Trey Hibbard 4/7/2025

# Blind Dating: Intermediate Report

## Background:

Back in spring of 2024, my friend Tyler held a blind dating night which hosted 11 dates (22 participants) all together at the same restaurant, at the same time. Data regarding the participants and their personalities were gathered via a Google Form and qualitative questions. There were mixed results from the dates, with most people parting ways, but there was one pairing who became close friends and another who is still together today! In fall 2024, the same event was held at the same place, although with much more participation. With 60 dates (120 participants), the task of manually pairing up as many dates as possible became a daunting one. With plans to do another round of blind dates this spring 2025, the number of submissions was expected to grow yet again. After I saw a video on YouTube, now titled "What is the Dot Product and How is it Used? Applied Linear Algebra (part 2)", the idea of using cosine similarity to determine correlation became an apparent solution for the future of this project. Together with another friend Garrett, the three of us split up our involvement in the project as follows: I would mainly focus on the mathematics, Tyler focused on the outreach and coordination, and Garrett made the code.

## Motivation:

The most tedious part of the blind dates had been individually pairing individuals based on personal judgement. As 22 participants became 60, it also becomes a challenge to see which two people may be a good fit for each other if their rows are far apart in a spreadsheet. Cosine similarity was seen as a solution to this. It would provide each pairing a score between -1 and 1 to determine compatibility based on numerous personality-based questions. After developing a code to compare every person against each other with cosine similarity the code would return a list of each person with a ranking of others based on their cosine similarity pairing.

Cosine similarity transforms our 38 questions, or indicators, into one metric that can be used in determining pairings. Questions focused on 6 categories: emotion, conflict, extraversion, lifestyle, communication, and partner interaction. In addition to the 38 questions, dealbreaker questions such as one's sexuality and age limits were used to eliminate conflicted pairings. The resulting ranking simplifies the final pairing process. There are several options we considered using for final pairings, such Gale-Shapley's algorithm or Kuhn's algorithm, but since these algorithms required bipartite graphs and our dating pool considered queer individuals who didn't necessarily fit in a bipartite graph, we ended up using a maximum weight matching to meet the needs of everybody.

Through our survey we gathered 192 submissions. That is 7,296 datapoints (38*192), but does not include dealbreaker questions and a number of qualitative questions asked for manual investigation. We had about a 44% retention rate by the end as a consequence of short notice and people dropping out. A proposed solution to these problems is made below.

# Methodology:

Without any vital quantitative data and results from the first two rounds of blind dates, this time served as round zero for what we intend to be an iterative process going forward. The process is as follows:

1) Determine a list of questions
2) Adjust the algorithm to work with these questions
3) Send the preliminary survey
4) Create a list of matchings
5) Have the blind dates
6) Send the post-date survey
7) Perform statistical analysis comparing the results of the post-date survey to the answers from the preliminary survey.

The beginning of the process again would use results from the past statistical analysis to determine which questions are the most important or should be emphasized more. There is certainly a large amount of personal judgement as to what makes a "good question" on our part, but we hope that by analyzing our results we can improve which questions we ask in the future.

<u>Cosine Similarity</u>

To begin deciding which questions should be asked and the best way to ask them, we sought out questions that would fit our cosine similarity model. To understand how to best fit the model, it's important to understand how it works.

$$\cos(\Theta) = \frac{A \bullet B}{|A|\,|B|}$$

$\cos(\Theta)$ tells us the cosine of the angle between two vectors A and B. For vectors in the space $\mathbb{R}^n$, where n is the dimension and furthermore the number of questions, the vector may be made up of positive and negative values. The dividend is a strictly positive value, so negative scores come directly from the dot product of A and B. This requires us to build questions where inputs in A and B when multiplied are positive if there is correlation and negative if there is not. To do this, we aimed to ask questions that could be answered on a seven-point scale for the values -3 to 3. An example of a question would be:

"You are often drained when at a big party."

The seven-point scaled ranged from "very unlike me" to "very like me" with a score in the middle being assumed as indifference, or mathematically 0. If person A answers this question with a score of -3 and person B answers with a score of -1, then those answers would be placed in the same index of their respective vectors and multiplied in the dot product, adding a positive score of +3 to the sum of A • B. This increases the total cosine similarity score, though it is important to note that -3 * -2 = +6 would've resulted in an even higher score. If they differed on the question, however, then -3 * 2 = -6 would decrease the cosine similarity score. This is an assumption we made that compatible individuals would be interested in attending similarly sized events. Statistical analysis after the dates will ideally allow us to see which questions were good indicators or not.

The question provided above is an example of a "reflexive" question as we called it. It assumes that people want the same thing. However, often people want somebody else to be different than themselves. We call these "dual" questions. It also presents the need for each person to have two different vectors. One to describe themselves personally, and one to describe what they are looking for in somebody else. Here is an example of a dual question:

Self: "You tend to be the center of attention."

Want: "You are more attracted to a quieter personality."

If person A answered 3 for the "Self" question and person B answered -2 for the "Want" question, then we would calculate the similarity as 3 * -(-2) = 6. The extra negative sign is specific to certain questions and addresses something we referred to as polarity: when one of the responses to a dual question needs its sign flipped to accurately show correlation in the dot product. The two different types of vectors, simplified, are shown below.

ASelf = […]

AWant = […]

BSelf = […]

BWant = […]

To determine person A's compatibility towards every other person, for example person B, we would apply cosine similarity between AWant and BSelf. Compatibility of B towards A would apply cosine similarity between BWant and ASelf. These two calculations would give different scores and answers the obvious problem that someone may have greater interest than somebody else. The number we used to determine mutual compatibility was the minimum of the two, since if the minimum was high, we assumed the match would be good.

Trey Hibbard

The following images show excerpts of our process handling csv files and python code. The first image shows examples of user responses to the question "You tend to be more attracted to people who are very outgoing." The response was received as a number from 1-7 and then shifted by four to fit our scale between -3 and 3. The second image shows an example from our code where we consider one our dealbreakers regarding alcohol usage. The third image provides an excerpt (with relabeling for privacy) of the beginning of a list of Person A's ranking of others based on their score.

| | |
|---|---|
| 1 | **You tend to be more attracted to people who are very outgoing. [Extraversion]** |
| 2 | 3 |
| 3 | 3 |
| 4 | 5 |
| 5 | 7 |
| 6 | 5 |
| 7 | 6 |
| 8 | 4 |
| 9 | 5 |
| 10 | 4 |
| 11 | 6 |

```
    # Alcohol and substances check
    if person1.dealbreakers["alcohol"][1] != person2.dealbreakers["alcohol"][0]:
        print(person1.dealbreakers["alcohol"][1])
        print(person2.dealbreakers["alcohol"][0])
        print(f"Alcohol issue {person1.name}--> {person2.name}")
        return False, common_days
```

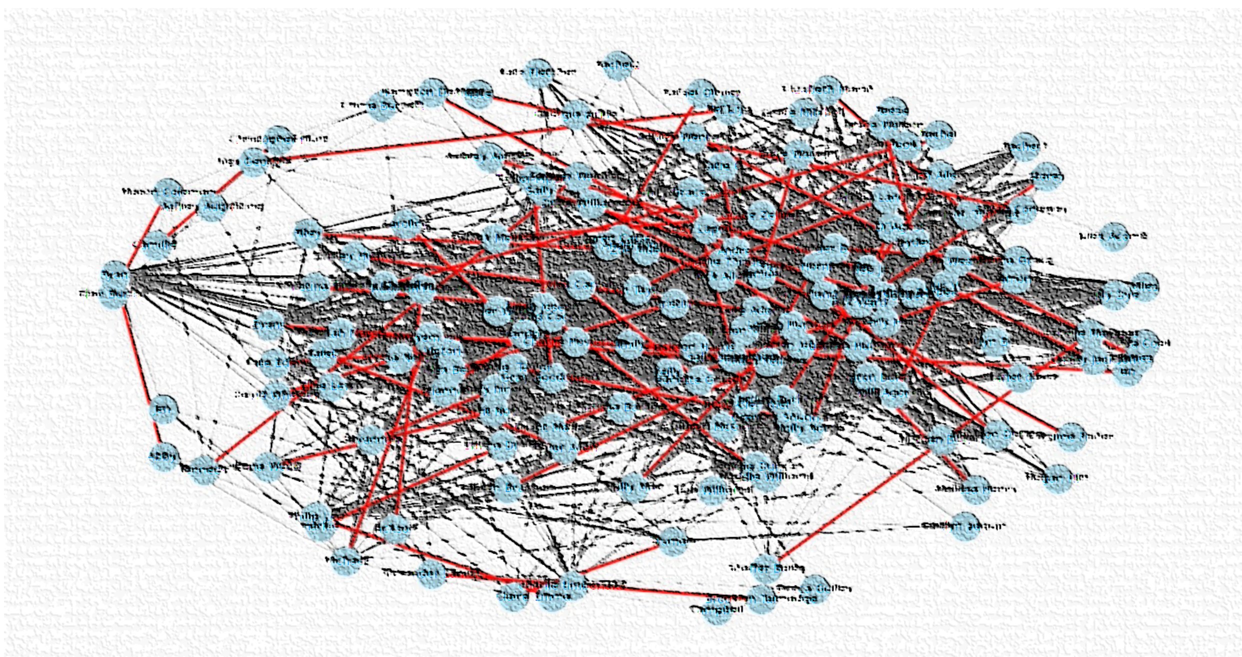
| | | | |
|---|---|---|---|
| 427 | **Person A** | Person B | 0.8347 |
| 428 | **Person A** | Person C | 0.7407 |
| 429 | **Person A** | Person D | 0.7224 |
| 430 | **Person A** | Person E | 0.6911 |
| 431 | **Person A** | Person F | 0.6803 |
| 432 | **Person A** | Person G | 0.6661 |

Maximum Weight Matching

After receiving every cosine similarity score of each person towards another, it was a separate question of how we would use that information to pair people. As mentioned in the "Motivation" section of this paper, the method we ended up using was a maximum weight matching. The NetworkX python package provides a simple way to incorporate this.

We set each person's name as a node, and the minimum cosine similarity between two people as a weighted edge. If there was a dealbreaker between the two people, no edge was assigned. The maximum weight matching algorithm then found the pairing list with the largest score sum from the weighted edges. It is important to note that this algorithm is not only maximal, but maximum, which means that there is not a greater combination of positive edge weights that could have been made. In other words, you wouldn't be able to manually spot another edge that had a positive score which nodes weren't already paired elsewhere.

The following graph represents a visual for the matching network. Red lines indicate pairs. The graph has been blurred to maintain privacy.



The pairs gathered from this algorithm were the primary ones used, with manual changes being made when people dropped out after confirming their availability for the date. Manual changes have been noted and will be considered carefully or excluded in the statistical analysis. Compatibility scores exist for all pairings however and still provide a practical way to apply linear regression.

# Hurdles:

Many of the challenges we had during the project pertained to a quick timeline, short notices sent to participants, and poor data collection techniques that made the code tedious and impractical.

Luckily, since our new methods have already been discussed and investigated, such as cosine similarity and maximum weight matching, timing is expected to be less of an issue. However, we do plan to have our code written before a survey is sent out to participants next time. This would also solve the issue of short notices and people dropping out.

Regarding data collection, our use of a Google Form created several problems. When changing questions for word choice, we learned it would create an entirely new column in the resulting Google Sheet, which led to an unnecessarily large spreadsheet with empty columns. Also, when making questions, there wasn't a convenient way to pair dual questions to each other aside from directly indexing them. This is understandably poor practice and solutions such a key identifier in front of the question (such as [A1] and [A2] for dual question A) have already been proposed to solve this problem and make survey expansion easier.

Additionally, to promote having more blind dates, we felt that we could remove some dealbreakers. For example, we could likely remove a category such as height, while offering an alternative to women who desire a tall man by simply asking if their partner should be taller than them or not. At the end of the day, this is a blind dating service.

# Future Statistical Analysis:

A post-date survey has already been sent out to everyone who went on a blind date. Questions from this survey are mostly asked on a seven-point scale and will be compared to the number derived from each pair's score multiplied for "reflexive" and "dual" questions.

Questions asked on the survey aim to measure compatibility in personalities, romantic tension, and physical attraction. Additionally, we asked participants a few binary questions: did they want a second date, if a second date was lined up, and if they expect to be friends with the other in the future.

For questions asked on the seven point scale, I will be seeing how impactful a question was to determine the quality of a blind date using a stepwise multivariate linear regression model. Other regression models will be analyzed as well. For binary questions, logistic regression will be primarily used with additional investigation using clustering methods.

Although our pairings were made exclusively using the cosine similarity metric, all questions fed into the metric, and more decisive questions would lead to better accuracy in the scores. Results from our statistical analysis will inform us on how to change our survey to better reflect the interests of others.

Sources:

https://www.youtube.com/watch?v=BKwKRIUKv64