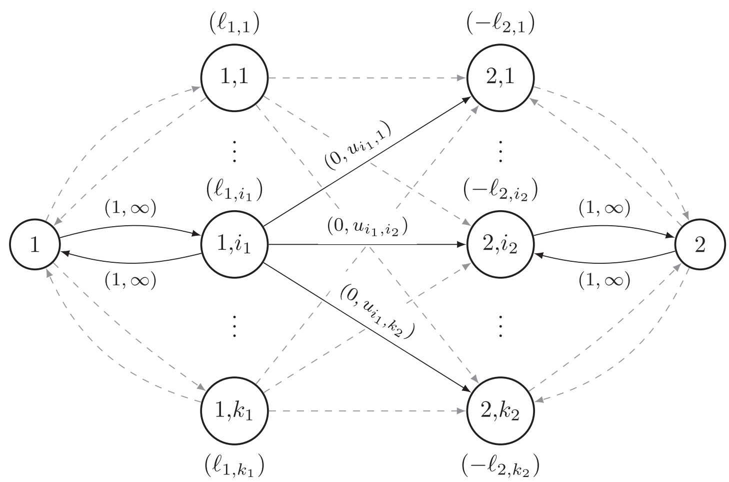

Fig. 1. MCNF graph corresponding to formulation (3). Arc labels have the form (cost, upper bound), and non-zero supplies and demands are displayed next to each node.

$$\sum_{i_2=1}^{k_2} d_{2,i_2} - \sum_{i_2=1}^{k_2} e_{2,i_2} = 0 \tag{3e}$$

$$e_{p,i}, d_{p,i} \geq 0 \qquad p = 1, 2,\ i = 1, \ldots, k_p \tag{3f}$$

$$0 \leq x_{i_1,i_2} \leq u_{i_1,i_2} \qquad \begin{array}{l} i_1 = 1, \ldots, k_1, \\ i_2 = 1, \ldots, k_2. \end{array} \tag{3g}$$

Formulation (3) is a minimum cost network flow problem. The corresponding network is shown in Fig. 1, where all capacity lower bounds are 0, and each arc has a cost per unit flow and upper bound associated with it. Nodes of type $(1, i_1)$ each have supply of ℓ_{1,i_1}. Nodes of type $(2, i_2)$ each has demand of ℓ_{2,i_2}. For each i_1 and i_2, the flow on the arc between node $(1, i_1)$ and node $(2, i_2)$ represents variable x_{i_1,i_2}. The arc from node 1 to node $(1, i_1)$ represents the excess e_{1,i_1}. The arc to node 1 from any node $(1, i_1)$ represents the deficit d_{1,i_1}. The arc from node 2 to any node $(2, i_2)$ represents the deficit d_{2,i_2}. The arc to node 2 from any node $(2, i_2)$ represents the excess e_{2,i_2}. The per unit cost is 1 for arcs between node 1 or 2 and any node in $\{(1, 1), (1, 2), \ldots, (1, k_1)\} \cup \{(2, 1), (2, 2), \ldots, (2, k_2)\}$; all other arcs have cost 0. It is easy to verify that constraints (3b) correspond to flow balance at nodes $(1, i_1)$ for all i_1, constraints (3c) correspond to flow balance at nodes $(2, i_2)$ for all i_2. Constraint (3d) corresponds to the flow balance at node 1, and constraint (3e) corresponds to flow balance at node 2. A small numerical example can be found in Appendix E. (Note that (3a) can also be written with only the excesses such as in formulation (5).)

Theorem 1. *The 2-covariate min-imbalance problem with a selection size of n is solved as a minimum cost network flow problem in $O(n \cdot (n' + n \log n))$ time.*

Proof. We choose the algorithm of *successive shortest paths* that is particularly efficient for a MCNF with "small" total supply to solve the network flow problem of the 2-covariate min-imbalance problem.

The successive shortest path algorithm iteratively selects a node s with excess supply (supply not yet sent to some demand node) and a node t with unfulfilled demand and sends flow from s to t along a shortest path in the residual network (Busacker & Gowen, 1961; Iri, 1960; Jewell, 1958). The algorithm terminates when the flow satisfies all the flow balance constraints. At each iteration, the number of remaining units of supply to be sent is reduced by at least one unit, so the number of iterations is bounded by the total amount of supply. For our problem the total supply is $O(n)$.

At each iteration, the shortest path can be solved with Dijkstra's algorithm of complexity $O(|A| + |V| \log |V|)$, where $|V|$ is the number of nodes and $|A|$ is the number of arcs (Edmonds & Karp, 1972; Tomizawa, 1971). In our formulation, $|V|$ is $O(k_1 + k_2)$, which is at most $O(n)$. Because the number of nonempty sets $L'_{1,i_1} \cap L'_{2,i_2}$ is at most $\min\{n', k_1k_2\}$, the number of arcs $|A|$ is $O(\min\{n', k_1k_2\})$. Hence, the total running time of applying the successive shortest path algorithm with node potentials on our formulation is $O(n \cdot (n' + n \log n))$. □

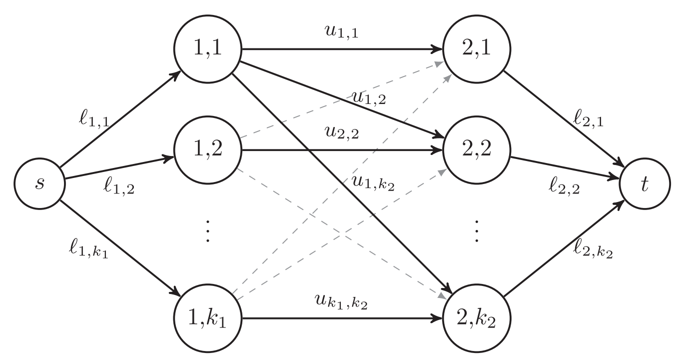

Fig. 2. Maximum flow graph. Arc labels indicate upper bounds.

5. Maximum flow formulation for $P = 2$

Here we show a maximum flow (max-flow) formulation (see Appendix D for a generic formulation of max-flow problem) for the min-imbalance problem with 2 covariates and a selection size of $q = n$. Unlike the previous formulations, the maximum flow solution requires further manipulation in order to derive an optimal solution to the min-imbalance problem with 2 covariates. That max-flow graph is illustrated in Fig. 2 (see Appendix E for an example). The source node s can send at most ℓ_{1,i_1} units of flow to node $(1, i_1)$ for each $i_1 = 1, \ldots, k_1$, the sink node can get at most ℓ_{2,i_2} units of flow from node $(2, i_2)$ for each $i_2 = 1, \ldots, k_2$, and there can be a flow from node $(1, i_1)$ to node $(2, i_2)$ with amount bounded by u_{i_1,i_2}, for $i_1 = 1, \ldots, k_1$ and $i_2 = 1, \ldots, k_2$.

Let the maximum flow value for the max-flow problem presented in Fig. 2 be denoted by f^*, and let $\mathbf{x}^*$ be the optimal flow vector, with $x^*_{i_1,i_2}$ denoting the flow amount between node $(1, i_1)$ and node $(2, i_2)$. It is obvious that $\sum_{i_1=1}^{k_1} \sum_{i_2=1}^{k_2} x^*_{i_1,i_2} = f^* \leq$

$\sum_{i_1=1}^{k_1} \ell_{1,i_1} = n$. That means an initial selection S' generated by selecting the prescribed number of control units as in the optimal max-flow solution, i.e., selecting $x^*_{i_1,i_2}$ control units from $L'_{1,i_1} \cap L'_{2,i_2}$ is of size f^*. In order to get a feasible solution for the min-imbalance problem it is required to select $n - f^*$ additional control units. The selection S' has no positive excess, only non-negative deficits with respect to the levels of both covariates. This is because $\sum_{i_2=1}^{k_2} x^*_{i_1,i_2} \le \ell_{1,i_1}$ due to the upper bound of the arc from s to $(1, i_1)$ for each i_1, and $\sum_{i_1=1}^{k_1} x^*_{i_1,i_2} \le \ell_{2,i_2}$ due to the upper bound of the arc from $(2, i_2)$ to t for each i_2.

To recover an optimal solution for the min-imbalance problem from the initial set S', we add up to $n - f^*$ unselected control units, one at a time, each corresponding to a level with positive deficit under either covariate 1 or 2. This process is repeated until either $n - f^*$ such control units are found, or until no such control unit exists. In the latter case, to complete the size of the selection, any randomly selected control units are added. Algorithm 1 is a formal

Algorithm 1 .

Initialization step: Select $x^*_{i_1,i_2}$ control units from $L'_{1,i_1} \cap L'_{2,i_2}$ in set S'.
while $|S'| < n$ **do**
  if there exists a control unit $j \notin S'$ whose covariate 1 level is i_1 and covariate 2 level is i_2, such that $|S' \cap L'_{1,i_1}| < \ell_{1,i_1}$ or $|S' \cap L'_{2,i_2}| < \ell_{2,i_2}$ **then**,
    $S' \leftarrow S' \cup \{j\}$.
  else
    Let $S'' = S'$ and let S^+ be any $n - |S'|$ control units$\notin S'$.
    Set $S' \leftarrow S' \cup S^+$.
Output S'.

statement of this process of recovering an optimal solution of the min-imbalance problem from the initial selection S'.

To show that Algorithm 1 provides an optimal solution to the min-imbalance problem, we distinguish two cases of Algorithm 1: (1) $S^+ = \emptyset$ and (2) $|S^+| \ge 1$. In the first case, there is, at each iteration, at least one control unit that belongs to some level with positive deficit. In Theorem 2 we prove that the output S' of Algorithm 1 is an optimal solution in this case.

Theorem 2. *If $S^+ = \emptyset$ then the output selection S' of Algorithm 1 is optimal for the min-imbalance problem, with an optimal objective value of $2(n - f^*)$.*

Proof. First, we show that the total imbalance of the selection S' is $IM(S') = 2(n - f^*)$. At the initialization step the selection S' has only deficits for all levels, with total deficit for covariate 1, $\sum_{i_1=1}^{k_1} (\ell_{1,i_1} - \sum_{i_2=1}^{k_2} x_{i_1,i_2}) = n - f^*$, and total deficit for covariate 2, $\sum_{i_2=1}^{k_2} (\ell_{2,i_2} - \sum_{i_1=1}^{k_1} x_{i_1,i_2}) = n - f^*$. At each iteration, there is an added control unit, say in $L'_{1,i_1} \cap L'_{2,i_2}$, such that either L'_{1,i_1} or L'_{2,i_2} has a positive deficit with respect to S'. It is however impossible for both L'_{1,i_1} and L'_{2,i_2} to have a positive deficit with respect to S' since otherwise, there is an s,t-augmenting path, from s to node $(1, i_1)$, to node $(2, i_2)$, to t, along which the flow can be increased by at least one unit. This is in contradiction to the optimality of the max-flow solution $\mathbf{x}^*$. As a result, at each iteration where a control unit is added, the total deficit is reduced by one unit, and the total excess is *increased* by one unit. Thus, at each iteration of the **if** step, the sum of total deficit and excess remains the same, namely $2(n - f^*)$.

Suppose, by contradiction, that there exists a selection S^* for which the total imbalance is lower, $IM(S^*) < 2(n - f^*)$. We repeat the following iterative procedure of removing control units from S^* until there is no positive excess remaining: while there is a level of either covariate with positive excess with respect to S^*, we remove one control unit of S^* that belongs to this level. Each such iteration results in the total excess reducing by at least 1 unit and the total deficit increasing by at most 1 unit, and therefore the sum of total deficit and excess does not increase. So when this iterative procedure ends, the total excess is zero and the total deficit is at most $IM(S^*)$. Let x_{i_1,i_2} be the number of control units remaining in $S^* \cap L'_{1,i_1} \cap L'_{2,i_2}$ after this excess removing procedure. Because there is no positive excess, $\mathbf{x}$ is a feasible solution for the max-flow problem with the flow between node $(1, i_1)$ and node $(2, i_2)$ equal to x_{i_1,i_2}. The sum of deficits associated with this remaining set is $n - \sum_{i_1=1}^{k_1} \sum_{i_2=1}^{k_2} x_{i_1,i_2}$ for covariate 1 and $n - \sum_{i_1=1}^{k_1} \sum_{i_2=1}^{k_2} x_{i_1,i_2}$ for covariate 2, for a total of $2(n - \sum_{i_1=1}^{k_1} \sum_{i_2=1}^{k_2} x_{i_1,i_2})$, which is at most $IM(S^*)$. Therefore, the total flow value, $\sum_{i_1=1}^{k_1} \sum_{i_2=1}^{k_2} x_{i_1,i_2}$, satisfies that it is at least $n - \frac{IM(S^*)}{2}$. Because $n - \frac{IM(S^*)}{2} > n - (n - f^*) = f^*$, it follows that the value of the feasible flow induced by the set S^* is greater than the maximum flow value f^*, which contradicts the optimality of f^*. □

We now address the second case where $|S^+| \ge 1$ and $|S''| < n$. In this case, the total imbalance of S'' is, from the arguments in the proof of Theorem 2, $IM(S'') = 2(n - f^*)$. Each one of the $|S^+|$ control units selected adds 1 unit of excess to each covariate, resulting in the addition of 2 units of excess to the imbalance. Therefore, the total imbalance of the output solution is $2(n - f^*) + 2|S^+|$. We next show what the value of $|S^+|$ is, and then demonstrate that any feasible selection to the min-imbalance problem has total imbalance of at least $2(n - f^*) + 2|S^+|$. This will prove that the output of Algorithm 1, S', is an optimal solution to the min-imbalance problem.

It will be useful to consider an equivalent form of Algorithm 1. For each level i of covariate p that has $|S' \cap L'_{p,i}| < \ell_{p,i}$, we add the largest number possible of available control units in $L'_{p,i}$ so long as the total does not exceed n. This number is $\min\{\ell_{p,i} - |S' \cap L'_{p,i}|, \ell'_{p,i} - |S' \cap L'_{p,i}|\}$. Let $\bar{\ell}_{p,i} = \min\{\ell_{p,i}, \ell'_{p,i}\}$, then for each p, i that has $|S' \cap L'_{p,i}| < \ell_{p,i}$ we add $\bar{\ell}_{p,i} - |S' \cap L'_{p,i}|$ previously unselected control units to S'. The outcome of this equivalent procedure is exactly the same as that of Algorithm 1. In the case that $|S^+| \ge 1$ there is an insufficient number of control units to add to S' after the largest possible number has been added for all levels. Therefore, at the end of this process, the **if** step returns that another unselected control unit does not exist, and the total number of control units of S'', for each level i of covariate p, is $\bar{\ell}_{p,i}$.

Lemma 2. *If $|S^+| \ge 1$ (and $|S''| = n - |S^+| < n$) then $|S^+| = n - (\bar{\ell}_1 + \bar{\ell}_2 - f^*)$ where $\bar{\ell}_1 = \sum_{i_1=1}^{k_1} \bar{\ell}_{1,i_1}$ and $\bar{\ell}_2 = \sum_{i_2=1}^{k_2} \bar{\ell}_{2,i_2}$.*

Proof. At the initialization step of Algorithm 1, $|S'| = f^*$ and the total deficit is $2(n - f^*)$. Each time a control unit is added to S' in the **if** step, the total deficit is decreased by 1 unit. So we can derive the value of $|S''|$ when the algorithm terminates if we know the total deficit when the algorithm terminates. Note that the total excess may change, but we only consider here the deficit.

From the discussion above, the total number of control units of S'', for each level i of covariate p, is $\bar{\ell}_{p,i}$. We denote $\bar{\ell}_1 = \sum_{i_1=1}^{k_1} \bar{\ell}_{1,i_1}$, and $\bar{\ell}_2 = \sum_{i_2=1}^{k_2} \bar{\ell}_{2,i_2}$. Because the sum $\sum_{i_1=1}^{k_1} \ell_{1,i_1} = n$ and $\sum_{i_2=1}^{k_2} \ell_{2,i_2} = n$, the sum of deficits of set S'' under covariate 1 is $\sum_{i_1=1}^{k_1} \ell_{1,i_1} - \bar{\ell}_{1,i_1} = n - \bar{\ell}_1$, and the sum of deficits under covariate 2 equals $\sum_{i_2=1}^{k_2} \ell_{2,i_2} - \bar{\ell}_{2,i_2} = n - \bar{\ell}_2$. It follows that the sum of deficits of S'' is $2n - \bar{\ell}_1 - \bar{\ell}_2$.

Because the initial set S' that has total deficit of $2(n - f^*)$ has its deficit reduced through Algorithm 1 to $2n - \bar{\ell}_1 - \bar{\ell}_2$ in the set S'', the additional number of control units in S'' that were added to the initial f^* control units is $2(n-f^*)-(2n-\bar{\ell}_1-\bar{\ell}_2)=\bar{\ell}_1+\bar{\ell}_2-2f^*$. Therefore, the size of S'' is $f^*+(\bar{\ell}_1+\bar{\ell}_2-2f^*)=\bar{\ell}_1+\bar{\ell}_2-f^*$. This number is less than n and the size of S^+ then satisfies $|S^+|=n-(\bar{\ell}_1+\bar{\ell}_2-f^*)$. □

Corollary 1. *If $|S^+|\geq 1$ when Algorithm 1 terminates, the total imbalance of the output solution S' is $IM(S')=4n-2\bar{\ell}_1-2\bar{\ell}_2$.*

Proof. The imbalance of S'', as in the proof of Theorem 2, is equal to $2(n-f^*)$. Because each control unit in S^+ adds two units to the imbalance, the total imbalance of the output solution S' is $IM(S')=2(n-f^*)+2(n-(\bar{\ell}_1+\bar{\ell}_2-f^*))$, which is equal to $4n-2\bar{\ell}_1-2\bar{\ell}_2$, as stated. □

Next, we prove that this is the minimum total imbalance achievable.

Theorem 3. *For any selection of size n, the total imbalance must be greater or equal to $4n-2\bar{\ell}_1-2\bar{\ell}_2$.*

Proof. For the optimal selection S^* of size n, let $IM(S^*)$ be the total imbalance of S^*. We first classify the control units in S^* into three types, S_1, S_2, and S_3, that form a partition of S^*, using the 3-type Classification Procedure shown in Algorithm 2. In the procedure we use variable $dis(p,i)$ to denote the value of $dis(S,p,i)$, discrepancy for selection S and level i under covariate p. With this notation the excess of the corresponding level is $e(S,p,i)=\max\{0, dis(p,i)\}$, and the deficit is $d(S,p,i)=\max\{0,-dis(p,i)\}$.

Algorithm 2 .

procedure 3-TYPE CLASSIFICATION
```
/* Initialize */
```
$S\leftarrow S^*$, $S_1\leftarrow\emptyset$, $S_2\leftarrow\emptyset$, $S_3\leftarrow\emptyset$
Let $dis(p,i)\leftarrow|S\cap L'_{p,i}|-\ell_{p,i}$ for $p=1,2,\ i=1,\dots,k_p$;
```
/* S_1 selection */
```
while there exists a control unit j in S whose covariate 1 level is i_1, covariate 2 level is i_2, such that $dis(1,i_1)>0$ and $dis(2,i_2)>0$ **do**
$S_1\leftarrow S_1\cup\{j\}$, $S\leftarrow S-\{j\}$
$dis(1,i_1)\leftarrow dis(1,i_1)-1$ $dis(2,i_2)\leftarrow dis(2,i_2)-1$
```
/* S_2 selection */
```
while there exists a control unit j in S whose covariate 1 level is i_1, covariate 2 level is i_2, such that $dis(1,i_1)>0$ or $dis(2,i_2)>0$ **do**
$S_2\leftarrow S_2\cup\{j\}$, $S\leftarrow S-\{j\}$
$dis(1,i_1)\leftarrow dis(1,i_1)-1$, $dis(2,i_2)\leftarrow dis(2,i_2)-1$
```
/* S_3 selection */
```
$S_3\leftarrow S$

The output S_1,S_2,S_3 of the procedure is not unique because it depends on the order in which control units are picked. However, the statements of the theorem hold for any output of the procedure. Note that in the procedure, whenever a control unit is picked, any $dis(p,i)$ can only go down. For that reason, once S_1 selection ends, there will not be another control unit in S for which the discrepancy values of the corresponding levels under the two covariates are both positive. Furthermore, once the S_1 and S_2 selections are done, $dis(p,i)\leq 0$ for each p,i. That means, $|S_3\cap L'_{p,i}|\leq\ell_{p,i}$ for all p,i.

Let the sizes of the three subsets be denoted by $s_1=|S_1|$, $s_2=|S_2|$, $s_3=|S_3|$. We claim that the total imbalance of the control units in S_3 is $IM(S^*)-2s_1$. For the S_1 selection part of the procedure, each control unit picked in S_1 reduces the total excess by 2. For each control unit selected in the S_2 selection part of the procedure, the excess is reduced by 1 and the deficit is increased by 1, so the total imbalance does not change. Therefore, the total imbalance of the control units in S_3 is $IM(S^*)-2s_1$. On the other hand, the total imbalance of S_3, which equals the sum of deficits of both covariates (all excesses equal zero as $|S_3\cap L'_{p,i}|\leq\ell_{p,i}$), is $2(n-s_3)$. Then $IM(S^*)-2s_1=2(n-s_3)$, and therefore

$$\begin{aligned}IM(S^*)&=2(n-s_3)+2s_1=2(n-s_3)+2(n-s_2-s_3)\\&=4n-2s_2-4s_3.\end{aligned}$$

Here, the second equality comes from the fact that $s_1+s_2+s_3=n$.

Next, we show that $s_2\leq(\bar{\ell}_1-s_3)+(\bar{\ell}_2-s_3)$. Let the control units in S_2 be ordered according to the order they were picked, $j_1,j_2,\dots,j_{s_2}$. We now add these control units to S_3, in the reverse order $j_{s_2},\dots,j_1$. When each control unit j_q is added to S_3, the deficit is reduced by exactly 1 unit. Once all the control units from S_2 are added to S_3, the deficit at each level of $S_2\cup S_3$ is zero, or alternatively, $dis(S_2\cup S_3,p,i)=|(S_2\cup S_3)\cap L'_{p,i}|-\ell_{p,i}\geq 0$ for each p,i.

We now consider the total deficit of S_3: By the definition of $\bar{\ell}_1$ and $\bar{\ell}_2$, the positive deficit of S_3 under covariate 1 is at most $\bar{\ell}_1-s_3$ and that the positive deficit of S_3 under covariate 2 is at most $\bar{\ell}_2-s_3$. That means the size of S_2 is bounded by the amount of this deficit, $s_2\leq(\bar{\ell}_1-s_3)+(\bar{\ell}_2-s_3)$. Then we have

$$\begin{aligned}s_2\leq(\bar{\ell}_1-s_3)+(\bar{\ell}_2-s_3)&\Leftrightarrow\ s_2+2s_3\leq\bar{\ell}_1+\bar{\ell}_2\\&\Leftrightarrow\ IM(S^*)=4n-2s_2-4s_3\\&\geq\ 4n-2\bar{\ell}_1-2\bar{\ell}_2.\end{aligned}$$

We conclude that the total imbalance $IM(S^*)$ is at least $4n-2\bar{\ell}_1-2\bar{\ell}_2$. That implies that the selection output of Algorithm 1, S', which has a total imbalance of $4n-2\bar{\ell}_1-2\bar{\ell}_2$, is optimal. □

The conclusion from Corollary 1 and Theorem 3, is that for $|S^+|\geq 1$ when Algorithm 1 terminates, the output solution S' is an optimal selection to the min-imbalance problem. Together with Theorem 2, we have that Algorithm 1 outputs an optimal selection for the min-imbalance problem using the max-flow solution to the flow problem in Fig. 2 as input.

Theorem 4. *The maximum flow formulation of the 2-covariate min-imbalance problem with a selection size of n is solved in $O\left(n'\cdot\min\left\{n^{\frac{2}{3}},n'^{\frac{1}{2}}\right\}\cdot\log^2 n\right)$ time.*

Proof. We choose here the *binary blocking flow* algorithm of Goldberg & Rao (1998) for solving the max-flow problem because this algorithm depends on the maximum arc capacity which is a small quantity in our formulation.

The complexity of the binary blocking flow algorithm for a graph $G=(V,A)$ is $O(|A|\cdot\min\{|V|^{\frac{2}{3}},|A|^{\frac{1}{2}}\}\cdot\log\frac{|V|^2}{|A|}\log U)$ where $|V|$ is number of nodes, $|A|$ is number of arcs, and U is maximum arc capacity. As argued earlier for the minimum cost network flow formulation, the number of nodes in the network $|V|$ is $O(k_1+k_2)$, which is no more than $O(n)$; and the number of arcs is bounded by $\min\{n',k_1k_2\}$. Although u_{i_1,i_2} could be as large as n', a feasible flow to our maximum flow formulation can not have more than ℓ_{1,i_1} units of flow on the arc from node $(1,i_1)$ to node $(2,i_2)$. Thus the maximum arc capacity U is effectively $O(n)$. The ratio $\frac{|V|^2}{|A|}\leq\frac{(k_1+k_2)^2}{k_1+k_2}\leq n$. Hence, the running time of applying the binary blocking flow algorithm to our max-flow problem is $O\left(n'\cdot\min\left\{n^{\frac{2}{3}},n'^{\frac{1}{2}}\right\}\cdot\log^2 n\right)$. The complexity of Algorithm 1 is $O(n)$ as the number of iterations is bounded by n, and each iteration takes $O(1)$ steps. Therefore, the running time of solving the min-imbalance problem as a max-flow problem is $O\left(n'\cdot\min\left\{n^{\frac{2}{3}},n'^{\frac{1}{2}}\right\}\cdot\log^2 n\right)$. □

6. Network flow formulations for $P = 2$ and alternate selection size

The preceding results apply when the selection S is required to have size equal to n, the size of the treatment sample. If the size of the selection is required to be some value $q \neq n$, then the min-imbalance problem with $P = 2$ covariates can be solved using either the min-cost network flow or maximum flow formulations with appropriate modifications described in Appendix B.

With $q \neq n$, the min-proportional imbalance objective is more appropriate than min-imbalance for causal inference applications. We show here how to solve the min-proportional imbalance problem with two covariates and selection size $q \neq n$ using an alternate minimum cost network flow formulation (a variant of this formulation was provided in Sauppe, 2015). For any selection S of size q, we define the *scaled proportional discrepancy* at level i of covariate p as $spd(S, p, i) = |S \cap L'_{p,i}| - \frac{q}{n}\ell_{p,i}$. As before, the scaled discrepancy can be positive or negative, leading to excess or deficit, respectively, and defined as $e_{p,i}(S) = \max\{0, spd(S, p, i)\}$ and $d_{p,i}(S) = \max\{0, -spd(S, p, i)\}$. The min-proportional imbalance objective can then be written as $\frac{1}{q}\sum_{p=1}^{P}\sum_{i=1}^{k_p}\left(e_{p,i}(S) + d_{p,i}(S)\right)$. We can formulate the min-proportional imbalance problem as a mixed integer program using binary variables z_j for each control unit and non-negative variables $e_{p,i}$ and $d_{p,i}$ for excess and deficit, respectively, for all p and i:

$$\min \quad \frac{1}{q}\sum_{p=1}^{P}\sum_{i=1}^{k_p}\left(e_{p,i} + d_{p,i}\right) \tag{4a}$$

$$\text{s.t.} \quad \sum_{j\in L'_{p,i}} z_j + d_{p,i} - e_{p,i} = \frac{q}{n}\ell_{p,i} \qquad p = 1,\ldots,P, \quad i = 1,\ldots,k_p \tag{4b}$$

$$\sum_{j=1}^{n'} z_j = q \tag{4c}$$

$$e_{p,i}, d_{p,i} \geq 0 \qquad p = 1,\ldots,P, \quad i = 1,\ldots,k_p \tag{4d}$$

$$z_j \in \{0,1\} \qquad j = 1,\ldots,n'. \tag{4e}$$

Summing constraints (4b) across all values of i for covariate 1 and rearranging yields $\sum_{j=1}^{n'} z_j = q + \sum_{i=1}^{k_1}\left(e_{1,i} - d_{1,i}\right)$. This means constraint (4c) can be replaced with the constraint $\sum_{i=1}^{k_1}\left(e_{1,i} - d_{1,i}\right) = 0$. A similar argument can be used to show that $\sum_{i=1}^{k_p} e_{p,i} = \sum_{i=1}^{k_p} d_{p,i}$ for any covariate $p = 1,\ldots,P$. As such, the objective function can be reformulated to penalize twice the excess while omitting the deficit, which will be useful for further reformulation.

In the case of two covariates, we introduce variables $x_{i_1,i_2} = \sum_{j\in L'_{1,i_1}\cap L'_{2,i_2}} z_j$ as before. Combining this with the above observations and some additional modifications allows formulation (4) to be transformed into the following:

$$\min \quad \frac{2}{q}\sum_{p=1}^{2}\sum_{i=1}^{k_p} e_{p,i} \tag{5a}$$

$$\text{s.t.} \quad \sum_{i_2=1}^{k_2} x_{i_1,i_2} + d_{1,i_1} - e_{1,i_1} = \frac{q}{n}\ell_{1,i_1} \qquad i_1 = 1,\ldots,k_1 \tag{5b}$$

$$-\sum_{i_1=1}^{k_1} x_{i_1,i_2} - d_{2,i_2} + e_{2,i_2} = -\frac{q}{n}\ell_{2,i_2} \qquad i_2 = 1,\ldots,k_2 \tag{5c}$$

$$-\sum_{i_1=1}^{k_1} d_{1,i_1} + \sum_{i_1=1}^{k_1} e_{1,i_1} = 0 \tag{5d}$$

$$\sum_{i_2=1}^{k_2} d_{2,i_2} - \sum_{i_2=1}^{k_2} e_{2,i_2} = 0 \tag{5e}$$

$$e_{p,i}, d_{p,i} \geq 0 \qquad p = 1,2, \quad i = 1,\ldots,k_p \tag{5f}$$

$$0 \leq x_{i_1,i_2} \leq u_{i_1,i_2} \qquad i_1 = 1,\ldots,k_1,\ i_2 = 1,\ldots,k_2. \tag{5g}$$

$$x_{i_1,i_2} \in \mathbb{Z} \qquad i_1 = 1,\ldots,k_1,\ i_2 = 1,\ldots,k_2. \tag{5h}$$

In contrast to formulation (3) from Section 4, formulation (5) may have non-integral supply and demand values, and so we retain integrality constraints (5h) on the variables associated with the control unit selection. These constraints can be removed after some additional modifications. Specifically, we add new non-negative variables $a_{p,i}$, and $b_{p,i}$ along with new constraints $a_{p,i} \leq \left\lfloor \frac{q}{n}\ell_{p,i} \right\rfloor$, $b_{p,i} \leq \frac{q}{n}\ell_{p,i} - \left\lfloor \frac{q}{n}\ell_{p,i} \right\rfloor$, and $d_{p,i} = \frac{q}{n}\ell_{p,i} - a_{p,i} - b_{p,i}$ for all $p = 1,2$ and $i = 1,\ldots,k_p$. The last set of these new constraints decompose each deficit variable $d_{p,i}$ into a fixed "forward flow" $\frac{q}{n}\ell_{p,i}$, an "integer backward flow" $a_{p,i}$, and a "fractional backward flow" $b_{p,i}$. Because $a_{p,i}$ and $b_{p,i}$ are non-negative, this has the effect of imposing an upper bound of $\frac{q}{n}\ell_{p,i}$ on $d_{p,i}$ for each p and i. Constraints (5b) and (5c) can be relaxed to $d_{p,i} \leq \frac{q}{n}\ell_{p,i} + e_{p,i}$, so this last set of new constraints does change the feasible region of formulation (5). However, this does not impact optimality because any optimal solution will set at least one of $d_{p,i}$ and $e_{p,i}$ to 0. With these modifications, the revised formulation is:

$$\min \quad \frac{2}{q}\sum_{p=1}^{2}\sum_{i=1}^{k_p} e_{p,i} \tag{6a}$$

$$\text{s.t.} \quad \left(\sum_{i_2=1}^{k_2} x_{i_1,i_2}\right) - \left(a_{1,i_1} + b_{1,i_1} + e_{1,i_1}\right) = 0 \qquad i_1 = 1,\ldots,k_1 \tag{6b}$$

$$\left(a_{2,i_2} + b_{2,i_2} + e_{2,i_2}\right) - \left(\sum_{i_1=1}^{k_1} x_{i_1,i_2}\right) = 0 \qquad i_2 = 1,\ldots,k_2 \tag{6c}$$

$$\sum_{i_1=1}^{k_1}\left(a_{1,i_1} + b_{1,i_1} + e_{1,i_1}\right) = q \tag{6d}$$

$$-\sum_{i_2=1}^{k_2}\left(a_{2,i_2} + b_{2,i_2} + e_{2,i_2}\right) = -q \tag{6e}$$

$$a_{p,i}, b_{p,i}, e_{p,i} \geq 0 \qquad p = 1,2, \quad i = 1,\ldots,k_p \tag{6f}$$

$$a_{p,i} \leq \left\lfloor \frac{q}{n}\ell_{p,i} \right\rfloor \qquad p = 1,2, \quad i = 1,\ldots,k_p \tag{6g}$$

$$b_{p,i} \leq \frac{q}{n}\ell_{p,i} - \left\lfloor \frac{q}{n}\ell_{p,i} \right\rfloor \qquad p = 1,2, \quad i = 1,\ldots,k_p \tag{6h}$$

$$0 \leq x_{i_1,i_2} \leq u_{i_1,i_2} \qquad i_1 = 1,\ldots,k_1,\ i_2 = 1,\ldots,k_2. \tag{6i}$$

$$x_{i_1,i_2} \in \mathbb{Z} \qquad i_1 = 1,\ldots,k_1,\ i_2 = 1,\ldots,k_2. \tag{6j}$$

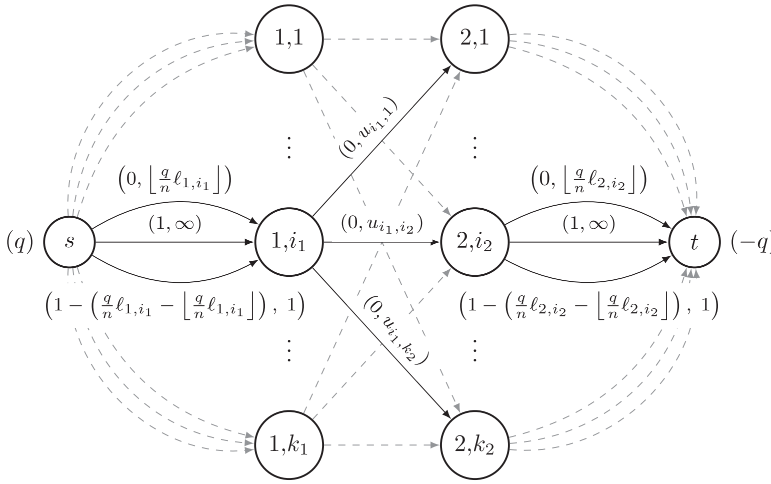

Fig. 3. MCNF graph for formulation (7). Arc labels have the form (cost, upper bound), and non-zero supplies and demands are displayed next to each node. Costs displayed in the figure omit a $2/q$ scaling.

Formulation (6) has integer supply and demand values, but the upper bound constraints (6h) on the $b_{p,i}$ variables may be non-integral. To address this, we raise the capacity on $b_{p,i}$ to 1, which allows us to send an additional $1-\left(\frac{q}{n}\ell_{p,i} - \left\lfloor \frac{q}{n}\ell_{p,i} \right\rfloor\right)$ units of flow along this edge, but we also add a cost of $\frac{2}{q} \cdot \left[1-\left(\frac{q}{n}\ell_{p,i} - \left\lfloor \frac{q}{n}\ell_{p,i} \right\rfloor\right)\right]$ per unit of flow sent on edge $b_{p,i}$. This ensures that all demands, supplies, and capacities are integral. Lastly, we drop the integrality requirements on the x_{i_1,i_2} variables to get the following MCNF formulation:

$$\min \quad \frac{2}{q}\sum_{p=1}^{2}\sum_{i=1}^{k_p}\left[\left(1-\left(\frac{q}{n}\ell_{p,i} - \left\lfloor \frac{q}{n}\ell_{p,i} \right\rfloor\right)\right)b_{p,i} + e_{p,i}\right] \tag{7a}$$

$$\text{s.t.} \quad \left(\sum_{i_2=1}^{k_2} x_{i_1,i_2}\right) - \left(a_{1,i_1} + b_{1,i_1} + e_{1,i_1}\right) = 0 \qquad i_1 = 1,\dots,k_1 \tag{7b}$$

$$\left(a_{2,i_2} + b_{2,i_2} + e_{2,i_2}\right) - \left(\sum_{i_1=1}^{k_1} x_{i_1,i_2}\right) = 0 \qquad i_2 = 1,\dots,k_2 \tag{7c}$$

$$\sum_{i_1=1}^{k_1}\left(a_{1,i_1} + b_{1,i_1} + e_{1,i_1}\right) = q \tag{7d}$$

$$-\sum_{i_2=1}^{k_2}\left(a_{2,i_2} + b_{2,i_2} + e_{2,i_2}\right) = -q \tag{7e}$$

$$a_{p,i}, b_{p,i}, e_{p,i} \ge 0 \qquad \begin{array}{l} p = 1,2, \\ i = 1,\dots,k_p \end{array} \tag{7f}$$

$$a_{p,i} \le \left\lfloor \frac{q}{n}\ell_{p,i} \right\rfloor \qquad \begin{array}{l} p = 1,2, \\ i = 1,\dots,k_p \end{array} \tag{7g}$$

$$b_{p,i} \le 1 \qquad \begin{array}{l} p = 1,2, \\ i = 1,\dots,k_p \end{array} \tag{7h}$$

$$0 \le x_{i_1,i_2} \le u_{i_1,i_2} \qquad \begin{array}{l} i_1 = 1,\dots,k_1, \\ i_2 = 1,\dots,k_2. \end{array} \tag{7i}$$

The network associated with formulation (7) is shown in Fig. 3 (see Appendix E for an example). Because all supplies, demands, and capacities are integral, formulation (7) has an integer optimal solution.

Theorem 5. *Let $\mathcal{S}^{(7*)}$ be an integer optimal solution to formulation (7). Then the solution $\mathcal{S}^{(6)}$ defined as*

$$x^{(6)}_{i_1,i_2} = x^{(7*)}_{i_1,i_2} \qquad b^{(6)}_{p,i} = b^{(7*)}_{p,i} - \left[1-\left(\frac{q}{n}\ell_{p,i} - \left\lfloor \frac{q}{n}\ell_{p,i} \right\rfloor\right)\right]b^{(7*)}_{p,i}$$

$$a^{(6)}_{p,i} = a^{(7*)}_{p,i} \qquad e^{(6)}_{p,i} = e^{(7*)}_{p,i} + \left[1-\left(\frac{q}{n}\ell_{p,i} - \left\lfloor \frac{q}{n}\ell_{p,i} \right\rfloor\right)\right]b^{(7*)}_{p,i}$$

for all respective indices is optimal for formulation (6).

Proof. By construction, $\mathcal{S}^{(6)}$ satisfies the flow balance constraints (6b)–(6e) as well as the bounds constraints on all variables $a_{p,i}$, $e_{p,i}$, and x_{i_1,i_2}. Additionally,

$$b^{(6)}_{p,i} = b^{(7*)}_{p,i} - \left[1-\left(\frac{q}{n}\ell_{p,i} - \left\lfloor \frac{q}{n}\ell_{p,i} \right\rfloor\right)\right]b^{(7*)}_{p,i}$$
$$= \left(\frac{q}{n}\ell_{p,i} - \left\lfloor \frac{q}{n}\ell_{p,i} \right\rfloor\right)b^{(7*)}_{p,i},$$

so it follows that $0 \le b^{(6)}_{p,i} \le \frac{q}{n}\ell_{p,i} - \left\lfloor \frac{q}{n}\ell_{p,i} \right\rfloor$. Hence, $\mathcal{S}^{(6)}$ is feasible for formulation (6). We also have

$$\frac{2}{q}\sum_{p=1}^{2}\sum_{i=1}^{k_p}\left[\left(1-\left(\frac{q}{n}\ell_{p,i} - \left\lfloor \frac{q}{n}\ell_{p,i} \right\rfloor\right)\right)b^{(7*)}_{p,i} + e^{(7*)}_{p,i}\right] = \frac{2}{q}\sum_{p=1}^{2}\sum_{i=1}^{k_p} e^{(6)}_{p,i},$$

so the costs of $\mathcal{S}^{(7*)}$ and $\mathcal{S}^{(6)}$ in their respective formulations are equal.

Suppose that $\mathcal{S}^{(6)}$ is not optimal for formulation (6), and let $\mathcal{S}^{(6*)}$ be an optimal solution with $\frac{2}{q}\sum_{p=1}^{2}\sum_{i=1}^{k_p} e^{(6*)}_{p,i} < \frac{2}{q}\sum_{p=1}^{2}\sum_{i=1}^{k_p} e^{(6)}_{p,i}$. We will use $\mathcal{S}^{(6*)}$ to construct a solution $\mathcal{S}^{(7)}$ for formulation (7). Before doing so, we make the following observation for any $p = 1,2$ and $i = 1,\dots,k_p$: the constraints (6b), (6c), and (6j) imply that the quantity $a^{(6*)}_{p,i} + b^{(6*)}_{p,i} + e^{(6*)}_{p,i}$ must be an integer. Then $\mathcal{S}^{(7)}$ is constructed as follows. Let $x^{(7)}_{i_1,i_2} = x^{(6*)}_{i_1,i_2}$ for all $i_1 = 1,\dots,k_1$ and $i_2 = 1,\dots,k_2$. The variables $a^{(7)}_{p,i}$, $b^{(7)}_{p,i}$, and $e^{(7)}_{p,i}$ for each $p = 1,2$ and $i = 1,\dots,k_p$ are determined based on two cases.

Case 1: $e^{(6*)}_{p,i} = 0$. Let $a^{(7)}_{p,i} = \left\lfloor a^{(6*)}_{p,i} + b^{(6*)}_{p,i} \right\rfloor$, $b^{(7)}_{p,i} = \left(a^{(6*)}_{p,i} + b^{(6*)}_{p,i}\right) - \left\lfloor a^{(6*)}_{p,i} + b^{(6*)}_{p,i} \right\rfloor$, and $e^{(7)}_{p,i} = e^{(6*)}_{p,i}$. By construction, these values satisfy the flow balance constraints (7b)–(7e). From constraints (6g) and (6h), we have

$$a^{(6*)}_{p,i} + b^{(6*)}_{p,i} \le \left\lfloor \frac{q}{n}\ell_{p,i} \right\rfloor + \left(\frac{q}{n}\ell_{p,i} - \left\lfloor \frac{q}{n}\ell_{p,i} \right\rfloor\right) = \frac{q}{n}\ell_{p,i},$$

so $a_{p,i}^{(7)}$ satisfies constraint (7g). Combining $e_{p,i}^{(6*)} = 0$ with the earlier observation that $a_{p,i}^{(6*)} + b_{p,i}^{(6*)} + e_{p,i}^{(6*)}$ is an integer implies that $a_{p,i}^{(6*)} + b_{p,i}^{(6*)}$ is an integer, which means that $b_{p,i}^{(7*)} = 0$ and hence it satisfies constraint (7h). We also have

$$e_{p,i}^{(6*)} = e_{p,i}^{(7)} = \left[1 - \left(\frac{q}{n}\ell_{p,i} - \left\lfloor \frac{q}{n}\ell_{p,i} \right\rfloor\right)\right] b_{p,i}^{(7)} + e_{p_i}^{(7)},$$

so $\mathcal{S}^{(6*)}$ and $\mathcal{S}^{(7)}$ have equal flow costs on the edges $a_{p,i}$, $b_{p,i}$, and $e_{p,i}$ in their respective formulations.

Case 2: $e_{p,i}^{(6*)} > 0$. Let $a_{p,i}^{(7)} = a_{p,i}^{(6*)}$, $b_{p,i}^{(7)} = b_{p,i}^{(6*)} + \left[1 - \left(\frac{q}{n}\ell_{p,i} - \left\lfloor \frac{q}{n}\ell_{p,i} \right\rfloor\right)\right]$, and $e_{p,i}^{(7)} = e_{p,i}^{(6*)} - \left[1 - \left(\frac{q}{n}\ell_{p,i} - \left\lfloor \frac{q}{n}\ell_{p,i} \right\rfloor\right)\right]$. By construction, these values satisfy the flow balance constraints (7b)–(7e). As $e_{p,i}^{(6*)} > 0$, we must have $a_{p,i}^{(6*)} = \left\lfloor \frac{q}{n}\ell_{p,i} \right\rfloor$ and $b_{p,i}^{(6*)} = \frac{q}{n}\ell_{p,i} - \left\lfloor \frac{q}{n}\ell_{p,i} \right\rfloor$, because otherwise the cost of $\mathcal{S}^{(6*)}$ could be decreased by moving some flow from edge $e_{p,i}$ to either $a_{p,i}$ or $b_{p,i}$. Therefore, $a_{p,i}^{(7)}$ and $b_{p,i}^{(7)}$ satisfy constraints (7g) and (7h), respectively, with $b_{p,i}^{(7)} = 1$. Additionally, through constraints (6b), (6c), and (6j), the quantity $a_{p,i}^{(6*)} + b_{p,i}^{(6*)} + e_{p,i}^{(6*)}$ must equal some integer z. Then $e_{p,i}^{(6*)} = z - \left(a_{p,i}^{(6*)} + b_{p,i}^{(6*)}\right) = z - \left(\frac{q}{n}\ell_{p,i}\right)$. Combining this with $e_{p,i}^{(6*)} > 0$ implies that $z - \left(\frac{q}{n}\ell_{p,i}\right) > 0$, or equivalently $z > \frac{q}{n}\ell_{p,i}$. Because z is an integer, $z > \frac{q}{n}\ell_{p,i}$ implies $z \ge 1 + \left\lfloor \frac{q}{n}\ell_{p,i} \right\rfloor$. Then

$$\begin{aligned} e_{p,i}^{(6*)} &= z - \left(\frac{q}{n}\ell_{p,i}\right) \ge \left(1 + \left\lfloor \frac{q}{n}\ell_{p,i} \right\rfloor\right) - \left(\frac{q}{n}\ell_{p,i}\right) \\ &= 1 - \left(\frac{q}{n}\ell_{p,i} - \left\lfloor \frac{q}{n}\ell_{p,i} \right\rfloor\right), \end{aligned}$$

and therefore $e_{p,i}^{(7)} \ge 0$. We also have

$$\begin{aligned} e_{p,i}^{(6*)} &= \left[1 - \left(\frac{q}{n}\ell_{p,i} - \left\lfloor \frac{q}{n}\ell_{p,i} \right\rfloor\right)\right] + e_{p_i}^{(7)} \\ &= \left[1 - \left(\frac{q}{n}\ell_{p,i} - \left\lfloor \frac{q}{n}\ell_{p,i} \right\rfloor\right)\right] b_{p,i}^{(7)} + e_{p_i}^{(7)}, \end{aligned}$$

so $\mathcal{S}^{(6*)}$ and $\mathcal{S}^{(7)}$ have equal flow costs on the edges $a_{p,i}$, $b_{p,i}$, and $e_{p,i}$ in their respective formulations.

Using the construction process outlined in these two cases ensures that $\mathcal{S}^{(7)}$ is feasible for formulation (7) and has the same cost as $\mathcal{S}^{(6*)}$, which is less than the cost of $\mathcal{S}^{(7*)}$, which contradicts the optimality of $\mathcal{S}^{(7*)}$. □

Theorem 6. *The min-proportional imbalance problem with two covariates and a selection size of q is solved as a minimum cost network flow problem in* $O(q \cdot (n' + n\log n))$ *time.*

Proof. The network associated with formulation (7) has $O(n)$ vertices, $O(\min\{n', k_1k_2\})$ arcs, and a total supply of q. As such, the algorithm of successive shortest paths can be applied to solve this MCNF in $O(q \cdot (n' + n\log n))$ time. □

7. Conclusions

We present new insights to the min-imbalance problem that involves the selection of units from a control sample with the goal of minimizing covariate imbalance with respect to a treatment sample. We show that an integer programming formulation of the problem on two covariates has a totally unimodular constraint matrix. We then present and analyze two efficient approaches to solve the problem for two covariates and a selection size of n. The first approach is based on minimum cost network flow, and the second more efficient approach is based on a maximum flow formulation. In addition, we show how these results can be applied to a related two-stage problem involving minimum imbalance in the first stage and matching in the second. In the case that the selection size differs from n, we show how to solve both the min-imbalance and min-proportional imbalance problems efficiently with two covariates. In particular, proportional imbalance requires an alternate MCNF formulation to deal with non-integral supplies and demands. We also provide a proof that the min-imbalance problem is NP-hard for three or more covariates. The solutions for the two-covariate problems can be used in problems with three or more covariates, for example by aggregating covariates into two representative covariates or by providing bounds in a branch-and-bound algorithm; exploring these ideas is left for future work.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ejor.2021.10.041

References

Bennett, M., Vielma, J. P., & Zubizarreta, J. (2020). Building representative matched samples with multi-valued treatments in large observational studies. *Journal of Computational and Graphical Statistics, 0*(0), 1–29.
Busacker, R., & Gowen, P. (1961). A procedure for determining minimal-cost network flow patterns. *Technical Report ORO Technical Report*. Operational Research Office, John Hopkins University.
Edmonds, J., & Karp, R. (1972). Theoretical improvements in algorithmic efficiency for network flow problems. *Journal of the ACM, 19*(2), 248–264.
Goldberg, A., & Rao, S. (1998). Beyond the flow decomposition barrier. *Journal of the ACM, 45*(5), 783–797.
Hochbaum, D. S., & Rao, X. Network flow methods for the minimum covariates imbalance problem. arXiv preprint arXiv:2007.06828
Iri, M. (1960). A new method of solving transportation-network problems. *Journal of the Operations Research Society of Japan, 3*(1), 2.
Jewell, W. (1958). Optimal flows through networks. *Technical report*. Operations Research Center, MIT.
Kim, A., & Eisen, S. (2020). The contribution of the observational research design to COVID-19 research. *The Lancet Rheumatology, 2*(11), e650–e652.
Nikolaev, A., Jacobson, S., Cho, W., Sauppe, J., & Sewell, E. (2013). Balance optimization subset selection (BOSS): An alternative approach for causal inference with observational data. *Operations Research, 61*(2), 398–412.
Rosenbaum, P. (1989). Optimal matching for observational studies. *Journal of the American Statistical Association, 84*(408), 1024–1032.
Rosenbaum, P., Ross, R., & Silber, J. (2007). Minimum distance matched sampling with fine balance in an observational study of treatment for ovarian cancer. *Journal of the American Statistical Association, 102*(477), 75–83.
Sauppe, J. (2015). *Balance optimization subset selection: A framework for causal inference with observational data*. University of Illinois at Urbana-Champaign Ph.D. thesis..
Sauppe, J., & Jacobson, S. (2017). The role of covariate balance in observational studies. *Naval Research Logistics, 64*(4), 323–344.
Sauppe, J., Jacobson, S., & Sewell, E. (2014). Complexity and approximation results for the balance optimization subset selection model for causal inference in observational studies. *INFORMS Journal on Computing, 26*(3), 547–566.
Stuart, E. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science, 25*(1), 1–21.
Tomizawa, N. (1971). On some techniques useful for solution of transportation network problems. *Networks, 1*(2), 173–194.
Yang, D., Small, D., Silber, J., & Rosenbaum, P. (2012). Optimal matching with minimal deviation from fine balance in a study of obesity and surgical outcomes. *Biometrics, 68*(2), 628–636.
Zubizarreta, J. (2012). Using mixed integer programming for matching in an observational study of kidney failure after surgery. *Journal of the American Statistical Association, 107*(500), 1360–1371.