

# Optimal Strategy in the Werewolf Game: A Theoretical Study

SHITONG WANG, University of Mannheim, Mannheim, Germany

---

In this article, we investigate the optimal strategies in the Werewolf Game—a widely played strategic social deduction game involving two opposing factions—from a game-theoretic perspective. We consider two scenarios: the game without a prophet and the game with a prophet. In the scenario without a prophet, we propose an enhanced strategy called “random strategy+” that significantly improves the werewolf group’s winning probability over conventional random strategies. In the scenario with a prophet, we reformulate the game as an extensive-form Bayesian game under a specific constraint, and derive the prophet’s optimal strategy that induces a Perfect Bayesian Equilibrium (PBE). This study provides a rigorous analytical framework for modeling the Werewolf Game and offers broader insights into strategic decision-making under asymmetric and incomplete information.

CCS Concepts: • **Theory of computation** → **Algorithmic game theory and mechanism design**;

Additional Key Words and Phrases: Werewolf Game, game theory, optimal strategy, random strategy+, extensive-form Bayesian game, PBE

**ACM Reference Format:**

Shitong Wang. 2025. Optimal Strategy in the Werewolf Game: A Theoretical Study. *ACM Games* 4, 1, Article 1 (December 2025), 25 pages. <https://doi.org/10.1145/3774646>

---

## 1 Introduction

### 1.1 Game Description

*Werewolf Game*, also known as *Mafia Game*, originated from a Russian social deduction game created by *Dimitry Davidoff* in 1986. It has since evolved into a widely played game of strategic reasoning, deception, and incomplete information. Typically involving 8–16 players, each is assigned a hidden role belonging to one of two opposing factions: the *citizen group* and the *werewolf group*.

The citizen group aims to eliminate all werewolves via public daytime voting, while the werewolf group strives to reduce the number of citizens to zero by secretly eliminating one player each night. To add complexity and entertainment value to the game, the citizen group is often assigned a prophet who can check one player’s group affiliation each night.

This game dynamic creates a rich environment for studying strategic decision-making under uncertainty. Players must reason not only with their own information, but also infer the intentions and beliefs of others, balancing short-term risks with long-term outcomes. In particular, the timing of the prophet’s information revelation plays a pivotal role: revealing too early risks premature death and wasted influence, while revealing too late may allow the werewolf group to seize control.

---

Author’s Contact Information: Shitong Wang, University of Mannheim, Mannheim, Germany; e-mail: shitongwang65@gmail.com.



This work is licensed under a Creative Commons Attribution 4.0 International License.

© 2025 Copyright held by the owner/author(s).

ACM 2832-5516/2025/12-ART1

<https://doi.org/10.1145/3774646>

## 1.2 Literature Review

Academic research on the Werewolf Game typically falls into three main categories:

- (1) **Probability and Game Theory:** Studying relevant strategies, equilibria, and winning rates of both groups under specific game settings.
- (2) **Social and Behavioral Sciences:** Studying phenomena in social psychology within games, such as group behavior, persuasion, and deception. This includes analyzing how players interact during the game, and examining the factors that influence their decisions and actions.
- (3) **Artificial Intelligence and Algorithms:** Using the Werewolf Game as a testbed for algorithm design, or leveraging in-game dialogue to train and evaluate artificial intelligence in assessing information authenticity.

This study models the Werewolf Game assuming fully rational agents and does not incorporate psychological or behavioral considerations. The subsequent literature review, therefore, concentrates on prior work in probability and game theory, which are most relevant to the present analysis.

Braverman et al. [2008] suggested that in the Werewolf Game without a prophet, if both the citizen group and the werewolf group adopt a random strategy, then the two groups have comparable winning probabilities when the werewolves' size is of the order of the square root of the total players' size. In the Werewolf Game with prophets, the two groups have comparable winning probabilities when the werewolves' size and total players' size are linearly related. Regarding the game without a prophet, Yao [2008] corrected the mathematical flaws of the probability boundary function in the original theorem of Braverman et al. [2008] and gave a more precise recursive formula and probability upper and lower bounds of players' winning probability. *Migdal* (2013) calculated the analytical solution of winning probability for the Werewolf Game without a prophet. The above articles (Braverman et al. [2008]; Yao [2008]; *Migdal* (2013)) all assumed that in the Werewolf Game without a prophet, "random strategy", i.e., the werewolf group eliminates a citizen randomly during the night and all players vote to eliminate a player randomly during the day is the optimal strategy for both groups, but we will revise this assumption. Bi and Tanaka [2016] calculated the Nash equilibrium of the game under certain limitations and concluded that the "stealth werewolf" strategy, i.e., werewolves pretending to be citizens, is not a good strategy. Xiong et al. [2017] used the *Game Refinement Measure*, a measure to qualify the sophistication of a game, to measure the Werewolf Game and concluded that too many players in a single game may make the game overly complex and less engaging.

This article investigates the optimal strategies for both groups in the Werewolf Game under settings with and without the presence of a prophet. For games without a prophet, our main contribution is the proposal of an improved strategy for the werewolf group, which achieves a higher winning probability than conventional approaches commonly used in practice. For games with a prophet, we develop a recursive Bayesian game model that captures the prophet's information structure and strategic considerations. Based on this model, the prophet can compute an optimal strategy under any given circumstance, including scenarios that rarely occur in standard gameplay.

## 2 Background of the Werewolf Game

In this section, we provide a formal introduction to the fundamental rules of the Werewolf Game, which lay the foundation for our subsequent modeling and analysis.

### 2.1 Role Set Configuration and Role Assignment

At the outset of the game, the total number of players and the composition of the role set (i.e., the number of players assigned to each role type) are common knowledge, while each player's specific role remains private information.

Table 1. Different Types of Roles

Role	Group	Special power
Villager	Citizen	No special power
Werewolf	Werewolf	No special power
Prophet	Citizen	Check one player's group affiliation every night

Players are typically divided into two opposing factions: the citizen group and the werewolf group. Within the citizen group, those without special power are referred to as *villagers*, while others may possess special powers. For example, a citizen who can check another player's group affiliation (citizen or werewolf) once per night is referred to as a *prophet*. Some versions of the game also permit werewolves to possess special powers, but such extensions are not considered in our model. A simple role set might be two villagers, one prophet, and two werewolves, forming a five-player game.

In some variants of the Werewolf Game, the exact role set may be uncertain. Nevertheless, even in these cases, the probability distribution over possible role sets is still common knowledge.

To facilitate subsequent analysis, players are assigned serial numbers in a clockwise order. When a player is eliminated from the game, the serial numbers of all remaining players with higher numbers are adjusted downward to close the gap. For example, if player 4 is eliminated, then player 5 becomes player 4, player 6 becomes player 5, and so on.

## 2.2 Gameplay Process

With the initial setup complete, players formally begin the game according to the established rules. Although there are many versions of the Werewolf Game, each with slightly different rule sets, we present a representative version that serves as the basis for our analysis. The game proceeds in a sequence of rounds, each consisting of a night phase followed by a day phase.

Before detailing the game procedure, we briefly introduce the possible roles in the version of the Werewolf Game considered in this article.

On the first night, the werewolves learn each other's identities. During the night phase, the werewolf group collectively selects one player to eliminate. This player is removed from the game at the beginning of the following day. If a prophet is present, they may check the group affiliation of one player each night. It is worth noting that the werewolves' elimination takes effect after the night; therefore, the prophet may end up checking the group affiliation of a player who is simultaneously being eliminated during that same night.

During the day phase, each player has one opportunity to publicly announce a message to all other players. These announcements occur simultaneously, and no private communication is permitted. Following the announcements, all players simultaneously vote to eliminate one player. The player receiving the highest number of votes is immediately removed from the game. In the event of a tie, one of the tied players is randomly chosen for elimination. The next night phase then begins.

Importantly, the group affiliations of eliminated players—whether eliminated by the werewolves or by public vote—are not revealed to the remaining players. This feature marks a key distinction between our game model and those studied in Braverman et al. [2008] and Yao [2008]. In their model, the prophet voluntarily sacrifices themselves in a specific round to validate their identity and securely convey checking information to the verified citizens, assuming the availability of private communication channels. Clearly, such a strategy is not feasible under our assumptions.

The game begins with a night phase and alternates between night and day until all players from one faction—either the citizen group or the werewolf group—are eliminated. The surviving group is then declared the winner.

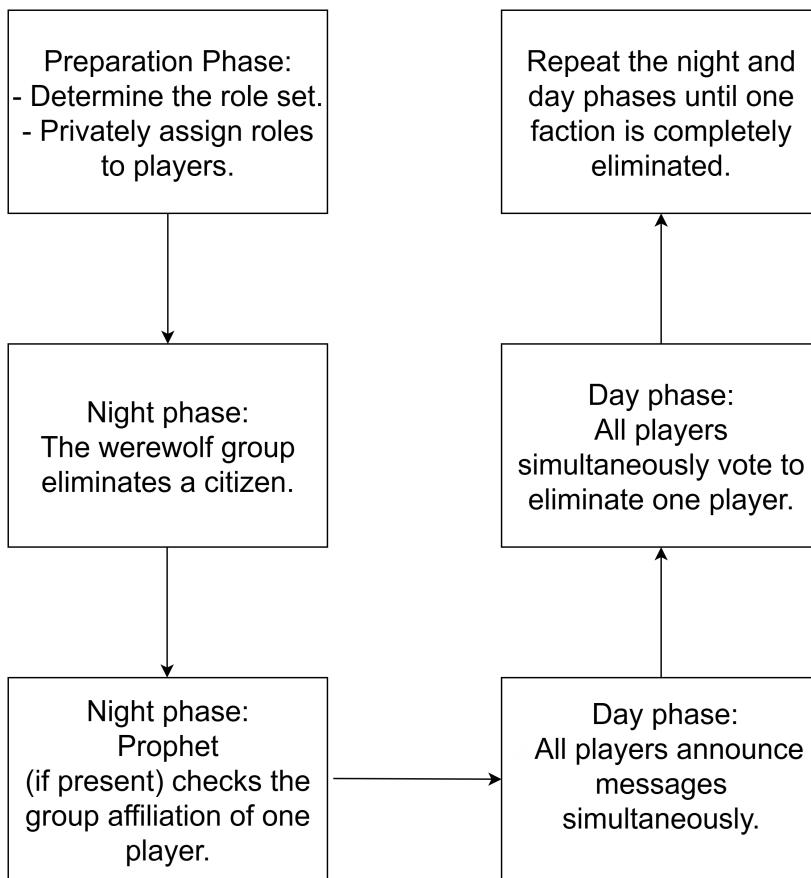


Fig. 1. Process of the Werewolf game.

### 3 Game without a Prophet

We now turn to the simplest and most widely discussed case: The Werewolf Game without a prophet. In this setting, we assume that all citizens are villagers and that all werewolves possess no special powers.

As mentioned above, in game without a prophet, many previous studies (eg., Braverman et al. [2008]; Yao [2008]; *Migdal* (2013)) have assumed that the optimal strategy for citizens (comprising only villagers) is to request all players to vote randomly during the day, while the optimal strategy for werewolves is to randomly eliminate a citizen during the night. This assumption stems from the absence of any information that would justify targeting specific players.

However, the assumption of “all players voting randomly” during the day encounters a practical challenge in real gameplay: it is difficult to verify whether players are truly voting at random. Werewolves, in particular, have incentives to avoid voting for their teammates and may collude to target specific players during the night. This behavior increases the likelihood of eliminating citizens while reducing the risk of werewolves being voted out. Nonetheless, werewolves can still claim that their votes were cast randomly—just like honest citizens—making it hard to distinguish between the two.

Fortunately, the citizen group has a natural mechanism to counter this issue. An implicit rule, often accepted by all players without the need for explicit agreement, is employed to enforce random voting. According to this rule, during the phase in which each player “simultaneously announces messages”, every player simultaneously selects a natural number. The sum of all selected numbers is then computed, and the result is taken modulo the total number of players. The player whose serial number corresponds to the result of the modulo operation is designated as the target for elimination in the subsequent vote. Any player who fails to support their declared number or votes against the modulo result is immediately identified as a werewolf and eliminated during the voting in the next round.

This mechanism ensures that even though the citizens (an uninformed majority) lack knowledge about others’ roles, they can still enforce fair and verifiable random voting. Meanwhile, the werewolves (an informed minority) are forced to conform to the rule to avoid suspicion and exposure.

More precisely, the process proceeds as follows: Each player  $i \in \{1, 2, \dots, n\}$  simultaneously selects a natural number  $x_i \in \mathbb{N}$  during the message announcement phase. All chosen numbers are then publicly disclosed, and their sum  $S = \sum_{i=1}^n x_i$  is computed. The player whose serial number equals  $S \bmod n$  (interpreting 0 as  $n$ ) is selected as the elimination target.

Formally, define

$$t = \begin{cases} S \bmod n, & \text{if } S \bmod n \neq 0, \\ n, & \text{if } S \bmod n = 0, \end{cases} \quad (1)$$

then player  $t$  becomes the designated elimination target.

The strategy described above, commonly referred to as the “random strategy”, has been regarded as optimal for both factions in prior works by Braverman et al. [2008]; Yao [2008]; and Migdal (2013). However, we propose an enhanced strategy for the werewolf group that improves their winning probability, particularly in games with a small number of players.

The improved strategy operates as follows: when the number of werewolves equals the number of remaining citizens during the voting phase, if the player designated for elimination via the modulo operation is a citizen, the werewolf group simply follows the voting rule and secures an immediate victory, as their number then exceeds that of the citizens. In contrast, if the designated player is a werewolf, the werewolf group may adopt an “all-in strategy”. This entails pre-coordinating during the previous night to unanimously vote for a specific citizen, thereby forcing a tie in the voting outcome. If the tie is resolved in favor of eliminating the targeted citizen (e.g., with probability  $\frac{1}{2}$ ), the werewolf group wins. If the werewolf is eliminated instead, the remaining werewolves eliminate a citizen during the following night and repeat the “all-in strategy” in the next round, continuing this cycle until the game concludes.

Naturally, once this strategy is employed, the citizens can infer the identities of all werewolves. In the following rounds, they can coordinate directly to vote for one of the werewolves, resulting in a tie with the werewolves’ votes.

We refer to the combination of the “random strategy” and the “all-in strategy” as the “random strategy+”. We will now demonstrate that “random strategy+” induces a **Perfect Bayesian Equilibrium (PBE)** in game without a prophet.

As discussed earlier, during the day phase, the citizen group lacks sufficient information to make informed decisions. Consequently, their only viable strategy is to ensure the integrity of the random voting process, enforced via the modulus mechanism. This makes the strategy the uniquely optimal choice for the citizen group.

We now demonstrate that “random strategy+” is also optimal for the werewolf group. Suppose that **at the end of the day phase and the beginning of the night phase**, there are  $n$  players

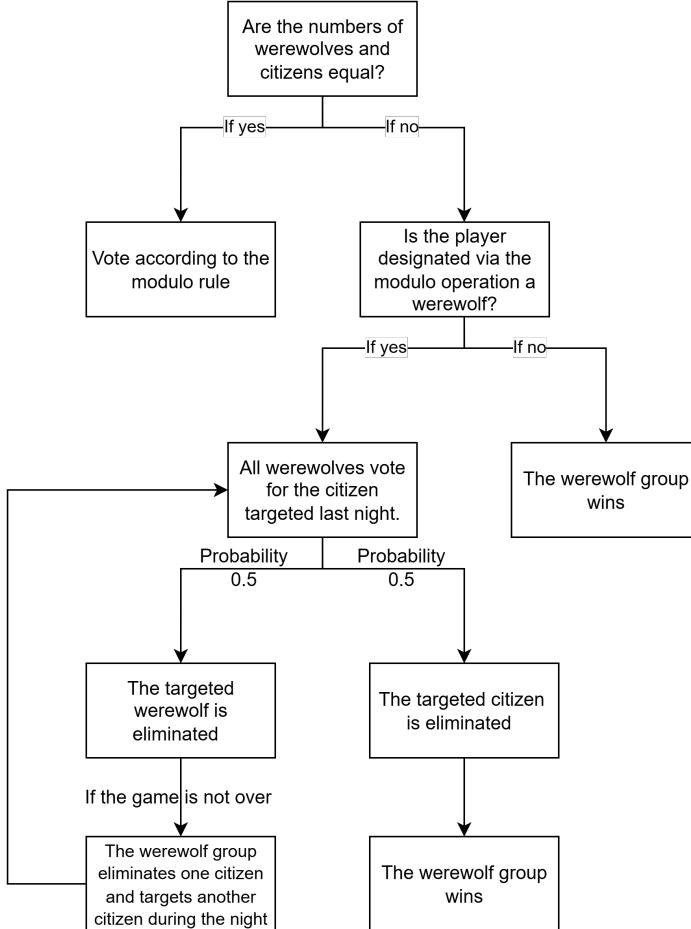


Fig. 2. Schematic diagram illustrating the werewolf group's “all-in strategy”.

remaining, of whom  $m$  are werewolves. Let  $w(n, m)$  denote the probability that the werewolf group ultimately wins the game when employing the “random strategy+”. Then  $w(n, m)$  satisfies the following recursive formula:

$$w(n, m) = \begin{cases} 0, & \text{if } m = 0, \\ 1, & \text{if } m \geq n - m, \\ 1 - \left(\frac{1}{2}\right)^{m+1} & \text{if } n - 1 = 2m, \text{ and } n \geq 5, \\ \frac{n-1-m}{n-1}w(n-2, m) + \frac{m}{n-1}w(n-2, m-1), & \text{otherwise.} \end{cases} \quad (2)$$

Interestingly, when the number of werewolves is equal to the number of citizens and the designated player is a werewolf during the voting phase, for the werewolf group, employing the “all-in strategy” immediately is mathematically equivalent to obeying the “random strategy” rule until the number of werewolves reduces to 2. From the perspective of obeying the “random strategy”

(assuming  $m > 2$ ), we observe

$$\begin{aligned} w(2m+1, m) &= \frac{2m+1-1-m}{2m} w(2m-1, m) + \frac{m}{2m+1-1} w(2m-1, m-1) \\ &= \frac{1}{2} + \frac{1}{2} w(2m-1, m-1). \end{aligned} \quad (3)$$

From the perspective of the werewolf group employing “all-in strategy” (assuming  $m > 2$ ), we have

$$w(2m+1, m) = 1 - \left(\frac{1}{2}\right)^m = \frac{1}{2} + \frac{1}{2} \left(1 - \left(\frac{1}{2}\right)^{m-1}\right) = \frac{1}{2} + \frac{1}{2} w(2m-1, m-1). \quad (4)$$

Therefore, in order to simplify the formula,  $w(n, m)$  can also be written as

$$w(n, m) = \begin{cases} 0, & \text{if } m = 0, \\ \frac{7}{8}, & \text{if } n = 5 \text{ and } m = 2, \\ 1, & \text{if } m \geq n - m, \\ \frac{n-1-m}{n-1} w(n-2, m) + \frac{m}{n-1} w(n-2, m-1), & \text{otherwise.} \end{cases} \quad (5)$$

Analogously, we can derive the recursion of the werewolf group winning probability when the werewolf group employs the “random strategy”. Suppose that **at the end of the day phase and the beginning of the night phase**, there are  $n$  players remaining, of whom  $m$  are werewolves. Let  $v(n, m)$  denote the probability that the werewolf group ultimately wins the game when employing the “random strategy”. Then  $v(n, m)$  satisfies the following recursive formula:

$$v(n, m) = \begin{cases} 0 & \text{if } m = 0 \\ 1 & \text{if } m \geq n - m \\ \frac{n-1-m}{n-1} v(n-2, m) + \frac{m}{n-1} v(n-2, m-1) & \text{otherwise.} \end{cases} \quad (6)$$

Now we prove that the “random strategy +” weakly dominates the “random strategy”:

When  $n$  is even, the “all-in strategy” would never happen, then  $w(n, m) = v(n, m)$  for all  $n, m$ . When  $n$  is odd, for all  $n \geq 5$  and  $m \geq 2$ ,  $w(n, m)$  or  $v(n, m)$  can be written as the linear form of  $w(5, 2)$  or  $v(5, 2)$ :

$$w(n_i, m_j) = c_{ij} + \alpha_{ij} \cdot w(5, 2), \quad (7)$$

$$v(n_i, m_j) = c_{ij} + \alpha_{ij} \cdot v(5, 2). \quad (8)$$

Since we get  $w(5, 2) = \frac{7}{8} \geq v(5, 2) = \frac{3}{4}$ , then  $w(n, m) \geq v(n, m)$  for all odd  $n \geq 5$  and  $m \geq 2$ . Easy to verify, when  $n < 5$  or  $m < 2$ ,  $w(n, m) = v(n, m)$ .

In summary,  $w(n, m) \geq v(n, m)$  for all  $n$  and  $m$ . We have proven that the “random strategy+” weakly dominates the “random strategy” in all cases for the werewolf group. The “random strategy+” is the optimal strategy for both groups and thus can indeed induce a PBE. The following figure illustrates the difference in the werewolf group’s winning probability under these two strategies. Bar colors are used solely for visual clarity.

Finally, we prove that the werewolf group’s strategy of self-killing during the night is strictly dominated. This proof is necessary because, in all previous analyses, we implicitly assumed that the werewolves would not adopt such a strategy. Providing a formal justification for this assumption is essential to completing the proof that the “random strategy+” is indeed the optimal strategy for the werewolf group.

**CLAIM 1.** *In a game without a prophet, the werewolf group’s strategy of killing themselves during the night is strictly dominated.*

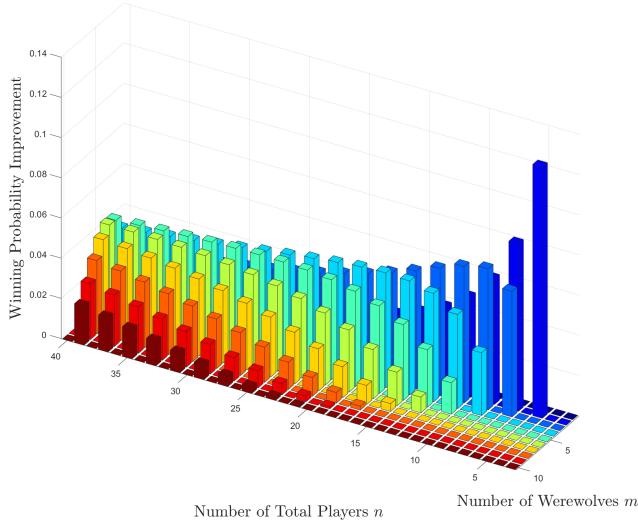


Fig. 3. Winning probability improvement for the werewolf group under the “random strategy+” compared to the “random strategy”.

PROOF. For  $n \geq 2m + 1 \geq 7$ , we have the following equation from (5):

$$(n-1-m)[w(n-2, m) - w(n, m)] = m[w(n, m) - w(n-2, m-1)]. \quad (9)$$

Rearranging the equation, we get:

$$(n-1)[w(n, m) - w(n-2, m-1)] = (n-1-m)[w(n-2, m) - w(n-2, m-1)]. \quad (10)$$

Therefore,

$$w(n-2, m) > w(n, m) \iff w(n, m) > w(n-2, m-1) \iff w(n-2, m) > w(n-2, m-1). \quad (11)$$

Suppose the werewolf group eliminates one of their own members during the night exactly once, and then proceeds with the “random strategy+” afterward. Let  $w'(n, m)$  denote their winning probability under this modified strategy. Then

$$w'(n, m) = \frac{m-1}{n-1}w(n-2, m-2) + \frac{n-m}{n-1}w(n-2, m-1). \quad (12)$$

We now compare  $w(n, m)$  and  $w'(n, m)$ :

$$\begin{aligned} w(n, m) - w'(n, m) &= \left(\frac{n-m-1}{n-1}\right)w(n-2, m) - \left(\frac{m-1}{n-1}\right)w(n-2, m-2) \\ &\quad + \left(\frac{2m-n}{n-1}\right)w(n-2, m-1). \end{aligned} \quad (13)$$

Since

$$w(n-2, m) > w(n-2, m-1) > w(n-2, m-2), \quad (14)$$

it follows that

$$w(n, m) - w'(n, m) > 0. \quad (15)$$

The remaining cases with smaller values of  $n$  or  $m$  can be easily verified through direct enumeration.

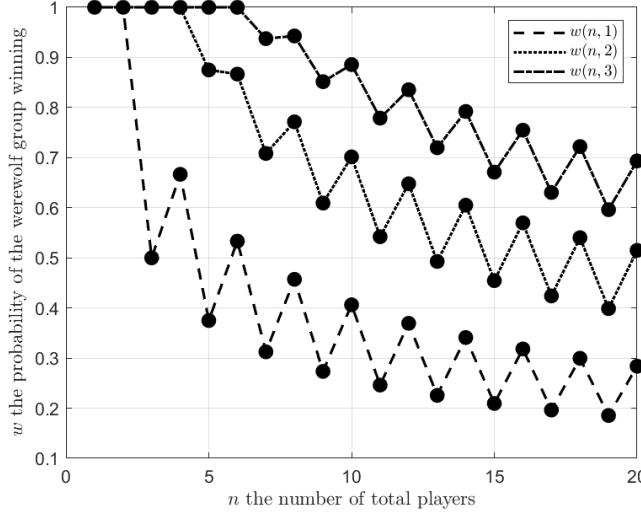


Fig. 4. Winning probability of the werewolf group in a game without a prophet.

Therefore, we conclude that for the werewolf group, any strategy involving self-killing during the night is strictly dominated. This completes the justification for the “random strategy+” as the optimal strategy.

We now conduct a quantitative analysis of the winning probability of the werewolf group employing the “random strategy+”.

From the analysis above, we have established that  $w(n, m) > w(n, m-1)$  and  $w(n+2, m) < w(n, m)$ . However, the relationship between  $w(n, m)$  and  $w(n-1, m)$  remains undetermined.

To illustrate this, we present a line plot showing the winning probability of the werewolf group in games with 1 to 3 werewolves and up to 20 total players.

It is easy to observe from the figure that  $w(2k, m) > w(2k-1, m)$  when  $k \geq 4$  and  $k > m$ . A similar property also holds under the “random strategy”, as noted by *Migdal* (2013).

From a game-theoretic perspective, we offer the following intuitive explanation: when the total number of players  $n$  is odd, adding one more citizen (making  $n$  even) does not trigger an additional round of voting. However, this added citizen reduces the per-round probability that a werewolf is voted out. Consequently, for a fixed number of werewolves  $m$ , the werewolf group’s winning probability increases as the total number of players changes from an odd to the next even number.

Now, we proceed to prove by mathematical induction that for all integers  $k \geq 4$  and  $m < k$ , the inequality  $w(2k, m) > w(2k-1, m)$  holds.

*Base Case.* Let  $k = 4$ . We verify the inequality for  $m = 1, 2, 3$  by explicit computation:

$$w(8, 1) = \frac{48}{105} > w(7, 1) = \frac{5}{16}, \quad (16)$$

$$w(8, 2) = \frac{27}{35} > w(7, 2) = \frac{39}{55}, \quad (17)$$

$$w(8, 3) = \frac{33}{35} > w(7, 3) = \frac{15}{16}. \quad (18)$$

Thus, the base case holds.

*Inductive Hypothesis.* Assume that for some  $k \geq 4$ , and for all  $m < k$ , we have:

$$w(2k, m) > w(2k - 1, m). \quad (19)$$

*Inductive Step.* We aim to prove:

$$w(2k + 2, m) > w(2k + 1, m), \quad \text{for all } m < k + 1. \quad (20)$$

We first show that the inequality holds for all  $m < k$ . From the recurrence relations, we have:

$$w(2k + 2, m) = \frac{2k + 1 - m}{2k + 1} w(2k, m) + \frac{m}{2k + 1} w(2k, m - 1), \quad (21)$$

$$w(2k + 1, m) = \frac{2k - m}{2k} w(2k - 1, m) + \frac{m}{2k} w(2k - 1, m - 1). \quad (22)$$

By the inductive hypothesis:

$$w(2k, m) > w(2k - 1, m) > w(2k - 1, m - 1), \quad (23)$$

$$w(2k, m - 1) > w(2k - 1, m - 1). \quad (24)$$

Moreover, the coefficients satisfy:

$$\frac{2k + 1 - m}{2k + 1} > \frac{2k - m}{2k}. \quad (25)$$

Combining these observations, we conclude:

$$w(2k + 2, m) > w(2k + 1, m), \quad \text{for all } m < k + 1. \quad (26)$$

We now consider the boundary case  $m = k$ . When  $k \geq 4$ , we have:

$$w(2k + 2, k) = \frac{k + 1}{2k + 1} w(2k, k) + \frac{k}{2k + 1} w(2k, k - 1) = \frac{k + 1}{2k + 1} + \frac{k}{2k + 1} w(2k, k - 1), \quad (27)$$

$$w(2k + 1, k) = \frac{1}{2} w(2k - 1, k) + \frac{1}{2} w(2k - 1, k - 1) = \frac{1}{2} + \frac{1}{2} w(2k - 1, k - 1). \quad (28)$$

To establish  $w(2k + 2, k) > w(2k + 1, k)$ , it suffices to show:

$$w(2k, k - 1) > w(2k - 1, k - 1). \quad (29)$$

Since  $w(8, 3) > w(7, 3)$ , the argument extends analogously to all larger  $k$ .

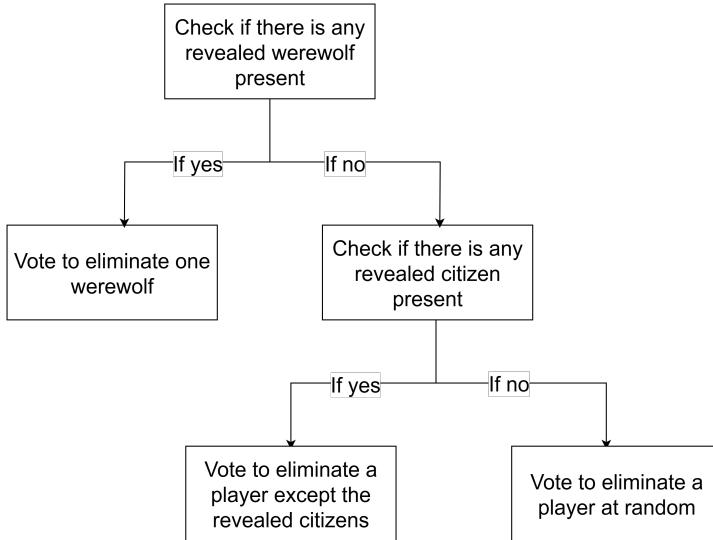


Fig. 5. The process of werewolf group choosing the eliminating target.

Therefore, by the principle of mathematical induction, we conclude that for all  $k \geq 4$  and all  $m < k$ , the inequality  $w(2k, m) > w(2k - 1, m)$  holds.

#### 4 Game with a Prophet

In this section, we focus on the game with a prophet and aim to derive the optimal strategies for both groups. These strategies collectively induce a PBE.

##### 4.1 Game under Honesty Rule

First, let us consider a simplified but instructive game setting. Assume that neither the villagers nor the werewolves can convey false information when announcing messages publicly during the day. We refer to this constraint as the *honesty rule*. Under this rule, it is evident that before the prophet actively reveals the checked information, the “random strategy +” discussed previously is the optimal strategy for both groups. Once the prophet reveals all information obtained from previous nights in a given round, due to the restriction that neither werewolves nor villagers can impersonate the prophet, all players will recognize the identity of the prophet and base their subsequent actions on the revealed information.

For the villagers, this means urging all players to vote out the revealed werewolves. Once all revealed werewolves are voted out, villagers will revert to random voting, excluding those who have been verified to be villagers by the prophet. The werewolves, in contrast, would prioritize eliminating the revealed prophet during the first night to prevent further information revelation. Subsequently, they will focus on eliminating the checked villagers during future nights, thereby reducing the likelihood of werewolves being voted out during the day. Naturally, if the number of werewolves equals the number of villagers and one of the werewolves is about to be voted out, the werewolf group will decisively employ the “all-in strategy”, as they do in the game without a prophet.

We illustrate the whole game process with a representative example as follows:

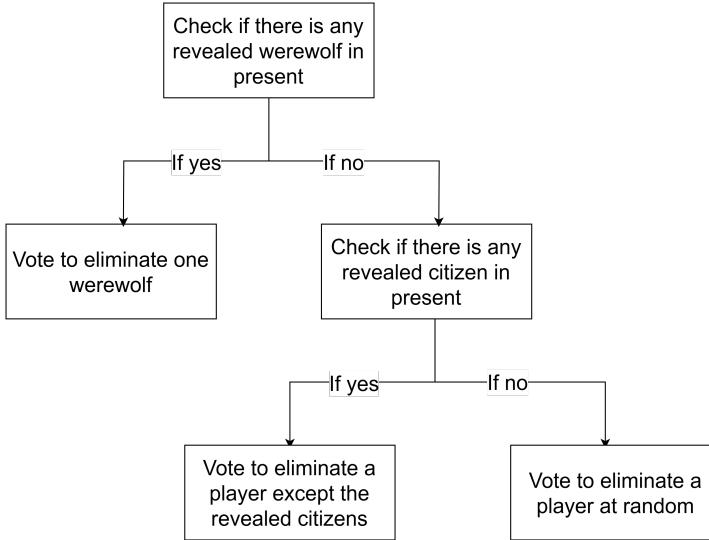


Fig. 6. The process of the citizen group choosing the voting target.

Table 2. One Possible Process of Game with 9 Villagers, 2 Werewolves, and 1 Prophet

<b>Initial role assignment:</b> Player 1: Villager, Player 2: Villager, Player 3: Villager, Player 4: Werewolf, Player 5: Villager, Player 6: Villager, Player 7: Villager, Player 8: Villager, Player 9: Werewolf, Player 10: Villager, Player 11: Villager, Player 12: Prophet
Night: The Werewolf group eliminated Player 6 (reason: random eliminating citizen)
Remaining Player Profile: 8 Villagers, 2 Werewolves, and 1 Prophet
Night: The Prophet checked Player 11, result: Villager
Day: Player 2 was voted out (reason: random voting out)
Remaining Player Profile: 7 Villagers, 2 Werewolves, and 1 Prophet
Night: The Werewolf eliminated Player 10 (reason: random eliminating citizen)
Remaining Player Profile: 6 Villagers, 2 Werewolves, and 1 Prophet
Night: Player 12, the Prophet checked Player 4, result: Werewolf
Day: The Prophet revealed information: Player 4 is a Werewolf, Player 11 is a Villager
Day: Player 4 was voted out (reason: voting out the revealed werewolf)
Remaining Player Profile: 6 Villagers, 1 Werewolves, and 1 Prophet
Night: The Werewolf group killed Player 12 (reason: prioritizing eliminating the revealed prophet)
Remaining Player Profile: 6 Villagers, 1 Werewolves, and 0 Prophet
Day: Player 1 was voted out (reason: random voting out except the revealed Villager Player 11)
Remaining Player Profile: 5 Villagers, 1 Werewolves, and 0 Prophet
Night: The Werewolf eliminated Player 11 (reason: prioritizing eliminating the revealed villager)
Remaining Player Profile: 4 Villagers, 1 Werewolves, and 0 Prophet
Day: Player 4 was voted out (reason: random voting out)
Remaining Player Profile: 4 Villagers, 0 Werewolves, and 0 Prophet
<b>Game Over: The Citizen group won!</b>

In this game, the prophet fortunately checked a werewolf and revealed it in time, which played a crucial role in the final victory of the citizen group.

#### 4.2 Rule of Thumb in Revealing Information

The central strategic challenge is to determine the optimal timing for the prophet to reveal their verified information. We begin by considering a fixed, *ex-ante* strategy in which the prophet commits in advance to disclosing all known information on the  $x$ th day of the game.

Table 3. Best Round  $f(h, m)$  of Prophet Revealing Information and the Citizen Group’s Winning Probability (with 95% CI)

Citizen Group Winning Probability (95% CI)		Number of werewolves			
		1	2	3	4
Number of villagers	4	R1, <b>70%</b> (68.9, 70.7)	R2, <b>37%</b> (35.8, 37.6)	R2, <b>17%</b> (16.3, 17.7)	R1, <b>5%</b> (4.5, 5.3)
	5	R2, <b>74%</b> (72.5, 74.2)	R2, <b>46%</b> (45.0, 46.9)	R2, <b>22%</b> (20.7, 22.4)	R3, <b>8%</b> (7.9, 8.9)
	6	R2, <b>77%</b> (76.4, 78.1)	R2, <b>49%</b> (48.3, 50.2)	R3, <b>27%</b> (26.0, 27.8)	R2, <b>12%</b> (10.9, 12.1)
	7	R2, <b>76%</b> (75.3, 77.0)	R3, <b>52%</b> (51.1, 53.1)	R3, <b>31%</b> (30.3, 32.2)	R3, <b>14%</b> (13.7, 15.1)
	8	R3, <b>76%</b> (75.4, 77.0)	R3, <b>56%</b> (55.0, 56.9)	R4, <b>33%</b> (32.4, 34.3)	R3, <b>19%</b> (17.9, 19.4)
	9	R3, <b>78%</b> (77.5, 79.1)	R3, <b>56%</b> (54.7, 56.7)	R3, <b>36%</b> (35.3, 37.2)	R4, <b>23%</b> (21.7, 23.4)
	10	R3, <b>79%</b> (78.3, 79.9)	R4, <b>58%</b> (57.0, 58.9)	R4, <b>40%</b> (38.5, 40.4)	R4, <b>26%</b> (24.7, 26.4)
	11	R3, <b>79%</b> (78.2, 79.8)	R4, <b>60%</b> (58.8, 60.8)	R4, <b>42%</b> (40.6, 42.5)	R5, <b>27%</b> (26.4, 28.1)
	12	R4, <b>80%</b> (79.0, 80.6)	R5, <b>61%</b> (60.2, 62.1)	R5, <b>44%</b> (42.9, 44.9)	R5, <b>30%</b> (29.1, 30.9)

Given a configuration  $(h, m)$ , where  $h$  denotes the number of villagers and  $m$  the number of werewolves, our objective is to define a mapping that maximizes the expected winning probability of the citizen group when the prophet reveals all checked information on day  $x$ .

Let  $H(h, m, x)$  denote the probability that the citizen group wins given  $h$  villagers,  $m$  werewolves, and a prophet who reveals all previously obtained information on day  $x$ . We define a function  $f : \mathbb{N}^2 \rightarrow \mathbb{N}$ , where the input  $(h, m)$  returns the optimal revelation round  $x \in \mathbb{N}$ . Formally:

$$f(h, m) = \arg \max_{x \in \mathbb{N}} \mathbb{E}[H(h, m, x)].$$

In essence, the mapping  $f(h, m)$  identifies the round  $x$  that maximizes the citizen group’s ex-ante probability of winning.

To evaluate the ex-ante optimal strategy for the prophet’s revelation, we conduct **10,000** independent Monte Carlo simulations for each  $(h, m)$  configuration. For each combination, we estimate the expected citizen winning probability under different revelation rounds and select the round that yields the highest expected probability. To account for sampling variability, we compute a 95% **confidence interval (CI)** for the citizen winning probability using the standard error of a Bernoulli process. The simulation code is available at: <https://zenodo.org/records/16366976>

In standard gameplay, even if the prophet does not strive to follow an interim optimal strategy, simply adhering to the “rule of thumb” outlined in the table above significantly increases the citizen group’s likelihood of winning, compared to scenarios where no prophet is present. Here, we provide a comparison between games with and without a prophet, given the same number of citizens.

Judging from the comparison results, the presence of a prophet significantly increases the winning probability of the citizen group.

#### 4.3 PBE under Honesty Rule

We aim to construct a strategy that induces a PBE and maximizes the citizen group’s winning probability under the honesty rule. To achieve this, we observe that both before and after the prophet reveals information, all players—including villagers and werewolves—adhere to predetermined strategic patterns as outlined in Section 4.1. Hence, the entire game can be equivalently reformulated as a dynamic game with incomplete information involving only the prophet as an active player. For the prophet, there exist multiple possible role profiles of other players in present (referred

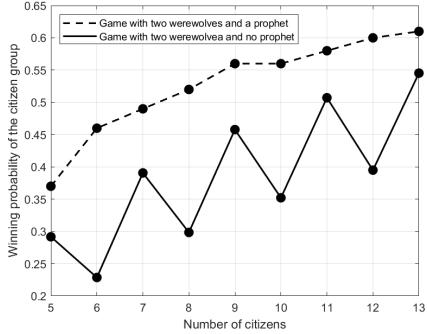


Fig. 7. Winning probabilities of the citizen group with and without a prophet in games with two werewolves.

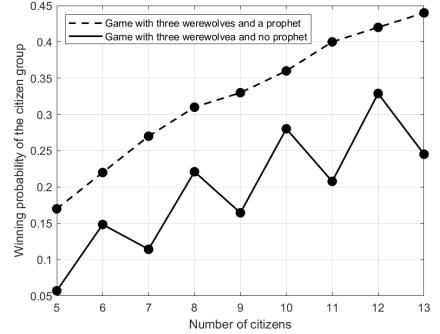


Fig. 8. Winning probabilities of the citizen group with and without a prophet in games with three werewolves.

to as *nodes*), and the probabilistic distribution over these profiles defines the prophet's current information set.

Let us denote the information set  $I_t$  possessed by the prophet (if still present) regarding the role profiles of other players at round  $t$  before the message announcement phase as

$$I_t = \sum_i \alpha_i (H_i, M_i, h_i, m_i), \quad (30)$$

where

- $(H_i, M_i, h_i, m_i)$  represents the different nodes in this information set.
- $H_i$  denotes the number of remaining villagers in node  $(H_i, M_i, h_i, m_i)$ .
- $M_i$  denotes the number of remaining werewolves in node  $(H_i, M_i, h_i, m_i)$ .
- $h_i$  denotes the number of checked villagers in node  $(H_i, M_i, h_i, m_i)$ .
- $m_i$  denotes the number of checked werewolves in node  $(H_i, M_i, h_i, m_i)$ .
- $\alpha_i$  is the probability of node  $(H_i, M_i, h_i, m_i)$  in information set  $I_t$ , with  $\sum_i \alpha_i = 1$ .

Any two nodes  $(H_i, M_i, h_i, m_i)$  and  $(H_j, M_j, h_j, m_j)$  within the same information set must satisfy the following conditions:

$$H_i + M_i = H_j + M_j, \quad (31)$$

$$h_i = h_j, \quad (32)$$

$$m_i = m_j. \quad (33)$$

For example, suppose a prophet currently holds the information set

$$I_t = \frac{1}{3}(5, 2, 1, 1) + \frac{2}{3}(6, 1, 1, 1).$$

Here,  $(5, 2, 1, 1)$  and  $(6, 1, 1, 1)$  represent different nodes in information set  $I_t$ . This means the prophet knows that among the remaining 7 players, there is a  $\frac{1}{3}$  chance of having 5 villagers and 2 werewolves, and a  $\frac{2}{3}$  chance of having 6 villagers and 1 werewolf. In either case, the prophet has already identified one specific player as a villager and another as a werewolf.

The requirement that  $h_i = h_j$  and  $m_i = m_j$  holds is straightforward: if the number of confirmed villagers and confirmed werewolves differs across nodes, then the prophet could distinguish between them. As a result, these nodes would not belong to the same information set.

Suppose there exists a mapping  $g : I_t \rightarrow \mathcal{A}$ , where  $\mathcal{A} = \{\text{Hiding}, \text{Revealing}\}$ . Then the prophet's optimal action is determined by

$$g(I_t) = \arg \max_{x \in \mathcal{A}} R(x, I_t), \quad (34)$$

where  $R(x, I_t)$  represents the winning probability of the citizen group if the prophet chooses action  $x$  given the information set  $I_t$ . For notational convenience, we denote  $R(g(I_t), I_t) := R(g, I_t)$ .

The detailed construction of  $R(\text{Hiding}, I_t)$  and  $R(\text{Revealing}, I_t)$  is provided in the Appendix. In summary, we have the analytical formula of  $R(\text{Revealing}, I_t)$ :

$$R(\text{Revealing}, I_t) = s(I_t) = s \left( \sum_i \alpha_i \cdot (H_i, M_i, h_i, m_i) \right) = \sum_i \alpha_i s(H_i, M_i, h_i, m_i), \quad (36)$$

$$s(H, M, h, m) = \begin{cases} \frac{M}{H-h+M} \left(\frac{1}{2}\right)^M, & \text{if } M = H + 1 \text{ and } m = 0, \\ \left(\frac{1}{2}\right)^M, & \text{if } M = H + 1 \text{ and } m \geq 1, \\ 1 - w(H + M + 2 - 2m, M - m), & \text{if } M < H + 1 \text{ and } m \geq h + 1, \\ u(H + 1 - m, M - m, h + 1 - m), & \text{if } M < H + 1 \text{ and } h + 1 > m \geq 1, \\ \frac{M}{H-h+M} u(H, M - 1, h) + \frac{H-h}{H-h+M} u(H - 1, M, h), & \text{if } M < H + 1 \text{ and } m = 0, \end{cases} \quad (37)$$

where

$$u(H', M', h') = \begin{cases} 1, & \text{if } M' = 0, \\ \frac{H'-h'}{H'+M'-h'} u(H' - 2, M', h' - 1) + \frac{M'}{H'+M'-h'} u(H' - 1, M' - 1, h' - 1), & \text{if } h' \geq 1, M' \geq 1, \text{ and } H' > M', \\ 1 - w(H' + M' + 1, M'), & \text{if } h' = 0, M' \geq 1, \text{ and } H' > M', \\ \frac{M'}{H'-h'+M'} \left(\frac{1}{2}\right)^{M'}, & \text{if } H' = M' \text{ and } M' \geq 1, \\ 0, & \text{if } H' < M' \text{ and } M' \geq 1. \end{cases} \quad (38)$$

Through recursive computation, all expressions of  $R(\text{Hiding}, I_t)$  can eventually be written in the following form:

$$R(\text{Hiding}, I_t) = e(I_t) + \sum_r \gamma_r R(g, I_{t+1}^r), \quad (39)$$

where  $e(I_t)$  denotes a constant term that differs for each information set  $I_t$ , and  $I_{t+1}^r$  denotes potential information sets in round  $t + 1$ .

In the above discussion, we assumed that the self-killing strategy is not viable for the werewolf group. Now we need a rigorous proof to demonstrate that the proposed strategy for the prophet,  $g(I_t)$ , indeed induces a PBE.

**CLAIM 2.** *Under the honesty rule, the strategy in which the werewolf group deliberately eliminates one of its own members is strictly dominated.*

The proof of Claim 2 is provided in the Appendix.

#### 4.4 Example

Now, we give an example in which we solve the strategy for a prophet in a specific information set inducing PBE.

Suppose the prophet is in the information set  $I_t = \frac{2-\sqrt{2}}{2}(5, 1, 1, 1) + \frac{\sqrt{2}}{2}(4, 2, 1, 1)$ . In general games, this information set is impossible to exist, but this does not prevent us from calculating the optimal strategy of the prophet under this information set.

First, we calculate  $R(\text{Revealing}, I_t) = s(I_t) = 0.6464$  through (37) and (38). Then we calculate  $R(\text{Hiding}, I_t) = 0.0918 + \sum_{r=1}^{11} \gamma_r R(g, I_{t+1}^r)$  through (39). The reason why  $r$  goes up to 11 is that one

Table 4. Summary of Information Sets  $I_{t+1}^r$ 

$r$	$\gamma_r$	Nodes in $I_{t+1}^r$	Coefficients of nodes in $I_{t+1}^r$
$r$	$\gamma_r$	Nodes in $I_{t+1}^r$	Coefficients of nodes in $I_{t+1}^r$
1	0.0273	(3,1,0,1), (2,2,0,1)	0.3064, 0.6936
2	0.0630	(3,1,1,1), (2,2,1,1)	0.3986, 0.6014
3	0.0189	(2,2,0,2)	1
4	0.0443	(4,0,1,0), (3,1,1,0)	0.3152, 0.6848
5	0.0734	(4,0,2,0), (3,1,2,0)	0.3803, 0.6197
6	0.0050	(3,1,0,1)	1
7	0.0152	(3,1,1,1)	1
8	0.2084	(3,1,1,1), (2,2,1,1)	0.5152, 0.4848
9	0.1579	(3,1,2,1), (2,2,2,1)	0.6800, 0.3200
10	0.0505	(2,2,1,2)	1
11	0.0253	(2,2,0,2)	1

information set constructed by nodes  $(3, 1, -1, 2)$  and  $(2, 2, -1, 2)$  is impossible.

After further calculation, we obtain that

$$g(I_{t+1}^r) = \text{Revealing}$$

for all  $r = 1, 2, \dots, 11$ . The corresponding values are:

$$\begin{aligned} R(\text{Revealing}, I_{t+1}^1) &= 0.5916, & R(\text{Revealing}, I_{t+1}^2) &= 0.6993, & R(\text{Revealing}, I_{t+1}^3) &= 1, \\ R(\text{Revealing}, I_{t+1}^4) &= 0.7717, & R(\text{Revealing}, I_{t+1}^5) &= 1, & R(\text{Revealing}, I_{t+1}^6) &= 1, \\ R(\text{Revealing}, I_{t+1}^7) &= 1, & R(\text{Revealing}, I_{t+1}^8) &= 0.7576, & R(\text{Revealing}, I_{t+1}^9) &= 1, \\ R(\text{Revealing}, I_{t+1}^{10}) &= 1, & R(\text{Revealing}, I_{t+1}^{11}) &= 1. \end{aligned}$$

Hence, we compute:

$$R(\text{Hiding}, I_t) = 0.0918 + \sum_{ii=1}^{11} \beta_{ii} R(g, I_{t+1}^{ii}) = 0.6903 > R(\text{Revealing}, I_t) = 0.6464.$$

Therefore, in the information set  $I_t = \frac{2-\sqrt{2}}{2}(5, 1, 1, 1) + \frac{\sqrt{2}}{2}(4, 2, 1, 1)$ , *Hiding* is an optimal strategy.

## 5 Conclusion

This study presents a rigorous analytical framework for modeling the Werewolf Game and opens promising directions for research on decision-making under asymmetric and incomplete information. The mechanisms and equilibria derived—particularly in contexts involving strategic signaling and partial observability—reflect challenges found in real-world domains such as cybersecurity, political negotiation, and competitive markets.

Moreover, the Werewolf Game embodies core features shared by a wide class of social deduction games: asymmetric information, hidden roles, strategic deception, and public communication. As such, the equilibrium analysis and signaling mechanisms developed here are not limited to the Werewolf Game but offer a general modeling framework applicable to other games like *The Resistance*, *Secret Hitler*, and *Among Us*. For instance, roles such as Liberals and Fascists in *Secret Hitler*, or Spies in *The Resistance*, mirror the citizen/werewolf dichotomy in the Werewolf Game. Across these games, strategic communication—often involving limited opportunities for truthful

revelation or deliberate misdirection through voting—is a central mechanic. This makes our prophet signaling model broadly applicable in capturing how private information can be credibly conveyed under constraints.

In addition, the analytical tools employed in this study—such as PBE—are suitable for any dynamic game involving evolving information sets and belief-based strategy selection. The algorithm developed to compute optimal strategies for the prophet, grounded in recursive belief updates and decision trees, is also transferable to other roles in similar games. For example, roles like the Engineer or Scientist in *Among Us*, who possess private insights but limited communication opportunities, face decision problems structurally analogous to the prophet's.

Finally, we propose several open questions to guide future research efforts:

#### (1) **Multiple Prophets**

The current analysis focuses primarily on scenarios involving a single prophet. When two or more prophets participate, the game dynamics become substantially more intricate. Key questions arise regarding how the prophets should coordinate their revelation strategies to maximize their collective effectiveness. Furthermore, what would characterize an optimal equilibrium strategy for the citizen group in this multi-prophet setting?

#### (2) **Absence of Honesty Rule**

Without the honesty rule, players are no longer bound to truthful public communication. This dramatically enlarges the strategic space, as all messages may be deceptive. Even in such more complex or realistic settings, the prophet's decision of when and with what probability to reveal truthful information remains critical. In a sequential speaking framework, the timing and order of communication may significantly influence the outcome. For the citizen group, a natural fallback strategy is to disregard all messages and revert to the “random strategy+” employed in prophet-free games. However, this raises several further inquiries:

- Are there alternative strategies that can outperform this fallback approach in terms of citizen winning probability? If such strategies exist, should they be pure or mixed, depending on the player's role and specific conditions? Intuitively, mixed strategies might provide greater robustness in deceptive environments, but formal analysis is required to substantiate this.
- How does the order of speaking affect the information structure and confer strategic advantages to different players?

#### (3) **Introduction of Additional Roles**

Beyond werewolves, villagers, and prophets, common gameplay often includes other roles such as the guard, who can protect a player from elimination by werewolves during the night. How do these additional roles alter the strategic landscape and dynamics of the game?

### **Declarations**

### **Funding and/or Conflicts of interests**

The author did not receive support from any organization for the submitted work. The author has no relevant financial or non-financial interests to disclose.

### **Data availability**

All data we demonstrate and analyze is generated through theoretical and mathematical approaches in this article.

### **Acknowledgments**

I would like to thank Professor Ding Li, who introduced me to the fascinating world of game theory.

## References

Wikipedia. 2024. Mafia (Party Game). *Wikipedia, the free encyclopedia*. [Online]. Retrieved June 1, 2024 from [https://en.wikipedia.org/wiki/Mafia\\_\(party\\_game\)](https://en.wikipedia.org/wiki/Mafia_(party_game))

Mark Braverman, Omid Etesami, and Elchanan Mossel. 2008. Mafia: A theoretical study of players and coalitions in a partial information environment. In *Proceedings of the 3rd International Conference on Algorithmic Game Theory*. 825–846.

E. Yao. 2008. A theoretical study of mafia games. *arXiv preprint arXiv:0804.0071*.

P. Migdał. 2010. A mathematical model of the Mafia game. *arXiv preprint arXiv:1009.1031*.

Xiaoheng Bi and Tetsuro Tanaka. 2016. Human-side strategies in the werewolf game against the stealth werewolf strategy. In *Proceedings of the International Conference on Computers and Games*.

S. Xiong, W. Li, X. Mao, and H. Iida. 2017. Mafia game setting research using game refinement measurement. In *International Conference on Advances in Computer Entertainment*. Cham: Springer International Publishing, 830–846.

H. Ri, X. Kang, M. N. A. Khalid, and H. Iida. 2022. The dynamics of minority versus majority behaviors: A case study of the mafia game. *Information* 13, 3 (2022), 134.

H. Tanioka and R. Kohri. 2021. Improving the winning percentage of the werewolf team through collusion strategies. In *Proceedings of the 2021 10th International Congress on Advanced Applied Informatics (IIAI-AAI)*. IEEE, 944–945.

John F. Nash. 1950. Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences*, 36, 1 (1950), 48–49.

John F. Nash. 1951. Non-cooperative games. *Annals of Mathematics* 54, 2 (1951), 286–295.

John C. Harsanyi. 1967. Games with incomplete information played by 'Bayesian' players, I-III part I. *Management Science* 14, 3 (1967), 159–182.

Reinhard Seltен. 1965. Spieltheoretische behandlung eines oligopolmodells mit nachfrageträgheit. *Zeitschrift für die Gesamte Staatswissenschaft/Journal of Institutional and Theoretical Economics* 121, 2 (1965), 301–324.

## Appendix

### Construction of $R(x, I_t)$

Denote  $R(\text{Revealing}, I_t) = s(I_t)$ . After the prophet reveals the information, the winning probability of the citizen group is the weighted sum of winning probabilities in each node, because the same action patterns are adopted by both the citizen group and the werewolf group:

$$s(I_t) = s\left(\sum_i \alpha_i \cdot (H_i, M_i, h_i, m_i)\right) = \sum_i \alpha_i s(H_i, M_i, h_i, m_i). \quad (35)$$

We consider two different cases of  $s(H, M, h, m)$ , namely, when the number of werewolves that have been checked is greater than or equal to the number of villagers that have been checked plus the number of prophets, and when the number of werewolves that have been checked is smaller than the number of villagers that have been checked plus the number of prophets. The former case is relatively simple and can be directly transformed into the form of the game without a prophet. The latter case is a little more complex, in which there exist some checked villagers and no checked werewolves after some rounds processed by the action patterns adopted by both groups. Then we obtain the recursive formula:

$$s(H, M, h, m) = \begin{cases} \frac{M}{H-h+M} \left(\frac{1}{2}\right)^M, & \text{if } M = H + 1 \text{ and } m = 0, \\ \left(\frac{1}{2}\right)^M, & \text{if } M = H + 1 \text{ and } m \geq 1, \\ 1 - w(H + M + 2 - 2m, M - m), & \text{if } M < H + 1 \text{ and } m \geq h + 1, \\ u(H + 1 - m, M - m, h + 1 - m), & \text{if } M < H + 1 \text{ and } h + 1 > m \geq 1, \\ \frac{M}{H-h+M} u(H, M - 1, h) + \frac{H-h}{H-h+M} u(H - 1, M, h), & \text{if } M < H + 1 \text{ and } m = 0. \end{cases} \quad (36)$$

where

$$u(H', M', h') = \begin{cases} 1, & \text{if } M' = 0, \\ \frac{H'-h'}{H'+M'-h'} u(H' - 2, M', h' - 1) + \frac{M'}{H'+M'-h'} u(H' - 1, M' - 1, h' - 1), & \text{if } h' \geq 1, M' \geq 1, \text{ and } H' > M', \\ 1 - w(H' + M' + 1, M'), & \text{if } h' = 0, M' \geq 1, \text{ and } H' > M', \\ \frac{M'}{H'-h'+M'} \left(\frac{1}{2}\right)^{M'}, & \text{if } H' = M' \text{ and } M' \geq 1, \\ 0, & \text{if } H' < M' \text{ and } M' \geq 1. \end{cases} \quad (37)$$

The function  $u(H', M', h')$  characterizes the behavior of both groups after the prophet reveals information identifying  $h'$  villagers, with no werewolves exposed. Here,  $H'$  denotes the total number of villagers,  $M'$  the total number of werewolves, and  $h'$  the number of villagers whose group affiliations have been verified by the prophet.

This function reflects the citizen group's strategy to avoid voting out the confirmed villagers during the day, while the werewolf group, in turn, prioritizes eliminating these revealed villagers during the night.

Now, we give the recursive formula of  $R(Hiding, I_t)$ .

$$\begin{aligned}
 R(Hiding, I_t) = & \sum_i \alpha_i P_{17}(H_i, M_i, h_i, m_i)(1 - w(H_i + M_i, M_i)) \\
 & + \sum_i \alpha_i P_{18}(H_i, M_i, h_i, m_i)(1 - w(H_i + M_i, M_i)) \\
 & + \sum_i \alpha_i P_{19}(H_i, M_i, h_i, m_i)(1 - w(H_i + M_i, M_i - 1)) \\
 & + \sum_i \alpha_i P_{20}(H_i, M_i, h_i, m_i)(1 - w(H_i + M_i, M_i - 1)) \\
 & + \sum_i \alpha_i P_{21}(H_i, M_i, h_i, m_i)(1 - w(H_i + M_i, M_i)) \\
 & + \sum_{ii=1}^8 \left( \sum_i \alpha_i P_{ii}(H_i, M_i, h_i, m_i) R \left( g, \sum_i \frac{\alpha_i P_{ii}(H_i, M_i, h_i, m_i) node_{ii}(H_i, M_i, h_i, m_i)}{\sum_i \alpha_i P_{ii}(H_i, M_i, h_i, m_i)} \right) \right), \\
 & + \sum_{ii=9}^{12} \left( \sum_i \alpha_i \left( P_{ii}(H_i, M_i, h_i, m_i) + P_{ii+4}(H_i, M_i, h_i, m_i) \right) \cdot \right. \\
 & \quad R \left( g, \sum_i \frac{\alpha_i P_{ii}(H_i, M_i, h_i, m_i) \cdot node_{ii}(H_i, M_i, h_i, m_i)}{\sum_i \alpha_i (P_{ii}(H_i, M_i, h_i, m_i) + P_{ii+4}(H_i, M_i, h_i, m_i))} \right) \\
 & \quad \left. + \sum_i \frac{\alpha_i P_{ii+4}(H_i, M_i, h_i, m_i) \cdot node_{ii+4}(H_i, M_i, h_i, m_i)}{\sum_i \alpha_i (P_{ii}(H_i, M_i, h_i, m_i) + P_{ii+4}(H_i, M_i, h_i, m_i))} \right) \right). \tag{38}
 \end{aligned}$$

where

$P_{ii}$  denotes the probability of  $(H, M, h, m)$  turning into  $node_{ii}(H, M, h, m)$ .

$P_{17}$  denotes situation where prophet is voted out during the day.

$P_{18}$  denotes situation where checked villager is voted out during the day and prophet is eliminated during the night.

$P_{19}$  denotes situation where checked werewolf is voted out during the day and prophet is eliminated during the night.

$P_{20}$  denotes situation where unchecked werewolf is voted out during the day and prophet is eliminated during the night.

$P_{21}$  denotes situation where unchecked villager is voted out during the day and prophet is eliminated during the night.

$node_1(H, M, h, m)$  is induced by Checked villager voted out-Villager checked-Checked villager eliminated;

$node_2(H, M, h, m)$  is induced by Checked villager voted out-Villager checked-Unchecked villager eliminated;

$node_3(H, M, h, m)$  is induced by Checked villager voted out-Werewolf checked-Checked villager eliminated;

$node_4(H, M, h, m)$  is induced by Checked villager voted out-Werewolf checked-Unchecked villager eliminated;

$node_5(H, M, h, m)$  is induced by Checked werewolf voted out-Villager checked-Checked villager eliminated;

$node_6(H, M, h, m)$  is induced by Checked werewolf voted out-Villager checked-Unchecked villager eliminated;

$node_7(H, M, h, m)$  is induced by Checked werewolf voted out-Werewolf checked-Checked villager eliminated;

$node_8(H, M, h, m)$  is induced by Checked werewolf voted out-Werewolf checked-Unchecked villager eliminated;

$node_9(H, M, h, m)$  is induced by Unchecked werewolf voted out-Villager checked-Checked villager eliminated;

$node_{10}(H, M, h, m)$  is induced by Unchecked werewolf voted out-Villager checked-Unchecked villager eliminated;

$node_{11}(H, M, h, m)$  is induced by Unchecked werewolf voted out-Werewolf checked-Unchecked villager eliminated;

$node_{12}(H, M, h, m)$  is induced by Unchecked werewolf voted out-Werewolf checked-Checked villager eliminated;

$node_{13}(H, M, h, m)$  is induced by Unchecked villager voted out-Villager checked-Checked villager eliminated;

$node_{14}(H, M, h, m)$  is induced by Unchecked villager voted out-Villager checked-Unchecked villager eliminated;

$node_{15}(H, M, h, m)$  is induced by Unchecked villager voted out-Werewolf checked-Unchecked villager eliminated;

$node_{16}(H, M, h, m)$  is induced by Unchecked villager voted out-Werewolf checked-Checked villager eliminated.

The specific expressions of the variable are as follows:

$$P_1(H_i, M_i, h_i, m_i) = \frac{h_i}{H_i + M_i + 1} \cdot \frac{(H_i - 1) - (h_i - 1)}{(H_i - 1) + M_i - (h_i - 1) - m_i} \cdot \frac{h_i}{H_i}, \quad (39)$$

$$P_2(H_i, M_i, h_i, m_i) = \frac{h_i}{H_i + M_i + 1} \cdot \frac{(H_i - 1) - (h_i - 1)}{(H_i - 1) + M_i - (h_i - 1) - m_i} \cdot \frac{H_i - 1 - h_i}{H_i}, \quad (40)$$

$$P_3(H_i, M_i, h_i, m_i) = \frac{h_i}{H_i + M_i + 1} \cdot \frac{M_i - m_i}{(H_i - 1) + M_i - (h_i - 1) - m_i} \cdot \frac{h_i - 1}{H_i}, \quad (41)$$

$$P_4(H_i, M_i, h_i, m_i) = \frac{h_i}{H_i + M_i + 1} \cdot \frac{M_i - m_i}{(H_i - 1) + M_i - (h_i - 1) - m_i} \cdot \frac{(H_i - 1) - (h_i - 1)}{H_i}, \quad (42)$$

$$P_5(H_i, M_i, h_i, m_i) = \frac{m_i}{H_i + M_i + 1} \cdot \frac{H_i - h_i}{H_i + (M_i - 1) - h_i - (m_i - 1)} \cdot \frac{h_i + 1}{H_i + 1}, \quad (43)$$

$$P_6(H_i, M_i, h_i, m_i) = \frac{m_i}{H_i + M_i + 1} \cdot \frac{H_i - h_i}{H_i + (M_i - 1) - h_i - (m_i - 1)} \cdot \frac{H_i - h_i}{H_i + 1}, \quad (44)$$

$$P_7(H_i, M_i, h_i, m_i) = \frac{m_i}{H_i + M_i + 1} \cdot \frac{(M_i - 1) - (m_i - 1)}{H_i + (M_i - 1) - h_i - (m_i - 1)} \cdot \frac{h_i}{H_i + 1}, \quad (45)$$

$$P_8(H_i, M_i, h_i, m_i) = \frac{m_i}{H_i + M_i + 1} \cdot \frac{(M_i - 1) - (m_i - 1)}{H_i + (M_i - 1) - h_i - (m_i - 1)} \cdot \frac{H_i - h_i}{H_i + 1}, \quad (46)$$

$$P_9(H_i, M_i, h_i, m_i) = \frac{M_i - m_i}{H_i + M_i + 1} \cdot \frac{H_i - h_i}{H_i + (M_i - 1) - h_i - m_i} \cdot \frac{h_i + 1}{H_i + 1}, \quad (47)$$

$$P_{10}(H_i, M_i, h_i, m_i) = \frac{M_i - m_i}{H_i + M_i + 1} \cdot \frac{H_i - h_i}{H_i + (M_i - 1) - h_i - m_i} \cdot \frac{H_i - (h_i + 1)}{H_i + 1}, \quad (48)$$

$$P_{11}(H_i, M_i, h_i, m_i) = \frac{M_i - m_i}{H_i + M_i + 1} \cdot \frac{(M_i - 1) - m_i}{H_i + (M_i - 1) - h_i - m_i} \cdot \frac{H_i - h_i}{H_i + 1}, \quad (49)$$

$$P_{12}(H_i, M_i, h_i, m_i) = \frac{M_i - m_i}{H_i + M_i + 1} \cdot \frac{(M_i - 1) - m_i}{H_i + (M_i - 1) - h_i - m_i} \cdot \frac{h_i}{H_i + 1}, \quad (50)$$

$$P_{13}(H_i, M_i, h_i, m_i) = \frac{H_i - h_i}{H_i + M_i + 1} \cdot \frac{(H_i - 1) - h_i}{(H_i - 1) + M_i - h_i - m_i} \cdot \frac{h_i + 1}{H_i}, \quad (51)$$

$$P_{14}(H_i, M_i, h_i, m_i) = \frac{H_i - h_i}{H_i + M_i + 1} \cdot \frac{(H_i - 1) - h_i}{(H_i - 1) + M_i - h_i - m_i} \cdot \frac{(H_i - 1) - (h_i + 1)}{H_i}, \quad (52)$$

$$P_{15}(H_i, M_i, h_i, m_i) = \frac{H_i - h_i}{H_i + M_i + 1} \cdot \frac{M_i - m_i}{(H_i - 1) + M_i - h_i - m_i} \cdot \frac{(H_i - 1) - h_i}{H_i}, \quad (53)$$

$$P_{16}(H_i, M_i, h_i, m_i) = \frac{H_i - h_i}{H_i + M_i + 1} \cdot \frac{M_i - m_i}{(H_i - 1) + M_i - h_i - m_i} \cdot \frac{h_i}{H_i}, \quad (54)$$

$$P_{17}(H_i, M_i, h_i, m_i) = \frac{1}{H_i + M_i + 1}, \quad (55)$$

$$P_{18}(H_i, M_i, h_i, m_i) = \frac{h_i}{H_i + M_i + 1} \cdot \frac{1}{H_i - 1 + 1}, \quad (56)$$

$$P_{19}(H_i, M_i, h_i, m_i) = \frac{m_i}{H_i + M_i + 1} \cdot \frac{1}{H_i + 1}, \quad (57)$$

$$P_{20}(H_i, M_i, h_i, m_i) = \frac{M_i - m_i}{H_i + M_i + 1} \cdot \frac{1}{H_i + 1}, \quad (58)$$

$$P_{21}(H_i, M_i, h_i, m_i) = \frac{H_i - h_i}{H_i + M_i + 1} \cdot \frac{1}{H_i - 1 + 1}, \quad (59)$$

$$node_1(H_i, M_i, h_i, m_i) = (H_i - 2, M_i, h_i - 1, m_i), \quad (60)$$

$$node_2(H_i, M_i, h_i, m_i) = (H_i - 2, M_i, h_i, m_i), \quad (61)$$

$$node_3(H_i, M_i, h_i, m_i) = (H_i - 2, M_i, h_i - 2, m_i + 1), \quad (62)$$

$$node_4(H_i, M_i, h_i, m_i) = (H_i - 2, M_i, h_i - 1, m_i + 1), \quad (63)$$

$$node_5(H_i, M_i, h_i, m_i) = (H_i - 1, M_i - 1, h_i, m_i - 1), \quad (64)$$

$$node_6(H_i, M_i, h_i, m_i) = (H_i - 1, M_i - 1, h_i + 1, m_i - 1), \quad (65)$$

$$node_7(H_i, M_i, h_i, m_i) = (H_i - 1, M_i - 1, h_i - 1, m_i), \quad (66)$$

$$node_8(H_i, M_i, h_i, m_i) = (H_i - 1, M_i - 1, h_i, m_i), \quad (67)$$

$$node_9(H_i, M_i, h_i, m_i) = (H_i - 1, M_i - 1, h_i, m_i), \quad (68)$$

$$node_{10}(H_i, M_i, h_i, m_i) = (H_i - 1, M_i - 1, h_i + 1, m_i), \quad (69)$$

$$node_{11}(H_i, M_i, h_i, m_i) = (H_i - 1, M_i - 1, h_i, m_i + 1), \quad (70)$$

$$node_{12}(H_i, M_i, h_i, m_i) = (H_i - 1, M_i - 1, h_i - 1, m_i + 1), \quad (71)$$

$$node_{13}(H_i, M_i, h_i, m_i) = (H_i - 2, M_i, h_i, m_i), \quad (72)$$

$$node_{14}(H_i, M_i, h_i, m_i) = (H_i - 2, M_i, h_i + 1, m_i), \quad (73)$$

$$node_{15}(H_i, M_i, h_i, m_i) = (H_i - 2, M_i, h_i, m_i + 1), \quad (74)$$

$$node_{16}(H_i, M_i, h_i, m_i) = (H_i - 2, M_i, h_i - 1, m_i + 1), \quad (75)$$

$$node_{17}(H_i, M_i, h_i, m_i) = (H_i - 2, M_i, h_i, m_i), \quad (76)$$

$$node_{18}(H_i, M_i, h_i, m_i) = (H_i, M_i, h_i, m_i + 1). \quad (77)$$

Now we introduce the transformation criteria for information sets. If the current information set  $I_t$  satisfies any of the following conditions, it should first be transformed into an equivalent form, as defined below, before being substituted into (39).

If no valid nodes remain within the information set after applying this transformation, the prophet's decision-making process is considered complete. The game effectively terminates for the prophet.

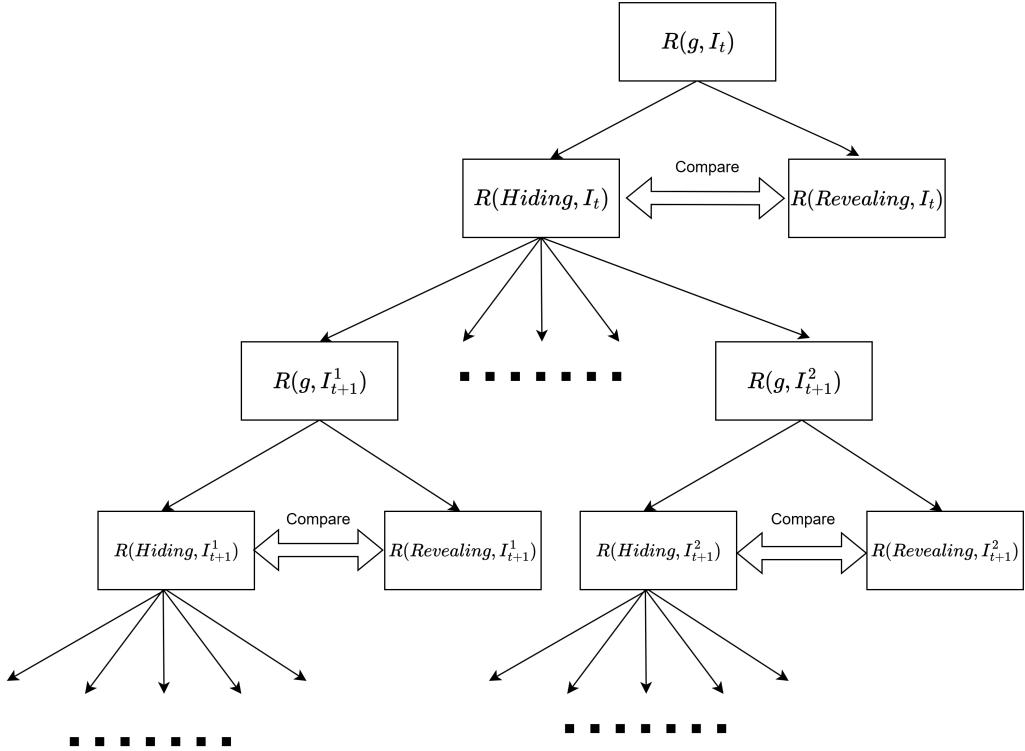
- (1) There exists some node in the information set  $I_t$  such that the werewolf group wins directly. For the information set  $I_t = \sum_i \alpha_i \cdot (H_i, M_i, h_i, m_i)$ , if there exists any  $i$  such that  $H_i + 1 < M_i$ , let  $J$  be the set of such indices. Then we rewrite:

$$R(Hiding, I_t) = \left( \sum_{i \notin J} \alpha_i \right) R \left( Hiding, \sum_{i \notin J} \frac{\alpha_i \cdot (H_i, M_i, h_i, m_i)}{\sum_{i \notin J} \alpha_i} \right). \quad (78)$$

Naturally, if the prophet instead chooses the *Revealing* strategy, the werewolf group will still win in these nodes, which is already captured in  $R(Revealing, I_t)$ .

- (2) There exists some node in the information set  $I_t$  such that the citizen group wins directly. For the information set  $I_t = \sum_i \alpha_i \cdot (H_i, M_i, h_i, m_i)$ , if there exists any  $i$  such that  $M_i = 0$ , let  $K$  be the set of such indices. Then we rewrite:

$$R(Hiding, I_t) = \sum_{i \in K} \alpha_i + \left( \sum_{i \notin K} \alpha_i \right) R \left( Hiding, \sum_{i \notin K} \frac{\alpha_i \cdot (H_i, M_i, h_i, m_i)}{\sum_{i \notin K} \alpha_i} \right). \quad (79)$$

Fig. 9. Flowchart of calculating  $R(Hiding, I_t)$ .

(3) There exists some node in the information set  $I_t$  such that the werewolf group would directly employ the “all-in strategy”.

For the information set  $I_t = \sum_i \alpha_i \cdot (H_i, M_i, h_i, m_i)$ , if there exists any  $i$  such that  $H_i + 1 = M_i$ , let  $L$  be the set of such indices. Then we rewrite:

$$R(Hiding, I_t) = \sum_{i \in L} \alpha_i \left(\frac{1}{2}\right)^{M_i+1} + \sum_{i \notin L} \alpha_i R\left(Hiding, \sum_{i \notin L} \frac{\alpha_i \cdot (H_i, M_i, h_i, m_i)}{\sum_{i \notin L} \alpha_i}\right). \quad (80)$$

Combining the recursive formulation of  $R(Hiding, I_t)$  with the transformation criteria discussed above, we observe that each information set  $I_t$  can generate at most twelve successor information sets at time  $t + 1$ . Denote these potential information sets by  $\{I_{t+1}^r\}_r$ . Then, (39) can be written as

$$R(Hiding, I_t) = e(I_t) + \sum_r \gamma_r R(g, I_{t+1}^r), \quad (39)$$

where  $e(I_t)$  denotes a constant term that differs for each information set  $I_t$ .

The value  $R(g, I_t)$  is defined as the maximum between  $R(Revealing, I_t)$  and  $R(Hiding, I_t)$ . As shown in (39), the value of  $R(Hiding, I_t)$  is a linear combination of all  $R(g, I_{t+1}^r)$  for  $I_{t+1}^r \in \{I_{t+1}^r\}_r$ . For each  $R(g, I_{t+1}^r)$ , we apply the same process recursively, just as we did for  $R(g, I_t)$ .

Through dynamic programming, we can calculate the prophet’s optimal strategy given any information set in the game. Moreover, the entire process can be regarded as a **Markov Decision**

**Process (MDP)**, meaning that no matter how the prophet arrives at a certain information set, as long as the information set is the same, the optimal action should also be the same. We can even derive optimal actions in certain information sets that cannot be reached through regular gameplay, such as those involving irrational node probabilities. The entire proc

### Proof of Claim 2

PROOF. First, consider the scenario where the werewolf group might self-kill during the night, which can be divided into two cases:

- The prophet has revealed the checking results;
- The prophet has not yet revealed the checking results.

The case where the prophet has revealed their information is straightforward to analyze. Once the prophet has revealed the information set  $I_t = \sum_i \alpha_i(H_i, M_i, h_i, m_i)$ , the winning probability of the citizen group is  $\sum_i \alpha_i s(H_i, M_i, h_i, m_i)$ . For any  $s(H_i, M_i, h_i, m_i)$ , it can be regarded as a variation of the game without prophets. Therefore, any strategy involving self-killing for the werewolf group is dominated in this case.

The case where the prophet has not yet revealed their information can be further divided into two subcases:

- The prophet has been eliminated either by werewolf elimination or by being voted out;
- The prophet is still present.

If the prophet has been eliminated, the game reduces to one without a prophet. In this situation, the werewolf group self-killing during the night is a strictly dominated strategy, which has been proven in Section 3.

If the prophet is still present, the situation becomes more complex, and we provide a detailed proof below.

Given the complexity of the recursive model introduced earlier, along with its multiple termination conditions, directly analyzing the potential payoffs for the prophet across various information sets is challenging. To address this, we adopt a *Dynamic Adjustment Process* approach, similar to that used in the Cournot duopoly model, where both parties continuously adapt to each other's strategies, iteratively adjusting their actions until convergence to an equilibrium.

Suppose the information set of the prophet after the checking and before the werewolf elimination is  $\sum_{i=1}^n \beta_i \cdot (H_i, M_i, h, m)$ . Assume that in the prophet's belief, the probability of the werewolf group adopting a self-killing strategy at node  $(H_i, M_i, h, m)$  is  $\hat{q}_i$ . Let the vector  $\hat{Q} = [\hat{q}_1, \dots, \hat{q}_n]$  represent these independent probabilities. If the actual probability vector of the werewolf group's self-killing actions equals  $\hat{Q}$ , we denote the expected winning probability of the citizen group under strategy profile  $\hat{Q}$  as  $AR(\hat{Q})$ . Formally, this can be expressed as

$$\begin{aligned} AR(\hat{Q}) = & \sum_i \beta_i \hat{q}_i \frac{m}{M_i} R \left( g, \sum_i \frac{\beta_i \hat{q}_i \frac{m}{M_i} (H_i, M_i - 1, h, m - 1)}{\sum_i \beta_i \hat{q}_i \frac{m}{M_i}} \right) \\ & + \sum_i \beta_i \left( \hat{q}_i \frac{M_i - m}{M_i} + (1 - \hat{q}_i) \frac{H_i - h}{H_i + 1} \right) \\ & \cdot R \left( g, \sum_i \frac{\beta_i \left( \hat{q}_i \frac{M_i - m}{M_i} (H_i, M_i - 1, h, m) + (1 - \hat{q}_i) \frac{H_i - h}{H_i + 1} (H_i - 1, M_i, h, m) \right)}{\sum_i \beta_i \left( \hat{q}_i \frac{M_i - m}{M_i} + (1 - \hat{q}_i) \frac{H_i - h}{H_i + 1} \right)} \right) \end{aligned}$$

$$\begin{aligned}
& + \sum_i \beta_i (1 - \hat{q}_i) \frac{1}{H_i + 1} (1 - w(H_i + M_i + 1, M_i)) \\
& + \sum_i \beta_i (1 - \hat{q}_i) \frac{h}{H_i + 1} R \left( g, \sum_i \frac{\beta_i (1 - \hat{q}_i) \frac{h}{H_i + 1} (H_i - 1, M_i, h - 1, m)}{\sum_i \beta_i (1 - \hat{q}_i) \frac{h}{H_i + 1}} \right). \tag{81}
\end{aligned}$$

It is straightforward to see that  $AR(\hat{Q})$  is monotonically increasing with respect to each component of  $\hat{Q}$ . This implies that when the probability distribution of the werewolf group's actions is fully known to the prophet, self-killing strategies cannot mislead the prophet's judgment. Instead, such strategies only decrease the werewolf group's winning probability.

Suppose the actual probability of the werewolf group self-killing is  $Q = [q_1, \dots, q_n]$  and the actual expectation of the citizen group's winning probability, when the prophet takes the optimal strategy given belief  $\hat{Q}$ , is  $DR(Q, \hat{Q})$ . When  $Q \leq \hat{Q}$ , we have

$$DR(Q, \hat{Q}) \leq AR(Q) \leq AR(\hat{Q}), \tag{82}$$

with equality holding if and only if  $Q = \hat{Q}$ .

From the perspective of the werewolf group, given any  $\hat{Q}$  from the prophet, they could choose  $Q < \hat{Q}$  to reduce the citizen group's winning probability. Similarly, from the prophet's perspective, given any  $Q$  from the werewolf group, the prophet can ensure  $\hat{Q} = Q$ . Thus, we conclude  $Q = \hat{Q} = \vec{0}$ , meaning that in any case, the probability of the werewolf group adopting a self-killing strategy is zero.

**Proof complete.**

Received 27 November 2024; revised 25 July 2025; accepted 11 September 2025