Optimal tradeoffs in matched designs comparing US-trained and internationally-trained surgeons

Sam Pimentel, Department of Statistics, UC Berkeley

Rachel Kelz, Department of Surgery, Perelman School of Medicine, University of Pennsylvania

Partially supported by the Department of Defense (DoD) through the National Defense Science & Engineering Graduate Fellowship (NDSEG) Program

June 29, 2021

Based on

Application background: Zaheer, S., Pimentel, S.D., Simmons, K.D., Kuo, L.E.Y, Datta, J., Williams, N., Fraker, D.L., and Kelz, R.R. (2017). Comparing international and United States undergraduate medical education and surgical outcomes using a refined balance methodology. *Annals of Surgery*, 265 (5), 916-922.

Methods: Pimentel, S. D., & Kelz, R.R. (2020). Optimal tradeoffs in matched designs comparing US-trained and internationally-trained surgeons. *Journal of the American Statistical Association* 115 (532), 1675-1688.

IMGs vs. USMGs

- 15% of surgeons practicing in the United States are international medical graduates (IMGs).
- IMGs take a US certification exam and can then compete for US postgraduate training programs, or residencies.
- Compared to US medical graduates (USMGs), IMGs are much less likely to be placed in their desired training program, and are more likely to split their training among multiple programs (Datta et al., 2015).

- Does this difference in training impact patient outcomes?
- In particular, does receiving surgery from an IMG (vs. USMG) change risk of death in 30 days?

Randomized trial approach

- An **ideal** study design: randomized trial
- A coin is flipped for each pair of consecutive surgical operations to be performed:
  - If heads, operation 1 performed by an IMG and operation 2 performed by a USMG
  - If tails, vice versa
- Randomization ensures comparability of (large) groups on all pre-treatment covariates, measured and unmeasured.
- Measure "IMG effect" by comparing 30-day mortality rates of the two groups

Reality: observational study

- Our data: medical claims for inpatient general surgeries in greater Orlando, 2008-2011 (Florida Agency for Health Care Administration).
- Linked to AMA Masterfile (surgeon information).
- 80,000 records, $\sim$20,000 of the operations performed by IMGs.
- Operations are grouped in several hundred hospitals.
- Many measured covariates:
  - Surgical procedure (30+ categories)
  - Patient comorbidities and demographics, admission source, etc.
  - Surgeon training and experience.

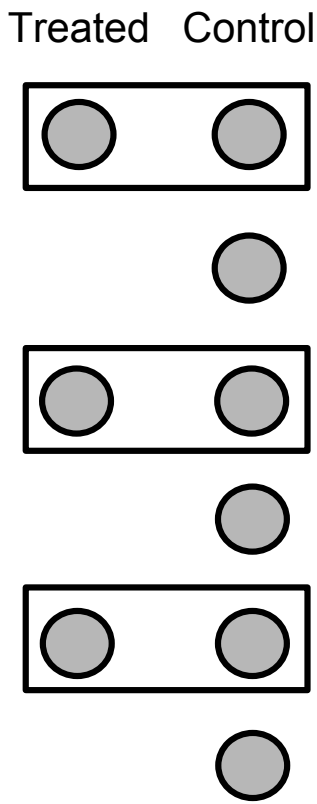
Surgeon ID	Hospital ID	Procedure	Age	Sex	ER	CHF	Cancer	Dementia	...
507	4341	Append.	67	Male	1	0	0	0	...
507	4341	Hernia Repair	78	Male	0	0	0	1	...
511	4482	L Knee Repl	77	Female	0	1	0	0	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱

*Data is artificial

Making a fair comparison

- Assignment to IMG or USMG is not random — in fact, they tend to practice in very different settings.
- Attributes related to mortality (**confounders**) are out of balance between groups: 58% of IMG surgeries are ER admissions, only 50% of USMG admissions.
- Difference in mortality rates is a mix of true effect and an effect due to confounders (bias).
- How to remove confounding?

Matching (Rosenbaum and Rubin, 1985)



- Pair each IMG operation (treated) to a similar USMG operation (control).
- Analyze paired outcomes as though they came from a randomized experiment.
- Requires absence of unmeasured confounders.
- Requires pairs to be homogenous on either propensity for treatment (Rosenbaum and Rubin, 1983) or expected potential outcome (Hansen, 2008).
- Compared to regression-based methods, matching keeps analysis and design separate and makes quality of confounding removal very transparent.

What makes a matched design good?

1. Balanced covariates: distribution of a covariate among matched IMG operations same as among matched USMG operations.
   - Similar to balance in a randomized experiment.
2. Close pairs: individual IMG-USMG operation pairs should have similar values of covariates (closeness on a covariate distance)
   - More stringent than marginal balance.
   - Reduces heterogeneity in responses and sensitivity to bias.
3. Large matched sample:
   - More power (all else being equal)
   - Using all/most of IMG operations (smaller group) makes estimand interpretable as effect of treatment on treated

- Existing methods optimize these criteria directly in some order.
- Different measures of quality may conflict.

Tradeoffs in IMG-USMG study

- One especially challenging variable: **surgical experience**.
- 25% of IMG operations performed by surgeon with 30+ years, only 4% of USMG operations were.
- Off-the-shelf approaches that strictly optimize some individual objective didn't fix this:

	Imbalance on experience (std. diff.)	Matched pairs retained
Sparse optimal matching	$\approx$0.25 (2.4 years)	$\approx$ 20,000
Exact match on experience	0	$\approx$ 6,000

- **Problem**: strict optimization of one measure of quality may force bad performance on others.
- How to produce and consider "intermediate" solutions that make tradeoffs differently?

Tradeoff analysis via multiobjective optimization

- **Our contribution**: multiobjective optimization methods to explore tradeoffs between design goals.

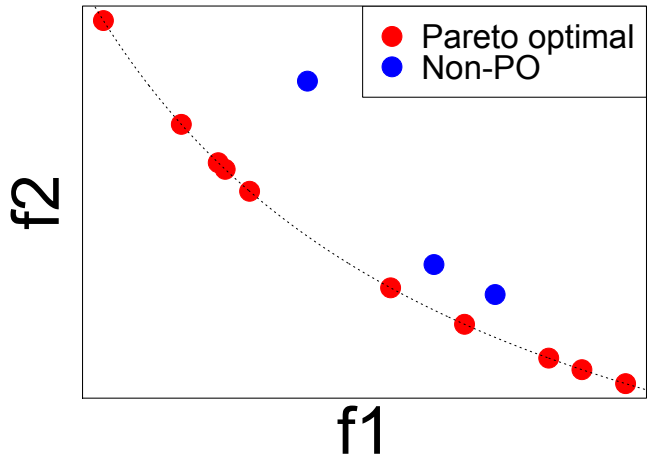
- How to precisely define class of interesting intermediate solutions? Use Pareto optimality.
- How to compute and explore these solutions? Iterative approach based on efficient single-objective problem.
- Application: IMG-USMG study.

Pareto optimality

- Look at **Pareto optimal** or non-dominated, solutions

Definition 1: A solution $\mathbf{x}$ to an optimization problem is Pareto optimal for objective functions $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$ if there exists no dominating $\mathbf{x}'$, i.e. a solution $\mathbf{x}'$ such that

$f_i(\mathbf{x}') \leq f_i(\mathbf{x})$ for all i and $f_i(\mathbf{x}') < f_i(\mathbf{x})$ for some $i \in \{1, 2\}$

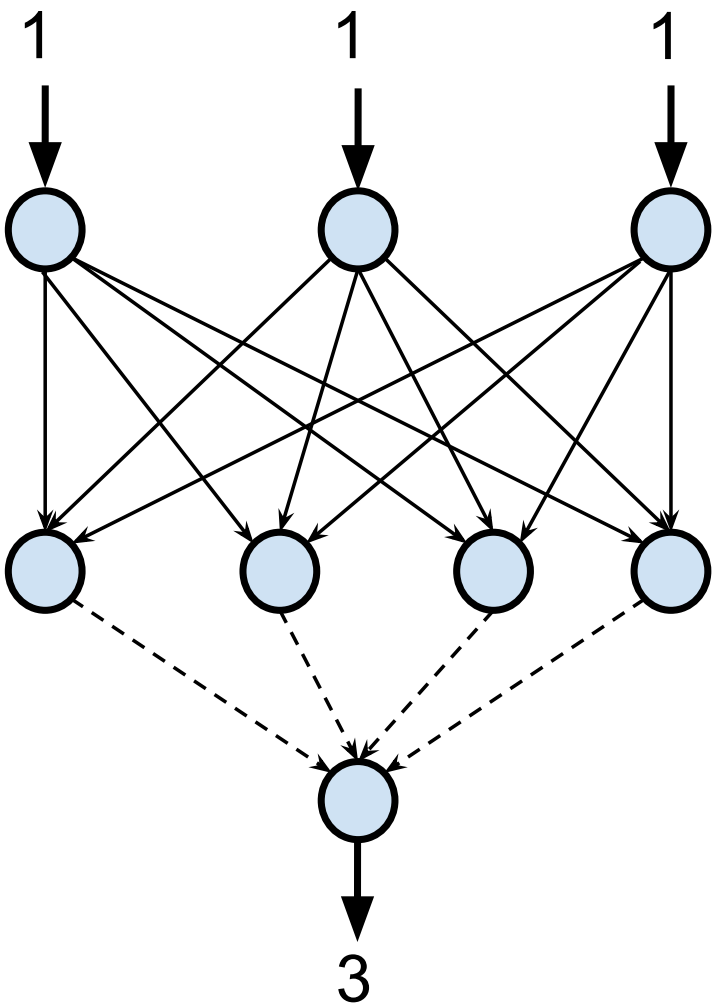


- **Our goal:** generate and explore Pareto optimal solutions to an optimal matching problem with two general objective functions.

Related methods

- Optimal subset matching (Rosenbaum, 2012).
  - Provides a method for continuously varying priority between close pairwise distances and sample size.
  - Proves Pareto optimality of solutions.
- Balance-sample size frontier (King et al., 2017).
  - Tools for exploring tradeoff between matched sample size and covariate imbalance.
  - Restricted to matching with replacement.
- We provide a more general framework based on viewing matching as a minimum-cost network flow optimization problem.

Review: minimum-cost network flow optimization



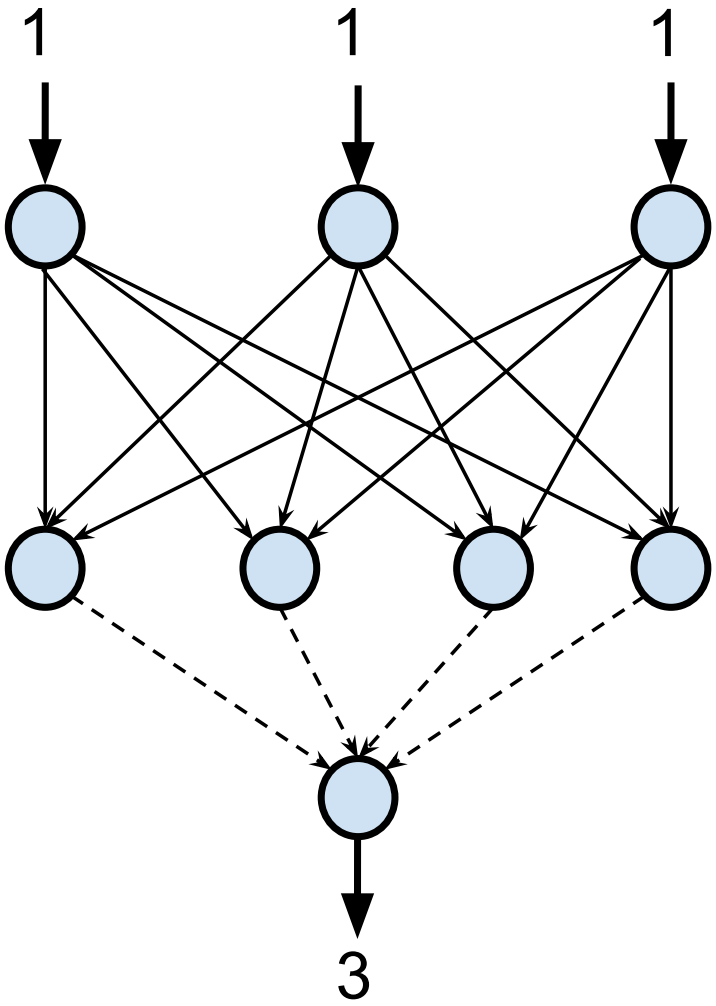
- Nodes $\mathcal{N}$ edges $\mathcal{E}$ with costs ψ_e and upper capacities.
- Decision variables x_e choose flow over edges.
- Solve:

$$\min \sum_{e \in \mathcal{E}} \psi_e x_e$$

$$\text{s.t.} \quad \mathbf{x} \in \mathcal{F}$$

- $\mathcal{F}$ contains all $\mathbf{x}$ such that $\mathbf{x}$ is nonnegative, edge capacities are satisfied, and flow is preserved.

Review: minimum-cost network flow optimization



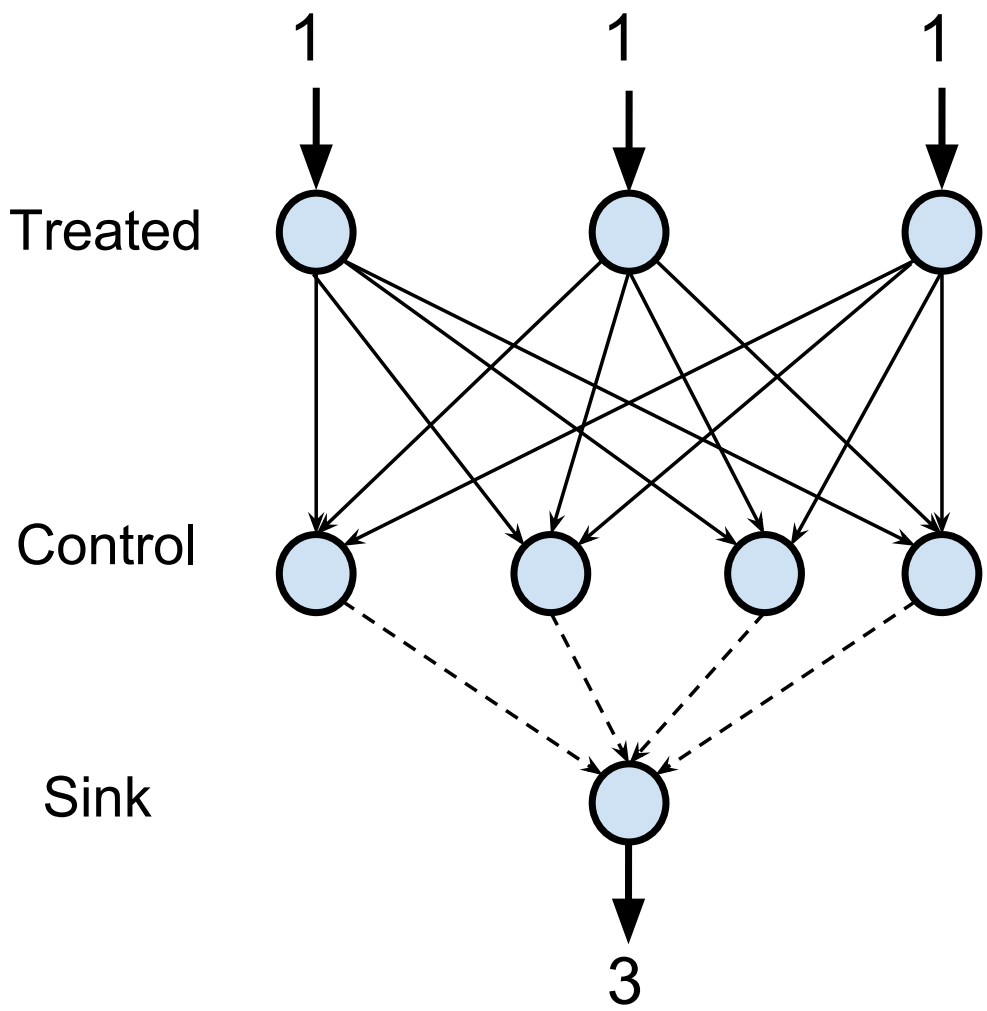
- Solve:

$$\min \sum_{e \in \mathcal{E}} \psi_e x_e$$

$$\text{s.t.} \quad \mathbf{x} \in \mathcal{F}$$

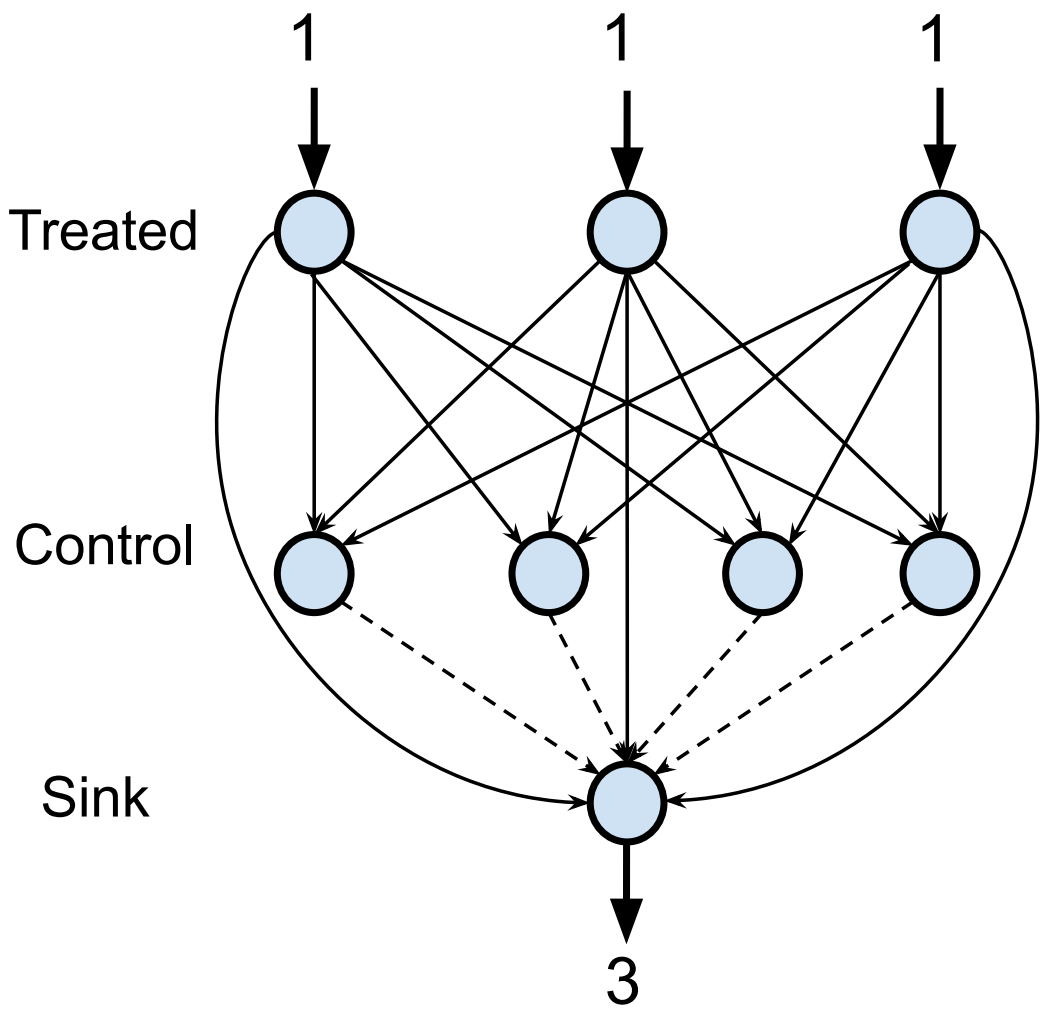
- Integer solutions can be obtained at low computational cost via network simplex algorithm.
- Additional side constraints destroy this property (Bertsekas, 1998, §8).

Review: matching as minimum-cost network flow



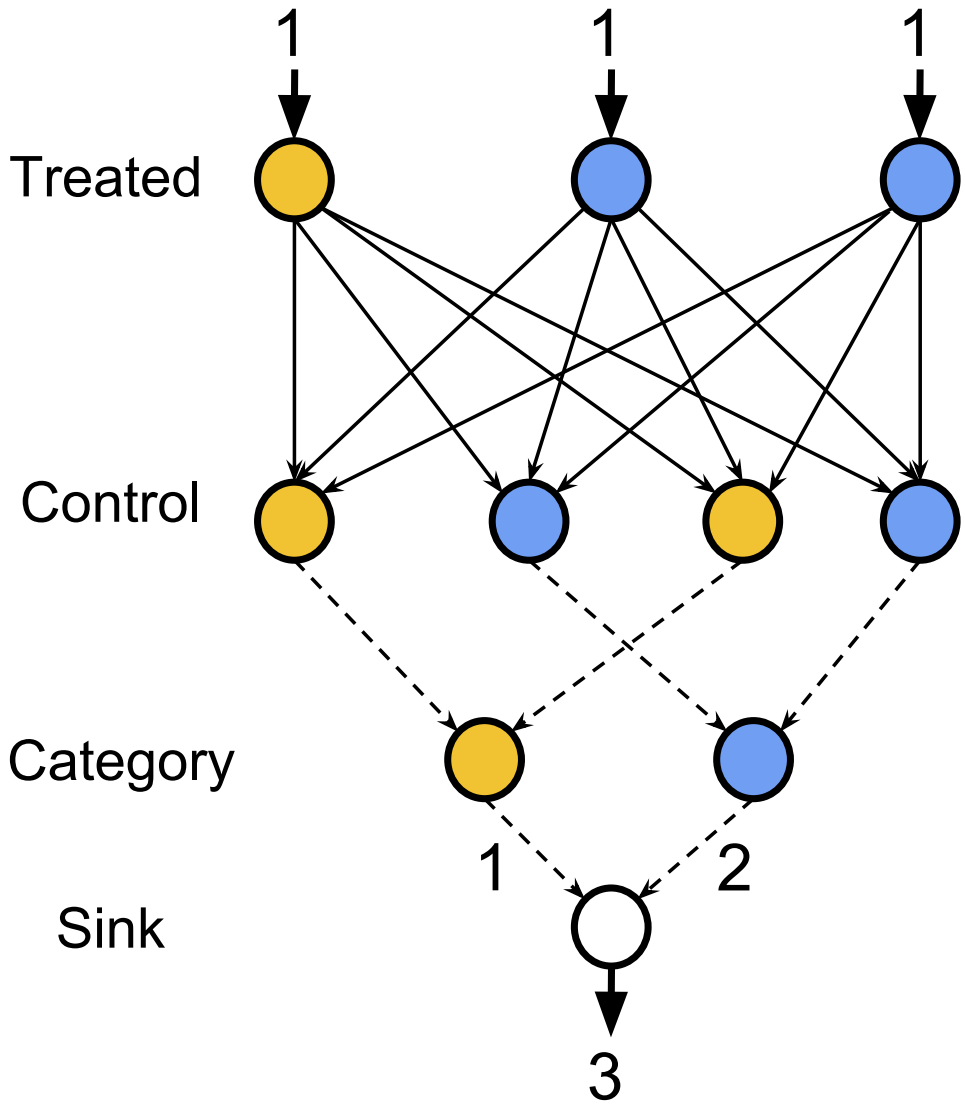
Fixed sample size, minimize within-pair differences

Review: matching as minimum-cost network flow



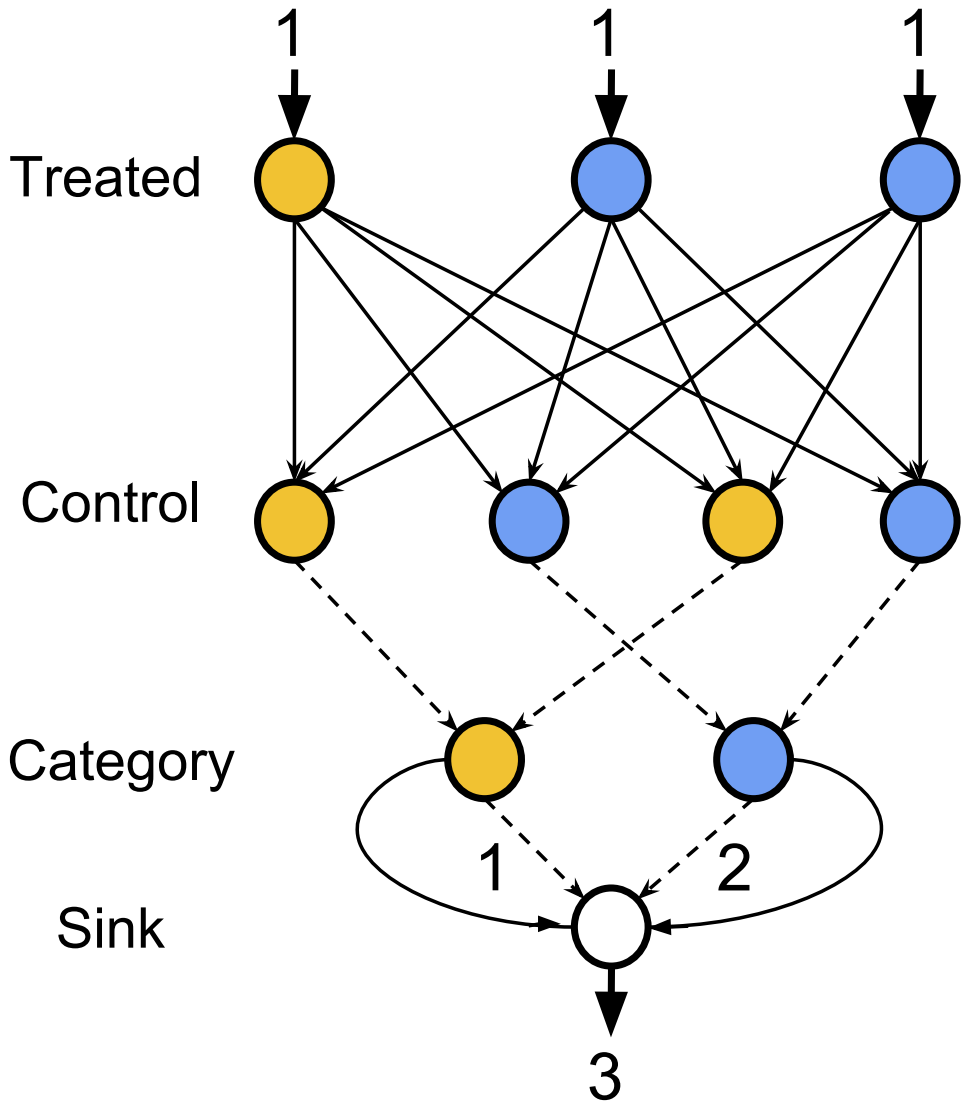
Variable sample size, minimize within-pair differences

Review: matching as minimum-cost network flow



Fixed sample size, constrain overall covariate imbalance

Review: matching as minimum-cost network flow



Fixed sample size, minimize overall covariate imbalance

Network flow with two objective functions

- Consider a matching problem represented as network flow with decision variables $\mathbf{x}$ and constraint set $\mathcal{F}$.
- Define **two** linear objective functions to minimize:

$$f_1(\mathbf{x}) = \sum_{e \in \mathcal{E}} \psi_e x_e \qquad \text{and} \qquad f_2(\mathbf{x}) = \sum_{e \in \mathcal{E}} \gamma_e x_e$$

- Wish to find Pareto-optimal solutions with distinct values (i.e. distinct Pareto points) for these two objectives

Producing Pareto optimal solutions

Directly-constrained problem $\mathcal{Q}(a)$: given a problem with network flow constraints $\mathcal{F}$ and edge set $\mathcal{E}$,

$$\min f_1(\mathbf{x}) \qquad \text{s.t.} \qquad \mathbf{x} \in \mathcal{F}, \quad f_2(\mathbf{x}) \leq a, \quad \mathbf{x} \in \mathbb{Z}^{|\mathcal{E}|}$$

Theorem 1:

1. Any Pareto optimal solution $\mathbf{x}'$ for $\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), f_2(\mathbf{x}))$ must also be optimal for $\mathcal{Q}(f_2(\mathbf{x}'))$.
2. All optimal solutions of $\mathcal{Q}(f_2(\mathbf{x}'))$ are also Pareto optimal with identical values of $\mathbf{f}$.

- Pro: Guaranteed recovery of all Pareto optimal solutions
- Con: integer constraint makes computation intractable for large problems

Producing Pareto optimal solutions

Penalized problem $\mathcal{P}(\rho)$: given a problem with network flow constraints $\mathcal{F}$ and edge set $\mathcal{E}$,

$$\min f_1(\mathbf{x}) + \rho f_2(\mathbf{x}) \qquad \text{s.t.} \qquad \mathbf{x} \in \mathcal{F}$$

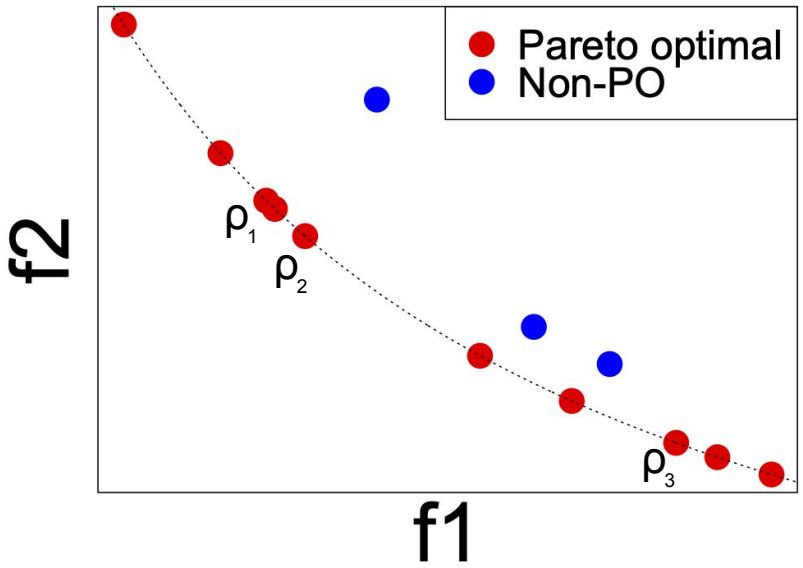
Theorem 2: Suppose $\mathbf{x}^*$ is an integer-valued optimal solution for the penalized problem $\mathcal{P}(\rho)$. Then $\mathbf{x}^*$ is Pareto optimal and is also optimal for the directly constrained problem $\mathcal{Q}(a^*)$ where

$$a^* = f_2(\mathbf{x}^*)$$

- Pros: Computationally tractable, any solution is Pareto optimal, interpretation via problem $\mathcal{Q}$ is available
- Con: doesn't guarantee recovery of every single Pareto optimal solution

From individual solutions to broader structure

- We don't usually want to look at all Pareto solutions:
  - Some are virtually identical to each other.
  - Many are not acceptable in practice.
- How to efficiently explore space?

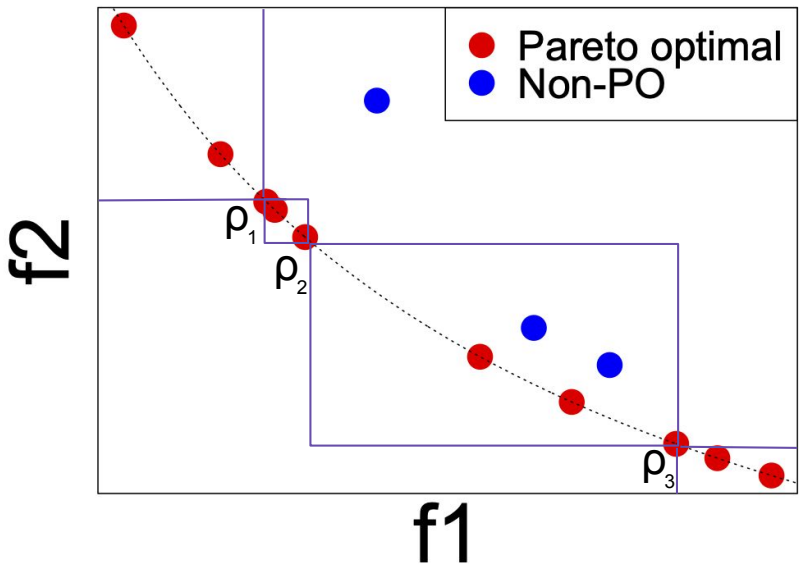


$$\min f_1(\mathbf{x}) + \rho f_2(\mathbf{x})$$
$$\text{s.t.} \quad \mathbf{x} \in \mathcal{F}$$

- Natural ordering: if we line up distinct Pareto optimal points so they are increasing in one objective, they are decreasing in the other (and also monotone in values of ρ used to produce them).

From individual solutions to broader structure

- We don't usually want to look at all Pareto solutions:
  - Some are virtually identical to each other.
  - Many are not acceptable in practice.
- How to efficiently explore space?



$$\min f_1(\mathbf{x}) + \rho f_2(\mathbf{x})$$
$$\text{s.t.} \quad \mathbf{x} \in \mathcal{F}$$

- Furthermore, any collection of Pareto points on the curve gives bounds on the locations of remaining points.

Solving in practice

- Search algorithm for generating representative family of Pareto optimal solutions:
  1. Solve penalized problem for a large value of ρ and a small value of ρ.
  2. Solve for a grid of a few intermediate ρ-values.
  3. For any neighboring pair of ρ-values for which points "between" would be interesting, repeat.
- Search can focus on areas where tradeoffs are important
- Computation is relatively efficient because each step is a standard network flow problem
- Implementation: forthcoming R package `MultiObjMatch` (with Shichao Han).

Design goals for IMG-USMG study

- Wish to balance surgical experience (in years) — minimize total variation imbalance on experience deciles between groups
- Wish to form close pairs on surgical procedure, Elixhauser index, age, sex, and emergency room status — minimize sum of pairwise robust Mahalanobis distances calculated using these variables
- Wish to retain as many IMG operations as possible — minimize number of IMG operations dropped

- Set up a network flow problem with an objective function for each
- Tradeoff #1: close pairing vs. balance
- Tradeoff #2: balance vs. sample size

Tradeoff #1: Close pairing vs. balance

Table: Summary of solutions to penalized problem for IMG-USMG study. All matches have 18,888 pairs.

	Attention to balance (vs. pair matching)				
	Very low ⟵			⟶ Very high	
Imbalance penalty ρ	0.01	1	5	8	50
Match label	A	B	C	D	E
f_1 (Sum of pair distances)	19241	19412	20960	22205	23242
f_2 (total imbalance)	4985	3829	3319	3112	3018
Prop. paired exactly on proc. type	0.92	0.92	0.91	0.91	0.90
Prop. paired on coarse Elix. index	0.96	0.96	0.95	0.95	0.95
Prop. paired on emergency admission	0.98	0.98	0.98	0.98	0.98
Std. diff., experience	0.47	0.32	0.26	0.23	0.22

- Good news: goals don't conflict much, get improved balance for free
- Bad news: balance is still too poor even when prioritized strictly

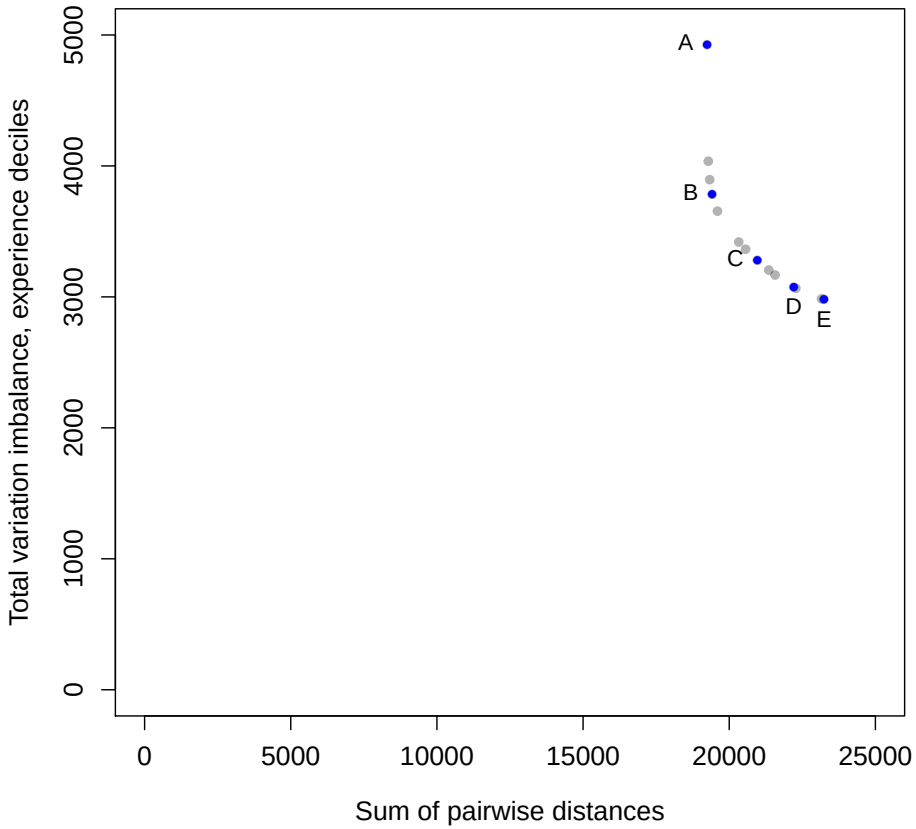
Tradeoff #2: Balance vs. sample size

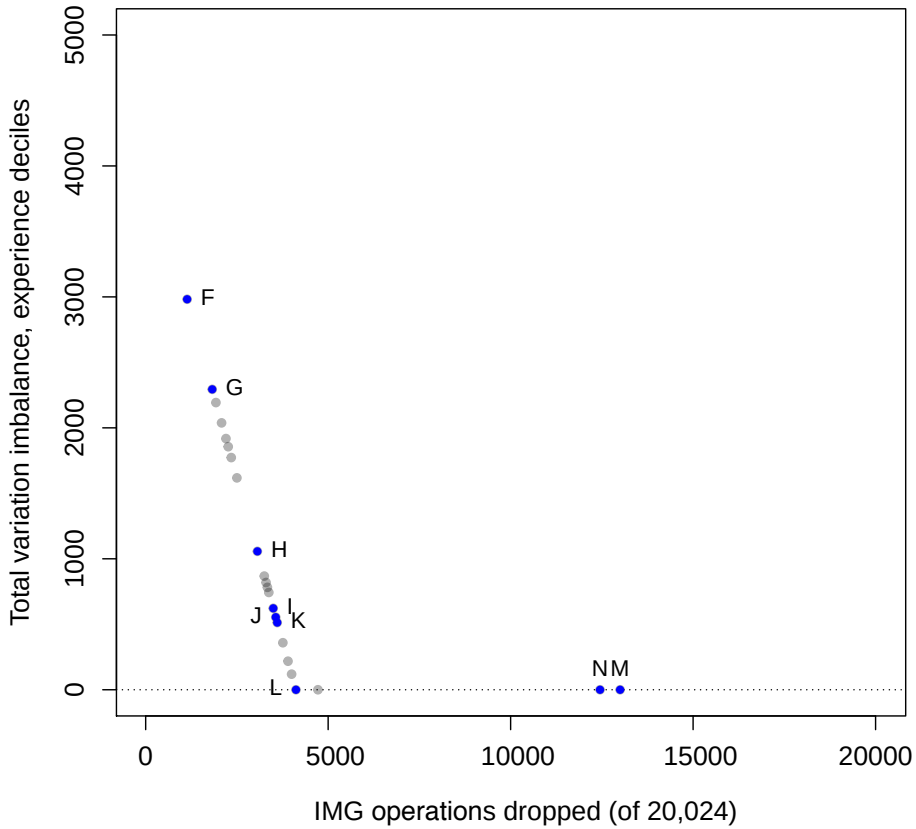
Table: Summary of solutions to penalized problem for IMG-USMG study.

	Attention to balance (vs. sample size)			
	Very low	$\longleftarrow$	$\longrightarrow$	Very high
Exclusion penalty ρ	500	449.9	448.7	0.01
Match label	F	H	K	M
f_1 (total imbalance)	2982	1057	513	0
f_2 (IMG operations dropped)	1136	3061	3605	13000
No. of pairs in matched sample	18,888	16,963	16,419	7,024
Std. diff., experience	0.23	0.05	0.00	0.00

- Much bigger tradeoff
- Standardized difference on experience stops improving around Match K with 80% of IMG operations retained

Tradeoffs visualized





Results

- Select Match K for analysis, with 16,419 matched pairs
- Mortality rate among IMG patients is 1.5%, among USMG patients it's 1.7%
- McNemar's test: assumes treatment is equally likely for paired individuals within matched pairs, permute labels to get reference distribution.
- Cannot reject null of zero effect (p= 0.14)

Equivalence testing

- Can we reject the possibility that any true IMG effect is large?
- Attributable effect $A =$ **number of lives saved** if USMGs had done all surgeries (Rosenbaum, 2002a)
- Choose an effect threshold A and see if we can reject effects above the threshold.
- Two one-sided tests: against $A \leq -\iota$ (IMGs better) and against $A \geq \iota$ (IMGs worse)
- If both reject, we have shown "equivalence" (effects can be no larger than $|\iota|$)
- We set $\iota = 162$ or change in mortality rate of $162/16419 \approx 1\%$ (1/4 of age-75 mortality rate).

Equivalence testing

- Results (assuming no unobserved confounders):

Null hypothesis	$A = 0$	$A \leq -162$	$A \geq 162$
Modified McNemar p-value	0.14	$< 10^{-10}$	$< 10^{-10}$

- For either choice of effect threshold, we reject both inferiority and superiority of size 162 deaths or more
- Type I error rate is controlled in the 2 one-sided level-α tests, as well as the two-sided test
- Example of **three-sided testing** (Goeman et al., *Stat. in Med.* 2010)

Review: sensitivity analysis

- What if paired operations do not have identical propensities for treatment due to unobserved confounders? Test will be wrong.
- Instead of assuming treatment probabilites π_i and π_j for matched i and j are equal, assume

$$\Gamma^{-1} \leq \frac{\pi_i/(1-\pi_i)}{\pi_j/(1-\pi_j)} \leq \Gamma$$

- For $\Gamma > 1$, a family of possible true randomization distributions exist
- Following (Rosenbaum, 2002b, §4), we can calculate the worst-case (largest) p-value over this family of distributions
- Sensitivity analysis: repeat for larger and larger Γ, report value at which results are overturned

Results, continued

- Sensitivity analysis for IMG-USMG study: might a true large effect be masked by unmeasured bias?
- Repeatedly test all three hypotheses ($A = 0$, $A \leq -162$, $A \geq 162$) at higher and higher values of Γ.
- Test still rejects superiority/inferiority at $\Gamma = 1.7$, no longer rejects $A \leq -162$ (IMGs much better) at $\Gamma = 1.8$.
- **Moderate level of confounding** needed: roughly a (binary) unobserved confounder that increases the odds of treatment by an IMG by fivefold and simultaneously multiplies the odds of death by 2.5.

Discussion: applications and extensions

- Forms of penalized and directly constrained problems appear many places in the matching literature – our results add deeper interpretation.
- Possible extension: choosing a match from the Pareto optimal set based on a model for inference.
  - e.g. look at a tradeoff between pairing closely on a propensity score and on a prognostic/outcome risk score.
  - Project expected mean-squared error of the estimator at each Pareto point under a model, choose the minimizing case.
  - Alternatively, compute power of a sensitivity analysis for each point to enforce robustness to unmeasured bias.
- Multidimensional tradeoffs — theory is straightforward, practical exploration of space of solutions is not.

Selected References

Bertsekas, D. P. (1998), *Network optimization: continuous and discrete models*, Belmont, MA: Athena Scientific.

Datta, J., Hoffman, R. L., Kelz, R. R., Morris, J. B., and Williams, N. N. (2015), "Preliminary residency in general surgery: comparative outcomes of International and US Medical Graduates," *The American Surgeon*, 81, 219–221.

Hansen, B. B. (2008), "The prognostic analogue of the propensity score," *Biometrika*, 481–488.

King, G., Lucas, C., and Nielsen, R. A. (2017), "The balance-sample size frontier in matching methods for causal inference," *American Journal of Political Science*, 61, 473–489.

Rosenbaum, P. R. (2002a), "Attributing effects to treatment in matched observational studies," *Journal of the American Statistical Association*, 97, 183–192.

— (2002b), *Observational Studies*, New York, NY: Springer.

— (2012), "Optimal Matching of an Optimally Chosen Subset in Observational Studies," *Journal of Computational and Graphical Statistics*, 21, 57–71.

Rosenbaum, P. R. and Rubin, D. B. (1983), "The central role of the propensity score in observational studies for causal effects," *Biometrika*, 70, 41–55.

— (1985), "Constructing a control group using multivariate matched sampling methods that incorporate the propensity score," *The American Statistician*, 39, 33–38.