15.093 Final Project: Optimal Dating

Daniel Chung
Joseph Lu

Fall 2022

"If you're looking for a bad boy, look no further because I'm bad at everything"
- Dating App User

1. Problem

The formation of social groups defines human history. From hunting packs to early civilizations, this process has allowed humans to survive, thrive, and achieve far more than any individual could alone. Why do these groups form? Who gets to decide inclusion or exclusion from a group? The answer, until very recently, was humans themselves. Perfection, however, is outside of the human condition, leaving room for improvement that technology is uniquely suited to fulfill.

Today, algorithms for group formation are rapidly changing the ancient paradigm, and they are already ubiquitous as well. Facebook connects friend groups. LinkedIn builds professional networks. One method, however, stands out for its uniqueness of formation, dependence on individual preference, and volatility of group outcome quality–dating apps. Unfortunately, modern dating apps use heuristics to form their pairings based on "likes", "swipes", etc. which is suboptimal towards the intended result of forming good pairs. Therefore, in this project, we attempt to solve the problem of forming dating pairs (and more!) out of potential dating pools using MIO and achieve provable optimality for dating.

2. Motivation

Why use optimization for dating? Simply put, advances in matchmaking are critical to keeping up with the exponentially greater size of today's dating pool. While for most of history potential dates were limited to one's immediate community, today one can create a dating account and view 75 million profiles on Tinder alone.[1] Heuristics can find the needle in a haystack of several dozen, but they cannot hope to succeed in a haystack of millions. Why date one person knowing that there could be a better match still unrealized? Why commit when there is no guarantee that commitment is the best of all available choices? Choosing a marriage partner is regarded as perhaps the most important decision a person will make in his or her life, and dating is the critical first step. Why leave to chance and human error what optimization can achieve with a guarantee? Dating is too important a process and optimization is too apt a solution not to apply. This is precisely what we accomplish with our methodology.

3. Data

Our data comes from the results of a speed dating experiment performed by Columbia Business School.[2] 552 individuals, each indexed by a unique id, participated in speed dating rounds of 4 minutes each, self-reporting both personal details and how they felt about the people they saw.

For our purposes, we focused on participant reports of how likely they would be to go on another date with the people they saw, which ranged from 1 to 10. A 10 indicates they would very likely go on a date

with that person again. Each participant also reported how important various factors are to deciding a potential mate. For our model, we utilized responses to how important religion is in a dating relationship, with 1 being not important at all and 10 being non-negotiable. Data on participant religious affiliation did not exist, so we synthesized it by assigning people religions at random from the distribution of the 10 most common religions according to the latest U.S. demographic data.[3]

Before plugging the data into our optimization formulation, we needed to perform three cleaning steps. The first was to normalize people's preference scores, and the second was to impute missing preference scores, since each participant only evaluated a subset of the other participants during the speed-dating rounds and thus created a sparse matrix of scores. We imputed missing scores by drawing from a standard normal distribution, as we had normalized these scores in step 1. The third step was to convert the importance of religion from a categorical variable to a binary one, since our formulation requires it to be. To do this we set a cutoff at 6 such that if the importance of religion was 6 or greater for a participant, the binary equivalent was 1.

4. Approach

Our objective is, given a dating pool, to create as many pairs as possible among people who both had high preference scores for each other, penalizing any instances of unpaired singles. Our decision is which people to assign to a specific pair. Our data, as discussed, consisted of these preference scores as well as whether religion is important to each person and what religion each person affiliated with. Given this framework, we approached the optimal dating problem in several variations.

1. Make optimal speed-dating pairs among all members of the dataset so as to maximize the collective utility from being paired with someone to whom the individual has assigned a high "likeability" score, subject to the constraints of 1-1 pairings, respecting individual preferences for other features.
2. Improve on the pair model by allowing up to 3 people per group subject to new constraints such as having diverse groups in terms of gender and religion.
3. Throw out the assumption that likeability scores are correct, and introduce uncertainty based on the scores that have been given out to all other members of the group.

4.1. Identifying Baselines

To calibrate our results, we investigated three baseline approaches to matchmaking to see how they perform within our objective function. We provide a qualitative assessment in this section, as the quantitative results can be found in section 5.

4.1.1 No Pairing

The first and poorest benchmark for matchmaking is to make no matches at all. This results in a negative objective value because every participant is an unpaired single, and every unpaired single adds a penalty term to the objective. It would be a tragedy, intuitively, to want to date and not be paired with anyone at all. The impact on global utility, as expected, is devastating.

4.1.2 Random Pairing

The second benchmark for matchmaking is to pair at random, selecting a random person of gender 0, or gender 1, and assigning them to each other. This results in an objective value close to zero. Intuitively this makes sense since some random matches will be between people who like each other by chance while others will be between people who are not interested in each other at all, balancing out the resulting objective value.

4.1.3 Ordered Match Heuristic

To simulate the traditional, heuristic-based process of matchmaking, we implemented the third benchmark of sorting participants by order of how much other people preferred them, assigning people to each other based on that preference order. As expected, this boosted the objective value over the other benchmarks because "more desirable" people were being matched together. However, this heuristic is greedy because it simply assigns in order of desirability, overlooking potential matches between people who liked each other more but who were further apart in desirability order. Our optimization approach moves beyond this weakness because it optimizes over global preferences, but to see this edge we would need to first implement it, which is precisely what we did next.

4.2. Initial Formulation with Traditional Couples

For our initial formulation, we optimize the formation of heterosexual couples matched 1:1 with constraints on religion. For consistency, we always set penalty rho= -0.5 for any unpaired singles. This variant of the dating problem requires additional constraints on top of the usual assignment problem constraints i.e. each person is matched with one other person, which we will discuss below.

4.2.1 Gender

We chose not to treat the genders as though they were an assignment problem because we expected to expand our model beyond 2 person groups or allow LGBTQ+ couples, meaning gender_0 could be matched with other gender_0. Therefore we use the binary variables one-hot encoding each gender to make the constraint such that the two genders must be different if the couple is matched. In later steps, we can expand this to enforce constraints on gender in groups of >2, or remove this constraint to open matching with anyone (say if we moved away from the realm of dating).

4.2.2 Religion

The religion constraint is based on individual preferences stating the importance of religion in considering a date, represented as a binary variable (1 if important, 0 if not important). Our constraint is therefore that if either member of a pairing considers religion important, then to be matched, they must share the same religion, otherwise, religions are free. This constraint was also constructed in such a way that it could be easily generalized to other categorical constraints such as age group, race, and major, all of which were also in the dataset.

4.3. Initial Formulation with Robustness

Because participants reported preference scores after only a 4-minute speed date, there is uncertainty in the data regarding one's true feelings about another person. A bad first impression can belie great compatibility, while a good first impression can belie just the opposite. Worst-case uncertainty in our context is the latter case, where one's true preference for another person turns out to be far lower than what was initially self-reported from the speed date.

In our model (*See figure 1*), we assume uncertainty for preference scores is gaussian since first impressions can skew either positive or negative with equal probability. We implement robustness by subtracting one standard deviation from each original preference score, a "worst case" scenario in which preference scores are inflated by one standard deviation by human bias. We chose to robustify for one standard deviation because a gaussian distribution has infinite support, meaning robustifying against the entire uncertainty set would mean robustifying against normalized preference scores of negative and positive infinity. A cutoff is necessary. Using a cutoff of one standard deviation, we robustify the model against over

68% of possible uncertainties in preference scores, a finite support that can handle the majority of uncertainties.

4.4 Formulation with Closed Polyamory

We define closed polyamory as the following: given a maximum group size, every person within the group is connected to every other person. This manifests itself primarily in the objective function, where we must consider how every member feels about every other member. The closed polyamory also required that we change the penalty to appropriately punish groups of 1 and 2.

In the process of running the optimization we found that the number of constraints results in an exponential time increase per additional person in the dating pool so we had to limit our testing to only a dating pool of 110 individuals. In order to obtain a fair comparison with the other formulations which work for up to 550 individuals, we assume that in practical usage we would run the group formation for each 110 in order, and take a simple sum over the 5 separate groups. This will underestimate the actual utility gained from running the optimization over all 550 at once, but at least provides a lower bound.

4.5. Formulation with Open Polyamory

We define open polyamory as allowing each individual to pair with up to the N maximum, but not requiring that groups are closed. For instance, if person 1 pairs with persons 3 and 6, persons 3 and 6 do not necessarily need to be paired. This allows for various sizes and shapes of groups and gives an easier problem with fewer constraints and higher overall utility because every match is contributing a large amount of utility, compared with closed polyamory, where the closing link may or may not be contributing utility. The formulation also results in fewer constraints so we can solve the problem for all 550 original participants in a manner of seconds.

5. Key Findings

The resulting objective values for each model, including baselines, are compared below. It is important to note that the original objective function is a minimization problem, so the results are inverted in this visualization for the sake of intuition: better assignment models should have higher global utility.

First, we consider only 2 person pairs (*See Figure 3)* Given a penalty of -0.5 utility per unpaired person, it should be obvious that no pairings are the worst performing out of all models/heuristics. Random pairing is roughly at 0 utility, which also makes sense given that the majority of liking values were generated with mean 0. The ordered match performs significantly better than the previous 2 and is the most realistic baseline candidate. We see that using the heuristic we reach roughly 600 global utility. Finally, we come to our formulations, both the non-robust and robust versions. The version without robustness results in nearly a 60% increase in utility compared to the ordered match heuristic, and even when we account for uncertainty while making matches, the robust formulation returns a global utility slightly better than the heuristic.

Next, we look at groups of 3 (*See figure 4*) for open and closed polyamory. For open polyamory, it makes sense that the global utility will be increasing significantly beyond the 2 person groups because each person has double the matches as before. For closed polyamory, the closed constraint results in lower global utility as expected. Robustness comes at a steep price, but we can see that at least in the case of open polyamory, the total global utility still exceeds that of pairwise matching, showing the utility of open choice.

6. Impact

Our results speak for themselves: optimization has an edge in romantic match-making. The optimization we employ here is different from the approach of major companies like Tinder, which has a swiping system to gauge binary interest in a date. Our setup is more costly since it requires a set pool of participants to progress through speed dating rounds with each other, reporting how likely they would be to go on a follow-up date with each person they see. Our payoff, however, is more satisfactory since it ensures a globally optimal match for any given pair. As demonstrated in our results, optimization has an edge over random and heuristic-based match-making methods, increasing the amount of global happiness in the system. Not only that, there comes with every match a peace in knowing that pairing was guaranteed to be the best possible given every participant's preferences. Platforms like Tinder can implement this by offering speed dating trials that users can opt-in for, processing the data from these trials to optimally assign dates.

Dating is, of course, among the more exciting uses of this model, but it can be generalized to any situation in which two or more groups have preferences for each other and must be paired in a way that maximizes satisfaction. Other applications include project partner assignments, dormitory roommate assignments, and student-company matching for co-op programs.

7. Future Steps

7.1. Practical applications

We generated data about preference scores because in the original dataset not everybody met everybody else at a round, and the dataset also contained the results of multiple experiments taking place on different days. We would be interested in imputing the like scores based on partner attributes using a more rigorous matrix completion method and running the optimization that way.

The next step is to verify the quality of our solutions. Liking may not be the most important feature for a successful pairing, so if we could run an experiment to test the assignments given by the optimization results, we can confirm the real impact that our model is providing over some other assignment method.

7.2. Expanding the Objective

We consider expanding the objective function to include more than just preference scores. For instance, an individual may value attractiveness over likeability, so using individual weights for different partner attributes may result in a better, more realistic objective function.

7.3. Additional Constraints

We chose not to include some hard constraints that may play a part in forming pairs such as race and geographical location. We relaxed the former constraint because we did not feel comfortable deciding how to fix the importance of race per person, and we relaxed the second because the data was not available, but we are certain that it could be an important factor. In the future, if we were able to conduct the speed dating surveys ourselves, we could ask more questions to help inform these decisions to include or exclude features.

8. References

1. https://worldpopulationreview.com/country-rankings/tinder-users-by-country
2. https://www.kaggle.com/datasets/annavictoria/speed-dating-experiment
3. https://www.pewresearch.org/religion/2021/12/14/about-three-in-ten-u-s-adults-are-now-religiously-unaffiliated/

9. Appendix

1. **Data:**
   (a) $\ell_{i,j}$: how much person i likes person j
   (b) $\ell_{j,i}$: how much person j likes person i
   (c) $r_{i,f}$: binary, 1 if person i identifies with religion f, 0 otherwise
   (d) $r_{j,f}$: binary, 1 if person j identifies with religion f, 0 otherwise
   (e) $f \in \{agnostic, atheist, buddhist, catholic, hindu, jewish, mormon, muslim, protestant, unaffiliated\}$
   (f) m_i: binary, 1 if religion is an important dating factor for person i, 0 otherwise
   (g) m_j: binary, 1 if religion is an important dating factor for person j, 0 otherwise
   (h) N: number of people of gender 0
   (i) M: number of people of gender 1
2. **Decision Variables:**
   (a) $z_{i,j}$: binary, 1 if person i and j are matched, 0 otherwise
   (b) u_i: binary, 1 if person i is unpaired, 0 otherwise
   (c) u_j: binary, 1 if person j is unpaired, 0 otherwise
3. **Hyperparameters:**
   (a) ρ: penalty term for unpaired singles in the final formulation
4. **Objective:**
   (a) $\min_{z,u} -\sum_{i=1}^{N} z_{i,j}(\ell_{i,j} + \ell_{j,i}) + \rho(\sum_{i=1}^{N} u_i + \sum_{j=1}^{M} u_j) \quad \forall \ell_{i,j} \in U_i, \ \forall \ell_{j,i} \in U_j$
5. **Uncertainty Sets:**
   (a) $U_i = \{\ell_{i,j} \in \mathbb{R} : \ell_{i,j} - \sigma_i \leq \ell_{i,j} \leq \ell_{i,j} + \sigma_i\}$ (Uncertainty for person i's preference scores)
   (b) $U_j = \{\ell_{j,i} \in \mathbb{R} : \ell_{j,i} - \sigma_j \leq \ell_{j,i} \leq \ell_{j,i} + \sigma_j\}$ (Uncertainty for person j's preference scores)
6. **Constraints:**
   (a) $\sum_{j=1}^{M} z_{i,j} \leq 1 \quad \forall i$ (Max 2 people per pair)
   (b) $\sum_{i=1}^{N} z_{i,j} \leq 1 \quad \forall j$ (Max 2 people per pair)
   (c) $r_{i,f} + r_{j,f} \geq 2(z_{i,j})t_{i,j} \quad \forall i, j, f$ (Pairs must share religion if it is important to at least one party)
   (d) $t_{i,j} \geq m_i \quad \forall i, j$ (Auxiliary variable t $\geq \max(m_i, m_j)$)
   (e) $t_{i,j} \geq m_j \quad \forall i, j$ (Auxiliary variable t $\geq \max(m_i, m_j)$)
   (f) $u_i = 1 - \sum_{j=1}^{M} z_{i,j} \quad \forall i$ (Definition of unpaired single)
   (g) $u_j = 1 - \sum_{i=1}^{N} z_{i,j} \quad \forall j$ (Definition of unpaired single)

Figure 1: Dating Assignment Formulation for Traditional Pairs of 2, with Robustness.

1. **Data:**
   (a) $\ell_{i,j}$: how much person i likes person j
   (b) $\ell_{i,k}$: how much person i likes person k
   (c) $\ell_{j,i}$: how much person j likes person i
   (d) $\ell_{j,k}$: how much person j likes person k
   (e) $\ell_{k,i}$: how much person k likes person i
   (f) $\ell_{k,j}$: how much person k likes person j
   (g) g_i: gender of person i (0-2)
   (h) g_j: gender of person j (0-2)
   (i) g_k: gender of person k (0-2)
   (j) $N = M = O$: number of people in the dating pool
2. **Decision Variables:**
   (a) $z_{i,j,k}$: binary, 1 if person i and j and k are matched, 0 otherwise
3. **Objective:**
   (a) $\min_{z,u} -\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k=j+1}^{O} z_{i,j}(\ell_{i,j} + \ell_{i,k} + \ell_{j,i} + \ell_{j,k} + \ell_{k,i} + \ell_{k,j})$
4. **Constraints:**
   (a) $\sum_{i=1}^{N}\sum_{j=1}^{M} z_{ijk} \leq 2 \quad \forall k$ (Max 3 people per pair)
   (b) $\sum_{j=1}^{M}\sum_{k=1}^{O} z_{ijk} \leq 2 \quad \forall i$ (Max 3 people per pair)
   (c) $\sum_{k=1}^{O}\sum_{i=1}^{N} z_{ijk} \leq 2 \quad \forall j$ (Max 3 people per pair)
   (d) $z_{ijk} = z_{jik} \quad \forall\, i \leq j \leq k$ (Pair consistency)
   (e) $z_{ijk} = z_{jki} \quad \forall\, i \leq j \leq k$ (Pair consistency)
   (f) $z_{ijk} = z_{kji} \quad \forall\, i \leq j \leq k$ (Pair consistency)
   (g) $z_{ijk} = z_{kij} \quad \forall\, i \leq j \leq k$ (Pair consistency)
   (h) $z_{ijk} = z_{ikj} \quad \forall\, i \leq j \leq k$ (Pair consistency)

   Case: 2 People in Pair
   (i) $(g_i + g_j - 1) \cdot z_{ij}$

   Case: 3 People in Pair
   (j) $(g_i + g_j + g_k - 2) \cdot z_{ijk} \leq 0$ (Gender for 3 people)
   (k) $(g_i + g_j + g_k - 1) - z_{ijk} \geq 0$ (Gender for 3 people)

Figure 2: Dating Assignment Formulation for Pairs of 3 Maximum

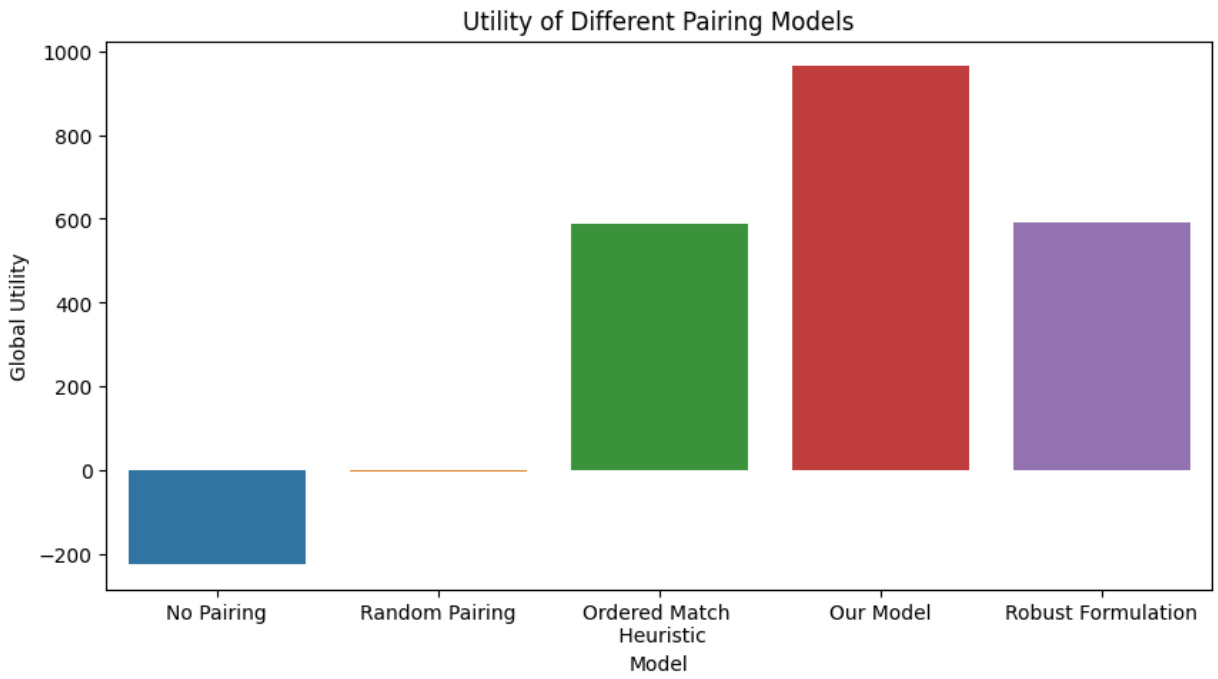

Figure 3: Objective Value of Each Baseline and Matchmaking Model (Inverted for Clarity)

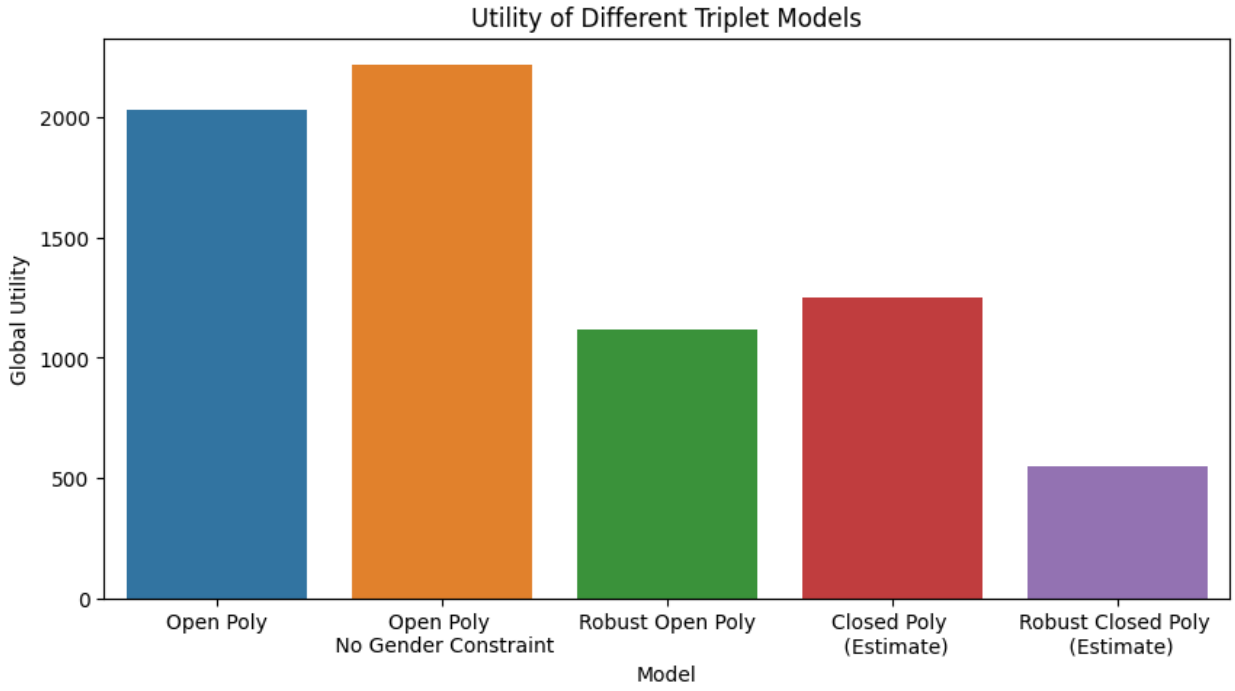

Figure 4: Objective Value of Each Triplet Matchmaking Model (Inverted for Clarity)