Aligning with Heterogeneous Preferences for Kidney Exchange

By Rachel Freedman
Center for Human-Compatible AI (CHAI)
UC Berkeley
Kidney Exchanges
Preference Aggregation
Methodology and Simulations
Kidney Exchanges
Preference Aggregation
Methodology and Simulations
Thanks to AI, paired kidney donations in the US are on the rise

12% of living donations that came from paired donors

Source: Quartz, How AI changed organ donation in the US
Kidney Exchanges
Preference Aggregation
Methodology and Simulations
Value Alignment as Preference Aggregation
Value Alignment as Preference Aggregation

- “Multi-single Delegation” from ARCHES: AI Research Considerations for Human Existential Safety (Andrew Critch and David Krueger, 2020)
Value Alignment as Preference Aggregation

- “Multi-single Delegation” from ARCHES: AI Research Considerations for Human Existential Safety (Andrew Critch and David Krueger, 2020)
- Examples:
  - kidney allocation (ex. Adapting a Kidney Exchange Algorithm to Align with Human Values, Freedman et al. 2020)
  - self-driving cars (ex. The Social Dilemma of Autonomous Vehicles, Bonnefon et al. 2016)
Individual Preferences: Pairwise Comparisons

Patient A is 70 years old, has 1 alcoholic drink per month, and has no other major health problems.

Patient B is 30 years old, has 5 alcoholic drinks per day, and has skin cancer in remission.

Who should get the kidney?

*Survey data graciously shared by the Duke Moral AI team.
Kidney Exchanges
Preference Aggregation
Methodology and Simulations
Model: BLP*

\[ v_1 = \text{BLP}_1 \]

\[ v_2 = \text{BLP}_2 \]

\[ w = \text{BLP}_3(\text{patient}_1) \]

\[ w = \text{BLP}_3(\text{patient}_2) \]

Algorithm: Modified LP

\[
\text{max } \sum_{c \in C(L)} x_c \quad \text{s.t.} \quad \sum_{c : v \in c} x_c \leq 1 \quad \forall v \in V.
\]

Our Modification:

\[
\text{max } \sum_{c \in C(L)} \left[ \sum_{(u,v) \in c} w_{\text{BLP}}(u,v) \right] x_c \\
\text{s.t.} \quad \sum_{c : v \in c} x_c \leq 1 \quad \forall v \in V \\
\sum_{c \in C(L)} |c| x_c \geq Q
\]
I built a simulation...

public class VariationDriver {

    // Experimental conditions
    public enum Condition {
        EQUAL_WEIGHTS, // All edges have weight 1.0
        PATIENT_WEIGHTS, // Edge weights depend on receiving patient only (BT model)
        DONOR_PATIENT_WEIGHTS // Edge weights depend on donor and receiving patient (BLP model)
    }

    // Probabilities generated based on a match frequency of 1 day
    static final int CYCLE_CAP = 3;
    static final int EXPECTED_PAIRS = 4;
    static final int ITERATIONS = 365*5;
    static final int NUM_RUNS = 50;
    static final double DEATH = 0.000580725433182381168050643691;

    public static void runExperiments(){
        String expID = Long.toString(System.currentTimeMillis());
        runSimulationWithEqualWeights(expID);
        runSimulationWithPatientWeights(expID);
        runSimulationWithDonorPatientWeights(expID);
    }
}
Considering Heterogeneity Improves Average Rankings
Most Desirable Profiles Still Prioritized
Acknowledgements

Thanks to the Duke Moral AI team for the human preference data.

Thanks to Professor John Dickerson for an earlier version of the simulation.

And thanks to Yunhao Huang for help on the project!
Thank you for listening!
Questions?