A Probabilistic Model for the Mafia Party Game

Nathan Holt

Rochester Institute of Technology
nxh7119@rit.edu

March 22, 2018

Abstract

In this paper, we develop a probabilistic partial difference equation model of a simplified version of the Mafia party game. This model provides different behavior based upon a parameter representing the likelihood of the players willing to lynch a player on a given day.

I. Introduction

Mafia is a social deduction party game. Players of the game are assigned a secret role: either a citizen or a mafia member. The game then progresses in rounds, with each round corresponding to a day and night cycle. During the night, the mafia members meet in secret and select a citizen to kill. During the day, the death in the village is discovered, and the players can discuss who they believe to be members of the mafia. At the end of the day, the players can vote on who to lynch. If a consensus is reached, that player is lynched, and exits the game. Upon death, the player will announce whether the townsfolk lynched a citizen or a mafia member. The next round then starts with another day-night cycle. Gameplay progresses like this until only members of a single party win (e.g., all mafia members or all citizens).

This game is of interest to game theorists, as it can be used to study groupthink, minority vs. majority dynamics, and risk taking. Groupthink, the phenomena of a group member deciding something to conform to the rest of the group and "fit in", can be studied through Mafia by analyzing how the townsfolk make the decisions of who (if anyone) to lynch. Players will often bandwagon on a certain idea and let it obstruct their rational decision making. Risk taking is studied through Mafia by studying the likelihood that townsfolk are willing to make decisions based upon an incomplete data set (they are not sure of who the mafia member(s) are). Certain groups are willing to randomly lynch townsfolk, whereas other groups prefer to wait until more concrete evidence is given.

In this paper, I will present a probabilistic model for this simplified game of Mafia. It will be a partial difference equation that aims to capture the probability that there are a certain number of citizens and mafia members at a given round. The goal is to capture the core mechanics of mafia - the mafia killings during the night, and the possible lynching during the day - by introducing a parameter that represents the likelihood of the townsfolk to lynch someone.

II. The Model

Let $p(k, c, m)$ be the probability that there are c citizens and m mafia members at round k. We will be given an initial condition

$$p(0, c_0, m_0) = 1, \quad c_0, m_0 \in \mathbb{N}, \quad m_0 < c_0 \tag{1}$$

That is, at the beginning of the game there is some initial population c_0 of citizens, and m_0 of mafia members, with a probability of 1. Furthermore, we are subject to the constraint that there are initial more citizens than there are mafia members. In realistic games of Mafia, one can expect there to be a number of citizens more than double the number of mafia members at the beginning of the game. In this model, we only require that there are initially more citizens than there are mafia members, although we will run simulations with reasonable numbers in a later section.

In a given round of Mafia, the mafia members will kill off a citizen during the night. This happens every turn (since we are only allowing citizen and mafia roles, this is a guaranteed successful kill). During the day, the townsfolk can discuss and decide whether or not to lynch someone. Let us introduce the parameter λ_k which represents the probability of the townsfolk lynching someone during round k.

If the townsfolk decide to lynch someone at round k, there is a $\frac{c_k}{c_k+m_k}$ chance that the townsfolk lynched a citizen. Likewise, if the townsfolk decide to lynch someone at round k, there is a $\frac{m_k}{c_k+m_k}$ chance that the townsfolk lynched a mafia member. Now we have all of the components to devise our partial difference equation.

$$\begin{aligned} p(k+1, c, m) = {} & p(k, c+1, m)(1-\lambda_k) \\ & + p(k, c+2, m)\lambda_k\left(\frac{c+2}{c+m+2}\right) \\ & + p(k, c+1, m+1)\lambda_k\left(\frac{m+1}{c+m+2}\right) \end{aligned} \tag{2}$$

The first term is the probability of the townsfolk not lynching during the day at round k, multiplied by the probability that at round k, there are $c+1$ citizens, and m mafia members. The second term is the probability of lynching, multiplied by the probability of a lynched person being a citizen, multiplied

by the probability that there were 2 more citizens at round k than there are at round $k+1$. The third term is the probability of lynching, multiplied by the probability of a lynched person being a mafia member, multiplied by the proability of there being 1 more citizen and 1 more mafia member at round k than there is at round $k+1$.

This presents a well-posed problem with can be solved given an initial condition (1).

The equation (2) is thus a probabilistic model that captures the basic behavior of a simplified game of Mafia. The formulation of (2) allows for the "paranoia" parameter λ_k to vary with respect to each round, or remain fixed. It is difficult to develop a way to more explicitly define λ_k, as each player is unique, and makes lynching decisions in different ways. Some players are more likely to lynch at the beginning, some are more likely to lynch towards the end of the game. Some players are more likely to lynch when the citizens have many more than the mafia members, some are more likely to lynch when the numbers are close. As such, we can not constrain λ_k further without taking away from the "human" component of the game.

III. Mathematical Sidenotes

Analysis of this model is difficult due to the parameter λ_k and its lack of constraints beyond the fact that $\lambda_k \in [0, 1]$.

This model has no fixed points, as we are guaranteed that at each round, there is at least one death of a citizen, so the probability at any given grid point will change during every round. As such, $p(k+1, c, m) \neq p(k, c, m) \quad \forall c, m$.

This equation is dimensionless, as every quantity is a probability, which is unitless.

IV. Computational Results

We will be producing "heat maps" of the probabilities. We will run our model in a simple loop in Python and save it to a data structure. Then we will sum up the probabilities for every possible citizen-mafia pair. We will then produce a "heat map" which plots the total summed up probabilities of the given data points occuring during the game.

We will first test our model by setting the paranoia parameter λ to 0. This means that the townsfolk never vote to lynch anything (this, of course, would not happen in a real game, since this means the citizens are guaranteed to lose. However, it's a good sanity check for our model). We will simulate a game with 10 initial citizens, and 3 initial mafia members.

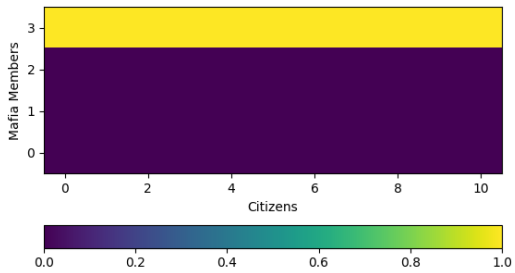

Figure 1: *A heat map from the model when there is no lynching of townsfolk.*

In figure 1, we see that there is a probability of 1 for all possible pairs of citizens and mafia members (c, m) such that $m = 3$, and $c \leq 10$. This means that in the game, no mafia members die, since it's always constant at $m = 3$, but that citizens will die off and lose the game. This matches our intuition, since if the townsfolk doesn't lynch since $\lambda = 0$, then mafia members can't die, and they are able to kill off citizens at night and win the game every time.

As such, our model reproduces the expected behavior of the game.

We will perform one other such sanity check, for when $\lambda = 1$. This means that every turn, the townsfolk will vote to lynch someone.

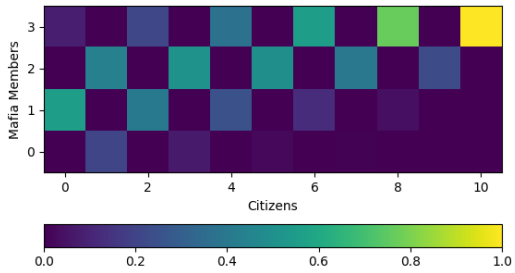

Figure 2: *A heat map from the model when* $\lambda = 1$

This produces in interesting plot which is matches intuition if thought through. Since we are given a starting population $(10, 3)$, we have probability 1 of reaching $(10, 3)$ over the course of our game. After that, we are guaranteed a lynching each turn. As such, we must move two squares to the left (they lynched a citizen) or one square left, and one square down (they lynched a mafia member). As such, this produces a checkerboard pattern, since certain states are inaccessible. Thus the model successfully reproduces the behavior expected from the game.

Now we can run our model for $\lambda = 0.5$ to represent what happens when there is a $50 - 50$ changes of the townsfolk deciding to lynch someone. Again, we will choose initial populations of 10 citizens, and 3 mafia members.

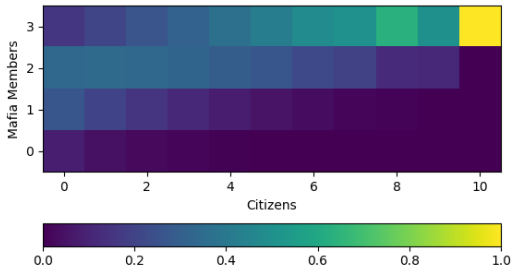

Figure 3: *A heat map from the model when* $\lambda = 0.5$

This shows the overarching trend of the game to be more favored towards mafia members, which is true in the real world. In real-world games, it is often the case that 2 or 3 times as many citizens die as do mafia members. This is due to the fact that the townsfolk do not know the identities of the mafia members, and when voting to lynch someone, the citizens outnumber the mafia members, so it's much more likely that a citizen is lynched. It is hard to quantify how well our model reproduces the real world game, as there is so much randomness inherent in the game, and this model is an oversimplification by nature. However, we see a probability of 1 at the starting location, we see a slight favoring of the game for mafia members, and we see decreasing probabilities as we approach 0 along either axis (corresponding to uncertainty in how the game will end). As such, this seems plausible.

V. Discussion

We have developed a mathematical model for the game of Mafia, in which we simplify the game down to only two roles (citizens and mafia members). Our model uses a "paranoia" factor λ which represents the likelihood to lynch on a given round. We have allowed this to be time-varying (changing as the game progresses), although we have only run simulations in which is remains constant for the entirety of the game. This allows us to see the basic behavior of various paranoia levels of players.

Our model successfully reproduces the expected outcomes of the extreme values for the parameter λ, which correspond to no lynching during the game, and lynching every round. Our model also seems to plausibly recreate the dynamics that happen when λ is a fixed value in between 0 and 1, although there's so much simplification in the model, and inherent randomness in the underlying problem we are studying, that it's hard to make any definitive, guaranteed conclusions about the success of the model in these cases.

More work is needed to be done, with several key aspects needing to be studied further. First, we would like to expand the model to incorporate other roles allowed by the game of Mafia, such as the doctor (who chooses someone to "heal" each night, and if the mafia attempts to kill the "healed" player overnight, that player is saved, and no one dies overnight that round). Second, we would like to come up with plausible ways to vary λ over time. This is a difficult task, as there are many possible ways to do it, and each player in a game may have different strategies when it comes to lynching, so it's difficult to create a definitive mathematical guideline for the value of the parameter. Also, experimental data is very scarce for Mafia (indeed, I had to gather up some friends and have them play 3 games and take notes just so I had some basis to go off of). As such, creating a data base of Mafia games would help better inform the model, and likely guide the optimal choice for a time-varying parameter λ.