

# The Fragility of Trust: Modeling Cooperation and Defection in a Noisy Networked Prisoner's Dilemma

Asma Daha

ENG-270: Computational Methods and Tools

GitHub repository link: <https://github.com/asmadaha/cmt-project.git>

## 1 Deviations from project proposal

The final project remains closely aligned with the original proposal. The central question; how stochastic noise affects cooperation in a networked repeated Prisoner's Dilemma was preserved, along with all core elements, including the use of a repeated Prisoner's Dilemma, a networked population of agents, four classical strategies (ALLC, ALLD, TFT, GRUD), stochastic noise, and payoff-based imitation dynamics.

I introduced two minor deviations. First, I implemented in C the full agent-based simulation, including interactions, noise, and imitation dynamics, while I used MATLAB exclusively for post-processing, data analysis, and visualization. Second, the original proposal included a systematic exploration of network connectivity but I omitted this to focus more deeply on the effects of noise, which already produces rich and interpretable dynamics.

These changes improved clarity and reproducibility without altering the scientific goals of the project.

## 2 Introduction to the problem

Trust and cooperation underpin many human and biological systems, from economic exchange to social relationships to microbial communities. The repeated Prisoner's Dilemma (PD) remains a central model for understanding how cooperation emerges among self-interested agents. Classic work by Axelrod demonstrated that reciprocal strategies such as Tit-for-Tat could sustain cooperation through repeated interaction [1].

However, real-world interactions are rarely perfect. Individuals sometimes defect by mistake, misread signals, or misunderstand intentions. This motivated later work, including Nowak and Sigmund's analysis showing that even small error rates destabilize cooperative strategies [2]. Structured population models reveal further complexity: depending on noise and network topology, cooperation may persist, fragment, or collapse entirely [3, 4].

An interesting exposition of misunderstanding-driven trust collapse appears in the Veritasium video "*This Game Theory Problem Will Change The Way You See The World*" [6], which helped shape the conceptual motivation for this project and illustrates how accidental defections can cascade into widespread breakdowns of cooperation.

From a computational perspective, the Prisoner's Dilemma provides an ideal testbed for exploring how simple local rules and stochastic perturbations can generate complex collective behavior.

### Scope

Included processes:

- A population of 100 agents arranged on a fixed random network.
- Four classical strategies: Always Cooperate (ALLC), Always Defect (ALLD), Tit-for-Tat (TFT), and Grudger (GRUD).

- Stochastic noise: with probability  $p_{\text{noise}}$ , an intended action flips.
- Payoff-based imitation dynamics: agents may adopt the strategy of a more successful neighbor.

Excluded processes:

- Dynamic network rewiring (graph topology is fixed)
- Learning beyond imitation

## 3 Approach used

### 3.1 Model Structure

The model is implemented as a forward-time agent-based simulation. Each of the  $N = 100$  agents occupies a node in a fixed random network. At each time step, agents play the PD with their neighbors, accumulate payoffs, and update their strategies via imitation. Interactions follow the standard Prisoner's Dilemma payoff matrix:

$$\begin{pmatrix} R & S \\ T & P \end{pmatrix} = \begin{pmatrix} 3 & 0 \\ 5 & 1 \end{pmatrix}$$

### 3.2 Strategies

I implemented four deterministic strategies:

- **ALLC**: always cooperates.
- **ALLD**: always defects.
- **TFT**: cooperates initially; then copies the opponent's previous action.
- **GRUD**: cooperates until the opponent defects, then defects permanently.

### 3.3 Noise Model

With probability  $p_{\text{noise}}$ , an agent's intended action is flipped. This models execution errors and misunderstandings. A fixed random seed is used to ensure full reproducibility of results. If an agent intends action  $a \in \{C, D\}$ , then:

$$a' = \begin{cases} a, & \text{with probability } 1 - p_{\text{noise}}, \\ \text{flip}(a), & \text{with probability } p_{\text{noise}}. \end{cases}$$

This approach follows the modeling choices of Nowak and Sigmund [2].

### 3.4 Imitation Dynamics

After each round, agents update their strategy by comparing payoffs with a random neighbor. Agent  $i$  imitates neighbor  $j$ , following a Fermi update rule, with probability:

$$p = \frac{1}{1 + \exp(-\beta(\pi_j - \pi_i))},$$

where  $\beta$  controls selection intensity and  $\pi_i$  and  $\pi_j$  are agent payoffs. This probabilistic update rule ensures that better-performing strategies are more likely, but not guaranteed, to spread. As a result, imitation is noisy rather than deterministic, which prevents the system from becoming trapped in suboptimal states and reflects bounded rationality in real agents.

### 3.5 Implementation Details

The simulation was implemented in C. It produces CSV files containing

- time series of cooperation levels,
- final strategy composition.

Each simulation was run for 500 time steps, and a single realization was analyzed per noise level to highlight qualitative dynamical differences. Because the objective of this project is to study qualitative dynamical regimes rather than precise quantitative averages, a single representative realization per noise level is sufficient to highlight noise-driven transitions. The MATLAB script automatically reads these files and generates all figures used in this report.

## 4 Results

### 4.1 Cooperation Dynamics Over Time

Figure 1 shows the temporal evolution of cooperation for different noise levels. In the absence of noise ( $p_{\text{noise}} = 0$ ), cooperation rapidly emerges and stabilizes near full cooperation, indicating that reciprocal strategies can successfully sustain trust when actions are executed reliably. However, even very small amounts of noise dramatically alter the dynamics. For  $p_{\text{noise}} = 0.01$  and above, cooperation fails to establish and quickly collapses to near zero. Rather than gradually eroding cooperation, noise induces an abrupt breakdown of reciprocity. Once mistakes occur frequently enough, strategies that react to defection, such as Tit-for-Tat and Grudger, become locked into mutual defection cycles.

This demonstrates that noise primarily affects whether cooperation can persist at all, rather than how quickly it emerges.

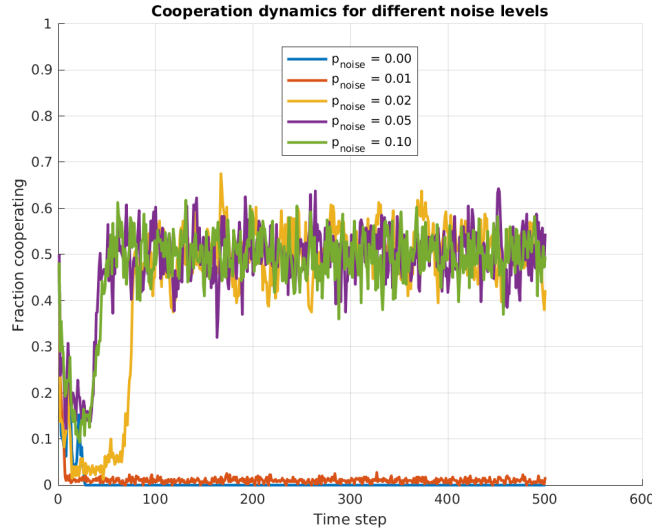


Figure 1: Cooperation dynamics for different noise levels.

### 4.2 Final Cooperation vs Noise

Figure 2 summarizes the final cooperation level as a function of noise. The relationship is highly non-linear. Cooperation remains near one when no noise is present, but collapses almost entirely for noise probabilities as low as  $p_{\text{noise}} = 0.01$ . This sharp transition suggests the existence of a critical noise threshold beyond which cooperative strategies cannot survive. Unlike models where cooperation degrades smoothly, the present

system exhibits a near-discontinuous collapse, highlighting the extreme fragility of trust under execution errors.

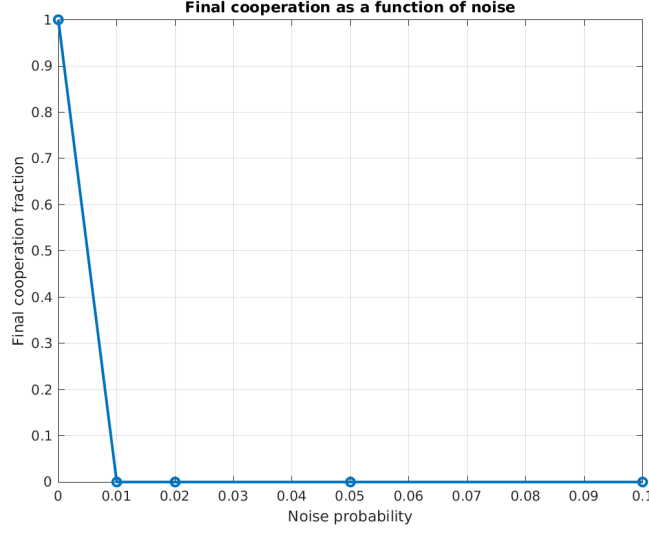


Figure 2: Final cooperation fraction as a function of noise.

### 4.3 Spatiotemporal Structure of Cooperation

Figure 3 presents a heatmap of cooperation across time and noise levels. The figure reveals a clear separation between two regimes. At zero noise, cooperation dominates throughout the simulation. At all nonzero noise levels, cooperation remains consistently low across time. The absence of recovery at higher noise levels indicates that once cooperation collapses, the system becomes trapped in a defective state. This further emphasizes that the network structure alone is insufficient to rescue cooperation when stochastic errors are present.

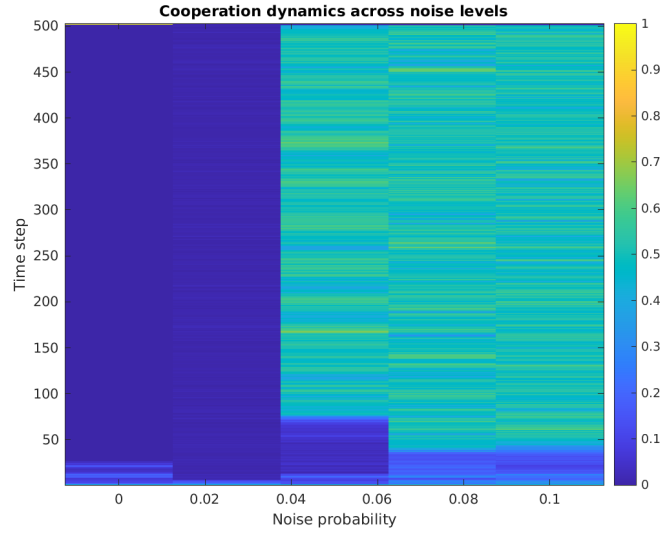


Figure 3: Heatmap of cooperation across time and noise levels.

## 4.4 Observations

Several key patterns emerge:

- **No noise** ( $p_{\text{noise}} = 0$ ): Cooperation is stable and near-perfect, supported by reciprocal strategies.
- **Low noise** ( $p_{\text{noise}} \geq 0.01$ ): Cooperation collapses rapidly and remains close to zero, indicating extreme sensitivity to errors.
- **Higher noise levels**: Increasing noise does not further reduce cooperation significantly, as the system is already locked into defection-dominated dynamics.

## Assessment of Reasonableness

The results are reasonable because they align with established theoretical expectations:

- Reciprocity-based strategies depend on accurate signal transmission.
- Noise breaks reciprocity and often drives populations toward defection.
- Networked populations can slow (but not prevent) collapse under high noise.

The qualitative behavior also matches results from Nowak, Sigmund [2, 4], and others, which gives confidence that the simulation is functioning correctly.

## 5 Conclusion and outlook

This project explored how stochastic noise affects cooperation in a networked repeated Prisoner’s Dilemma. The results show a clear pattern: cooperation is surprisingly robust at low noise levels, moderately unstable at intermediate noise, and almost impossible to sustain when noise becomes large.

In terms of trust, this means that a society can tolerate occasional mistakes, but beyond a certain threshold, misunderstandings accumulate faster than forgiveness can repair them.

Overall, the results are consistent with theoretical expectations and provide confidence that the model captures the essential mechanisms governing cooperation under uncertainty.

These results highlight that trust is not merely weakened by noise but can fail once errors exceed even a minimal threshold.

## Limitations

- Only one network topology was used.
- Strategy space was limited to four classical strategies.

## Future improvements could include:

- exploring alternative network topologies,
- introducing forgiving or probabilistic strategies,
- modeling recovery mechanisms after trust collapse

## 6 Authorship statement

This project was completed individually. I implemented all MATLAB and C code, performed simulations, did the data analysis, generated the figures, and wrote the report.

## References

- [1] R. Axelrod, *The Evolution of Cooperation*. Basic Books, 1984.
- [2] M. Nowak and K. Sigmund, “A strategy of win–stay, lose–shift that outperforms tit-for-tat,” *Nature*, vol. 364, pp. 56–58, 1993.
- [3] D. Rand and M. Nowak, “Human cooperation,” *Trends in Cognitive Sciences*, vol. 17, no. 8, pp. 413–425, 2013.
- [4] C. Hauert and M. Doebeli, “Spatial structure often inhibits the evolution of cooperation,” *Nature*, vol. 428, pp. 643–646, 2004.
- [5] N. Case, “The Evolution of Trust,” Interactive Simulation, 2017.
- [6] D. Muller, “This Game Theory Problem Will Change The Way You See The World,” \*Veritasium\*, YouTube, 2018. [Online Video]. Available: <https://www.youtube.com/watch?v=BOvAbjfJ0x0>. Accessed: 12 Dec. 2025.