Game of Hearts

Promoting Trust in Online Dating through an Evolutionary Game Model

Geoffrey Yang, Annie Tang, Conner Boehm, Wade Williams, Catherine Chu, Isaiah Bradner

Dartmouth College

Spring 2023

Abstract

In this paper, we present an evolutionary trust game to investigate the formation of trust on online dating platforms. Our model consists of four types of players: trustworthy Initiators, untrustworthy Initiators, trusting Recipients, and untrusting Recipients. Through two stages of analysis - first determining the theoretical long-run equilibria under various payoff assumptions, and then simulating a dating game on a finite social network - we determine the circumstances under which trust evolves. Our results show that while it is difficult to ensure trustworthiness, moderate penalties for untrustworthy Initiators result in frequent date acceptances and a fair number of trustworthy matches. These findings have important implications for designing policies on online dating platforms.

1. Introduction

Online dating is an increasingly popular way for young people to find romantic partners, with online dating apps expected to reach 441 million active users worldwide by the end of 2023 (Buchholz 2023). Additionally, 21.9 percent of the American population is on at least one online dating platform (Dixon 2023). Yet despite its growing popularity, online dating comes with a myriad of risks and dangers. Since daters fear that their potential partners are shopping through many attractive people online, online dating creates social pressure for individuals to present themselves in the way that they think makes them seem most desirable. The term "catfish" has been coined to describe someone that uses someone else's images and information to create a new identity online. Critically, 38 percent of men and 23 percent of women have catfished someone at some point in their life (Desimone 2022). Other methods of deception could include lying about one's height and age, "airbrushing" or otherwise editing one's profile photos, or using a false job title or alma mater.

This is problematic for multiple reasons: not only is the deceived person deterred from using dating apps again, but the deceiver is also likely to match with people they are severely incompatible with. The negative consequences of deception are exacerbated when more and more users do it. When everyone adds a couple inches to their height or touches up their photos to look more attractive, it creates pressure for other users on the app to do the same so as not to look worse by comparison. Even the dating apps themselves acknowledge the prevalence of deceptive profiles on their platforms. According to one dating platform, almost 80 percent of its users had been involved in

		Recipient	
		TR	UR
Initiator	TI	(R, R)	$(-N-S, -N)$
	UI	$(R+B, -L)$	$(-N, -N)$

Table 1: Payoff matrix

a conversation with a romance scammer at one point, a record high according to the Federal Trade Commission (Fletcher 2022).

As dating apps like Tinder, Bumble, and Hinge grow in popularity, it is important to understand the origins of deceptive behavior and how it can be prevented. Using evolutionary game theory, we model and analyze a dating environment as an ecosystem, determining the circumstances when untrustworthy behavior dominates and when trustworthiness evolves. We use our results to inform how real-world online dating platforms might promote trustworthiness and ensure a better experience for their users.

2. Methods

2.1. Game definitions and payoffs

We model online daters playing a sequential trust game in pairs. The Initiator moves first, presenting their online dating profile to the Recipient; the Recipient then decides whether or not to go on a date with the Initiator. Each player can choose from one of four possible strategies:

- TI: A **Trustworthy Initiator** presents their dating profile honestly.
- UI: An **Untrustworthy Initiator** exaggerates their profile to make themselves seem more attractive, hoping to fool higher quality Recipients into dating them.
- TR: A **Trusting Recipient** agrees to go on a date.
- UR: An **Untrusting Recipient** refuses to go on a date.

The payoff matrix in Table 1 displays the outcomes for playing each strategy. R is the base reward for Initiators and Recipients that go on a successful date. When a UI misleads a TR into a date, the UI receives R, plus a bonus B (for getting a date with a more attractive Recipient, in expectation). The TR who was misled receives a negative payoff $-L$, reflecting the painful emotional impact of being misled. For all players, the penalty for not getting a date is $-N$. Finally, Initiators that play TI yet fail to secure a date receive an additional penalty $-S$; this represents a "regret penalty" for thinking they might have gotten a date if they had chosen to play the UI strategy instead.

2.2. Fitness

In the evolutionary context, we are interested in the fitnesses of the Initiator and Recipient strategies, as they determine whether trustworthiness or untrustworthiness evolve in equilibrium. Let x denote the proportion of Initiators playing the TI strategy; then, the proportion playing UI is $1-x$. Likewise, let y represent the fraction of Recipients playing TR, and $1-y$ the fraction playing UR. Then, we can express the fitnesses of the

four strategies as follows:

$$f_{TI}(y) = Ry - (N + S)(1 - y) \tag{2.1}$$
$$f_{UI}(y) = (R + B)y - N(1 - y) \tag{2.2}$$
$$f_{TR}(x) = Rx - L(1 - x) \tag{2.3}$$
$$f_{UR}(x) = -N \tag{2.4}$$

We can then derive the average fitnesses for the Initiators and Recipients:

$$\overline{f_I} = xf_{TI}(y) + (1 - x)f_{UI}(y) \tag{2.5}$$
$$\overline{f_R} = yf_{TR}(x) + (1 - y)f_{UR}(x) \tag{2.6}$$

2.3. Strategy Update

As this trust game is played repeatedly, the values of x and y will change, as the strategies with higher relative fitness grow in frequency and potentially dominate the population. We formalize this process through replicator dynamics, which assume that strategies with higher fitness than the average will grow over time. For the Initiator and Recipient, the replicator dynamics are:

$$\dot{x} = x(f_{TI}(y) - \overline{f}_I) \tag{2.7}$$
$$\dot{y} = y(f_{TR}(x) - \overline{f}_R) \tag{2.8}$$

where $\dot{x}$ and $\dot{y}$ are the proportion of TI and TR strategies in the next time step, respectively.

3. Equilibrium Analysis

In the first part of our analysis, we use Dynamo 2×2 simulations to analyze three static scenarios for an online dating market. These are: a base case (3.1), a "desperation" scenario where Recipients are eager to go on dates (3.2), and a punishment scenario where Initiators face a heavy penalty for pursuing the UI strategy (3.3). In each case, we create distinct payoff matrices by varying the values of R, B, L, N, and S from our model. The results from this analysis inform us what behavior we would expect to emerge in equilibrium, under highly idealized assumptions - static payoffs, an unlimited time horizon, and a sufficiently large population that local "clusters" of daters don't matter.

3.1. Base case

Our base scenario incorporates what we believe are standard assumptions for an online dating market. Being misled is highly punitive for Recipients ($L = 5$), reflecting the negative emotional consequences of placing trust in an Initiator, risking an in-person date, only to realize they have been misled. Second, Initiators, perhaps having failed to secure dates in the past, receive a high bonus from lying ($B = 3$) because they believe exaggerating their profile will significantly increase their chances of success.

Under these assumptions, the payoff matrix (Table 2) indicates that UI-UR is a strict Nash equilibrium. That is, given UI, UR is the best response for the Recipient (payoff -1 versus -5 for TR); likewise, given UR, UI is the best response for the Initiator (payoff -1 versus -2 for TI). Since UI-UR is a *strict* Nash equilibrium, it is also an evolutionary

	TR	UR
TI	$(5, 5)$	$(-2, -1)$
UI	$(8, -5)$	$(-1, -1)$

Table 2: Base case payoff matrix

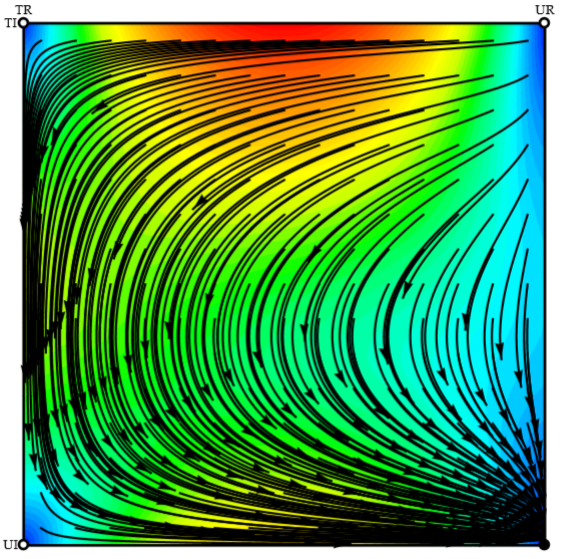

Figure 1: Base case phase diagram

stable strategy (ESS): once the population moves to the UI-UR strategy combination, no alternative strategy combination can invade. The phase diagram (Figure 1) confirms that no matter what the initial strategy frequencies are in the population, the population eventually converges to UI-UR. Finally, UI-UR is a stable rest point; small perturbations away from UI-UR will be pushed back by the dynamics of the system.

This analysis paints a bleak picture for online dating. According to this model, online dating platforms will eventually converge to an equilibrium where all users lie in their profiles, and no users accept going on dates. Furthermore, even if the dating platform can engineer a one-time shift away from UI-UR, the population will eventually return to the untrustworthy equilibrium, given that UI-UR is a stable rest point.

3.2. Desperation

In the real world, however, we do not observe the untrustworthy equilibrium all the time: online dating platforms *do* successfully facilitate dates, on occasion. We explore two possible explanations for what we observe. The first is that Recipients are desperate to go on dates. That is, for a Recipient, it is no worse to go on a date with an untrustworthy Initiator that it is to be stuck at home with no date at all ($L = 3$, and $N = 3$). Initiators are symmetrically desperate to go on dates ($N = 3$). As in the base case, there is an

	TR	UR
TI	$(5, 5)$	$(-4, -3)$
UI	$(8, -3)$	$(-3, -3)$

Table 3: Desperation payoff matrix

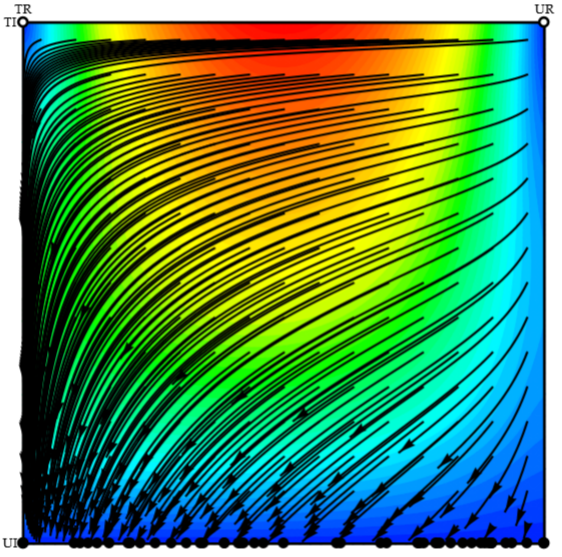

Figure 2: Desperation phase diagram

regret penalty $S = 1$ for trustworthy Initiators that fail to secure dates. R and B remain the same as the base case.

The payoff matrix (Table 3) indicates this game has two weak Nash equilibria: UI-TR, and UI-UR. Given that Initiators play UI, accepting the date (TR) or rejecting the date (UR) yield the same payoff for the Recipient (-3); hence both equilibria coexist. Weak Nash equilibria do not automatically qualify as ESS, and indeed the Dynamo simulations reveal there is no ESS for this system. However, there are multiple stable rest points along the bottom edge of the phase diagram (Figure 2). At all of these rest points, all Initiators play UI. These points differ in what proportion of Recipients play TR, with y ranging from 0.001 to 0.999.

At any of these stable rest points, minor disturbances will not change the strategic composition of the population. However, given none of these rest points are ESS, in the long-term, it is possible for an alternative strategy to invade and take over from any of these strategy combinations - provided that the alternative strategy has a sufficient initial frequency. Nevertheless, Figure 2 suggests that no matter what equilibrium we end up at, that equilibrium will have all Initiators playing UI. This seems plausible, given that over 80% of online daters admit to lying in their profiles (Toma et al. 2008).

	TR	UR
TI	$(5, 5)$	$(-2, -1)$
UI	$(4, -5)$	$(-1, -1)$

Table 4: Punishment payoff matrix

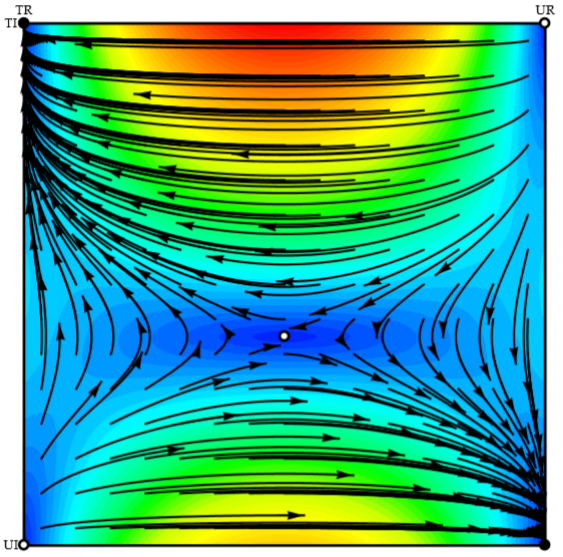

Figure 3: Punishment phase diagram

3.3. Punishment

In the third scenario, we model a dating platform where lying is punished - in other words, the bonus for UI is negative ($B = -1$). Conceptually, trustworthy Recipients that go on dates and are deceived are able to report the untrustworthy Initiators; these Initiators are subsequently punished by the algorithm, their profile being shown less often. Another possibility is that Initiators seek long-term relationships: they realize that luring someone into a date through deception makes it difficult to build a trusting relationship with them. All other assumptions are unchanged from the base case.

In the punishment case, the payoff matrix (Table 4) reveals that both TI-TR and UI-UR are Nash equilibria. They are also both ESS. The interpretation is that the system ends up at either TI-TR or UI-UR, depending on the initial frequencies of strategies in the population. The greater the frequency of trustworthy Initiators and Recipients we start with, the more likely we end up at a stable TI-TR equilibrium. Likewise, the more untrustworthy Initiators and Recipients we start with, the more likely we end up at UI-UR. Finally, there is a mixed Nash equilibrium, in the center of Figure 3, where 40% of Initiators play TI and 50% of Recipients play TR. However, this is an *unstable* rest point. That is, although players have no incentive to unilaterally change their strategies at this point, small changes in frequency can lead to a dramatic shift away from these frequencies - towards either TI-TR or UI-UR.

We conclude that under punishment, online dating becomes a bistable coordination game. The trustworthy equilibrium is better for all involved, yet the platform does not end up at the favorable equilibrium. This model seems to accurately capture certain *niches* within the online dating world: you can imagine that certain dating apps in certain cities might have all their users behaving trustworthily, for instance. However, it does not seem to accurately capture the reality of dating *platforms* as a whole, where there is always a mix between trustworthy and untrustworthy users at any given moment.

4. Network Simulation

Our analysis thus far does not seriously entertain a mixed strategy equilibrium. This conflicts with the empirical evidence from online dating platforms, where we always observe a mix of strategies. Our conclusions thus far may be put down to the idealized nature of our Dynamo simulations, that operate on an effectively infinite time horizon. Our Dynamo simulations also do not address interactions between players of the same type (i.e. between Initiators and other Initiators, or between Recipients and other Recipients), whereas in the real world we would expect individuals to assess the viability of alternative strategies outside of when they are actively engaging with others on the platform.

To address these limitations, we design a simulation that models a finite number of daters as nodes on a social network, who play the dating game over defined time steps. Our simulations establish a different equilibrium even while using the same base case payoff matrix as in subsection 3.1. We also analyze the impact of varying the bonus/punishment B for untrustworthy Initiators, and find a range of values where trustworthy behavior emerges. We first describe the experimental setup, and then discuss the simulation's results.

4.1. Experimental Setup

In our experiments, all agents in the finite population act as a social network, where the edges denote 'dates' between Initiator-Recipient, and 'interactions' between Initiator-Initiator and Recipient-Recipient, regardless of trustworthiness or untrustworthiness. These connections will later help in calculating the sum of payoffs, and strategy updates. These dates are facilitated by matching individuals on a chosen dating platform (i.e., Tinder, Bumble, Hinge, etc.) who have a mutual interest in going on a date and finding a partner, specifically through an online network. Each agent has a relationship with (a) an agent they go on a date with (a Recipient if the agent is an Initiator, or vice versa) and change their payoff or (b) an agent with the same position (Recipient or Initiator) who they can gather strategy information from. In other words, the network reflects both a matching of complementary Initiator-Recipient relationships of single individuals and an open-source network of information.

All agents in the population play the dating game over a fixed number of time steps (we simulate 1,000 time steps). Each agent i can choose one of the four possible strategies at every time step, such that $s(i) = \{TI, UI, TR, UR\}$ for all i. At every time step, each agent plays the game iteratively with another directly connected agent and decides whether to change their own strategy or not based on the payoffs. Payoffs of the agents were initially calculated based on the payoff matrix in Table 1.

The initial conditions model a population of $Z = 600$, where there are more trustworthy

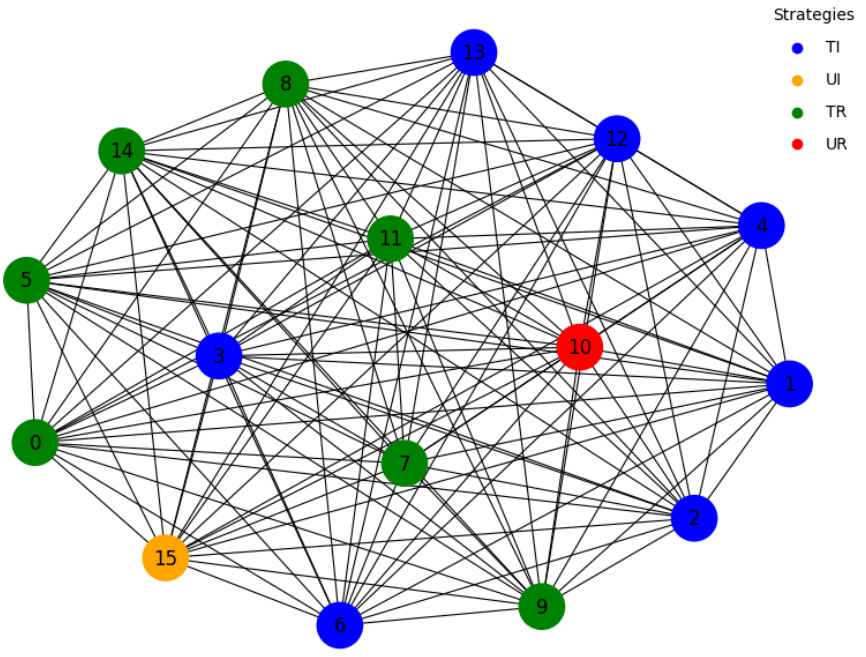

Figure 4: Initial social network of 16 agents

than untrustworthy players ($TI \approx TR \approx 0.45 \cdot Z$ and $UI \approx UR \approx 0.05 \cdot Z$). We set our population to have the majority of agents as trustworthy, as reflected in previous research (Desimone 2022). We set our static values to $R = 5$, $L = 6$, $N = 1$, and $S = 1$, chosen based on our base case values above.

We can see from the graph in Figure 4 our initial population with a size of 16 for ease of visualization.

4.2. Simulation Algorithm

With the setup defined, we now explore the logic of our simulation. In one game, all Initiators will go on a date with a random Recipient. As an example, in Figure 4, Agent 12 (TI) will go on a date with Agent 11 (TR). Then, based on the payoff matrix and who was trustworthy or untrustworthy, we update every agent's payoff value from the date they just went on to represent the "success" they're having on the dating platform. In our example $P(A_{12}) = 5$, and $P(A_{11}) = 5$. Then, we evaluate the "interactions" that Initiators have among Initiators, and Recipients have among Recipients. Each actor compares their payoff values with each other, simulating a reality in which an individual's friends know the results of how their dates went, with the logic that if an agent's connections have a higher payoff score, then that agent might be incentivized to switch strategies. The same scenario occurs with the Recipients. In our example, say Agent 12 sees their friend, Agent 10, who is an Untrustworthy Initiator, go on a date. Agent 10 receives a higher payoff from going on a date with Agent 9 ($P(A_{10}) = 8$, $P(A_9) = -1$). Agent 12 will then have a random probability of adjusting their strategy to Agent 10's strategy (i.e. $s(A_{12}) = UI$). We run these three steps 1,000 times for every simulation.

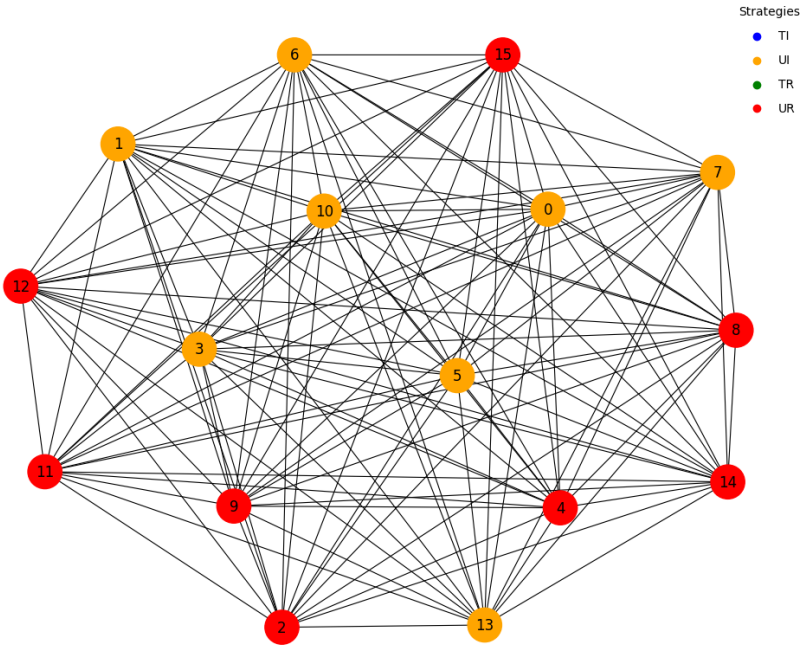

Figure 5: Network with pop = 16 and base case payoff matrix

Following our initial population simulation of 16 from Figure 4, our output is Figure 5. Our simulations corrobate our initial Dynamo calculations, where the base payoff matrix results in all agents converging to untrustworthy Initiator and untrusting Recipient strategies.

Our pseudocode for one payoff matrix simulation can be found in Algorithm 1.

4.3. Results on a Larger Population

For simplicity, we described our algorithm on a small population of 16. We then expanded our population to 600 to simulate a larger social network of agents. We assume the same initial distribution of strategies.

Results are shown in Figure 6. We see that the majority of agents are still untrustworthy, but with a larger population, we see a few deviants. Although every Initiator becomes untrustworthy, 84/300 Recipients are still trusting. This is likely due to each agent's probability of updating strategies. We believe this to be more reflective of reality, as some people still believe in the good of humanity and trust that not everyone on these dating platforms is untrustworthy. However, the majority of agents still eventually play UT or UI based on our base case payoff matrix.

4.4. Influence of Punishment on Initiators

In this section, we take our simulation algorithm and determine how varying levels of bonus/penalties applied to untrustworthy Initiators (UI) influence the dynamics of the game. These results are important to our conclusions, given we are trying to determine

Algorithm 1 Single Game Simulation

```
procedure singleSim(populationSize, payoffMatrix, graph)
    for i in range(1,000) do
        // simulate dates between Recipient and Initiators
        for edge in graph.edges do
            agent1, agent2 = edge
            if strategy(agent1), strategy(agent2) in payoffMatrix then
                agent1[payoff] += payoff(agent1)
                agent1[payoff] += payoff(agent1)
            end if
        end for
        // update strategies
        for agent in graph.nodes do
            for neighbor in graph.neighbors[agent] do
                if not areRecipients(agent, neighbor) then
                    continue
                end if
                if not areInitiators(agent, neighbor) then
                    continue
                    if neighbor[payoff] > agent[payoff] then
                        if rand(0, 1] <= agent[probability] then
                            agent[strategy] = strategy(neighbor)
                        end if
                    end if
                end if
            end for
        end for
    end for
end procedure
```

what interventions can cause the majority of agents on our social network to become trustworthy. We plot a frequency analysis graph on B and the number of agents, showing the different values of the penalty parameter B on the x-axis.

Results can be seen in Figure 7. We show the final states of the population with various values of B, in increments of 0.5. Examining the plot, we see that increasing the penalty for UI has two major effects. The first one is positive for promoting trust because it reduces the number of UR drastically if penalty levels are -1.5 or larger in magnitude. The second one is positive but less dramatic, as the amount of trustworthy and untrustworthy Initiators fluctuates depending on the punishment level. Although there seems to be an average of 20-30% of Initiators that remain untrustworthy when $-5 \leqslant B \leqslant 0$, we conclude that the majority of Initiators turn trustworthy when punishment is enforced. It is only at extreme levels of punishment that all Initiators and Recipients will be trustworthy. We believe these punishments are unrealistic for online dating platforms to implement in the real world, as they would discourage even trustworthy users from joining the platform.

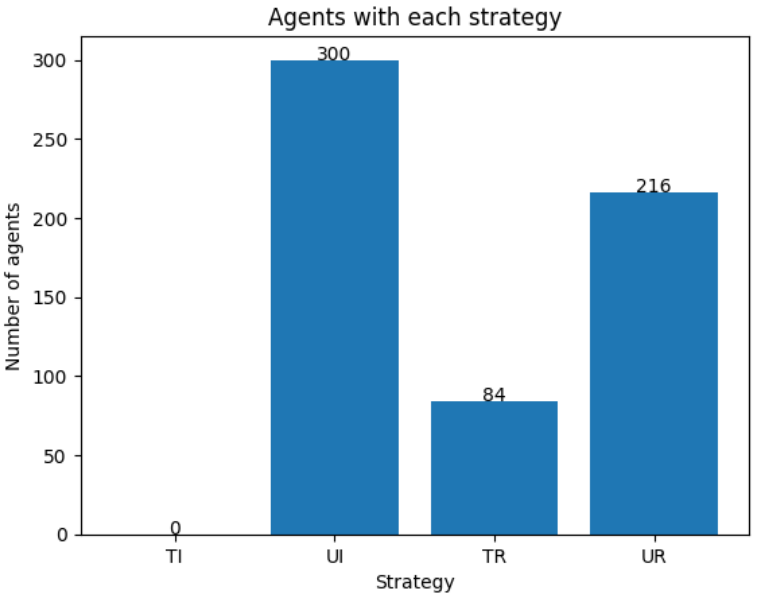

Figure 6: Distribution of strategies with pop = 600 and base case payoff matrix

5. Discussion and Conclusion

In this paper, we investigated the formation of trust on online dating platforms. We modeled dating as a two-player trust game, played between Initiators that present their profile honestly or dishonestly, and Recipients that decide whether or not to go on a date. Using evolutionary game theory, we analyzed when the population converges to trustworthy and untrustworthy equilibria.

In the first stage of our analysis, we utilized Dynamo simulations to study the long-run behavior of large populations of daters. Under our base case assumptions, we found that the population converges strongly to an untrustworthy equilibrium (3.1). Even when we introduce Recipients who are desperate for dates, Initiators stick to their dominant untrustworthy strategy (3.2). Only when untrustworthy Initiators are severely punished does trustworthiness emerge - even then, as one of two possible equilibria (3.3).

None of our Dynamo simulations raised the possibility of a mixed strategy equilibrium. However, in reality, dating platforms are populated with a mix of trustworthy and untrustworthy agents at any given moment. To provide an explanation, we turned to a second stage of analysis where we simulated a finite number of agents playing our dating game on a social network. In each time step, agents get matched with neighbors, choose a strategy, and update their strategy based on their neighbors' payoffs. Simulating over a finite number of time steps, mixed strategies emerged: under our base case assumptions, our simulations show Recipients playing a mix of trusting and untrusting strategies, in contrast to our Dynamo simulations which suggested they would only play the UR strategy.

Finally, we ran our simulations with varying levels of B - the bonus or penalty granted to untrustworthy Initiators. We found that only a severe penalty for lying ($B < -6$) could ensure that all Initiators and Recipients were trustworthy. However, moderate penalties

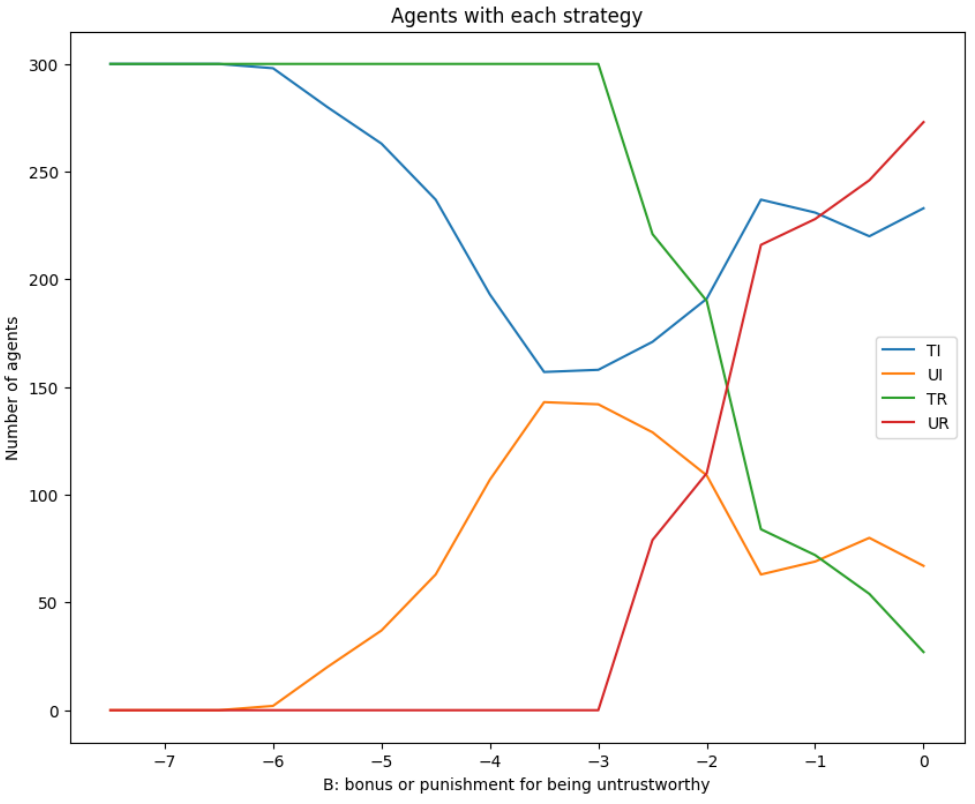

Figure 7: Varying agents in each strategy based on various penalty B values for the Untrustworthy Initiator. Simulations were run with the initial population having 45% trustworthy and 5% untrustworthy Recipients and Initiators in each subgroup.

$(-4 < B < -2.5)$ resulted in at least some Initiators being trustworthy, and a majority of Recipients playing TR and thus going on dates.

Our results have important policy implications for online dating platforms. First, they suggest that 100% trustworthiness is an unattainable goal for most dating platforms, given that severe punishment is unlikely be practical due to deterring even trustworthy users from joining. However, even moderate punishment for false profiles results in an environment where a majority of users are receptive to going on dates, and a significant portion of those dates happen between two trusting individuals. In sum, dating apps should explore avenues to punish dishonesty - perhaps a system where Recipients report false profiles, which are then penalized by the algorithm - in order to promote trustworthiness and ensure a more positive experience for their users.

5.1. Future research

The simulation component of this paper assumed fixed initial frequencies for each strategy. Further research could explore the effect of varying these frequencies. How sensitive are the outcomes to the initial frequencies, and over what time scale do these effects matter? Second, further research could explore the impact when successful dates (TI-TR in our nomenclature) leave the platform after forming a relationship. Does the

inherent nature of online dating result in a selection problem, where only untrustworthy players are left on the platform?

6. Data accessibility

Code is available on GitHub at https://github.com/annieetang/game-of-hearts

7. Competing interests

The authors have no competing interests to declare.

8. Authors' contributions

A.T, C.B., and G.Y discussed and conceived the models, and performed simulations and data analyses. All authors wrote and reviewed the manuscript and created the presentation. I.B. and C.C. created the poster.

9. Acknowledgements

The authors are extremely thankful to Professor Feng Fu, Associate Professor of Mathematics at Dartmouth College, for the continuous support, feedback, and encouragement to complete this research topic.

REFERENCES

Buchholz, Katharina 2023 Infographic: How the World Dates Online.

Desimone, Rachel 2022 Catfishing: The What, How, and Why.

Dixon, Stacy 2023 Topic: Online dating in the United States.

Fletcher, Emma 2022 Reports of romance scams hit record highs in 2021.

Toma, Catalina L., Hancock, Jeffrey T. & Ellison, Nicole B. 2008 Separating Fact From Fiction: An Examination of Deceptive Self-Presentation in Online Dating Profiles. *Personality and Social Psychology Bulletin* **34** (8), 1023–1036.