# A toy model for voting in the Resistance

Charles Zheng

October 31, 2015

## 1 Introduction

Resistance is a bluffing game where players have hidden roles. There are two teams: resistance, and spies. Players are secretly assigned to one of the two teams, with the majority of the players being assigned to the resistance, and are privately revealed their assignment. It is a zero-sum game between the two teams. The mechanics of the game are complicated and somewhat arbitrary, so we give an abstracted and oversimplified description. The object of the game for the resistance is for the every member of the resistance to correctly guess the identities of the spies at the end of the game. Since various events in the game leak information about the identities of the spies, the goal of the spies is to end the game before the resistance players can discover all of their identities. Each player takes actions at various points in the game which either advance the game towards ending (and therefore spy victory), or which reveal information about the players' identities, or both. Some actions are publically observed, while others are concealed to some degree. The spies have private information which allows them to know how much any particular action advances the game towards ending, but a resistance player only has private and partial information about how much any particular action made by themselves or another player advances the game towards ending.

If I am a resistance player, and I observe you to take many actions which appear to be advancing the game towards ending, then I will start to suspect you to be on the spy team. If the spies are too aggressive in taking such actions, then the information held by the resistance will accumulate, and the probability increases that the resistance will be able to correctly identify the spies. Furthermore, if the game progresses too long, spies are eventually forced to take increasingly suspicious actions, so there is no way a spy can avoid drawing suspicion to themselves.

The full dynamics of the game involve an interplay of strategy, teamwork, interrogation and complex conceits for players of both teams, and one where the reputation of the individual players cannot be disentangled from the individual games. It is an extremely human game which cannot be easily studied using simplified models. On the other hand, even without the human elements, the strategic aspects of the game remain nontrivial. The question we intend to investigate here is: if we exclude the human factor, how much information can be gained from the actions of the players under optimal play? To answer this question we introduce a toy model of the game and study its properties. The toy model has the minimal level of complexity to capture the fact that players take actions which either benefit or hurt the resistance team, and which also reveal information about their hidden roles.

## 2 Noisy votes model

The game includes one special player, a judge $J$, and $N$ other players, called *voters*. $N_G$ of the voters are on the Good team and $N_B$ of the voters are on the Bad team. The voters are randomly assigned, so that each individual voter has an $N_G/N$ probability of being on the Good team. Let $G_1, \ldots, G_{N_G}$ denote the Good voters and $B_1, \ldots, B_{N_B}$ denote the Bad voters.

The game consists of a voting round and a judging round. In the voting round, the voters simultaneously and private choose a vote $V \in \{-1, 1\}$. Each Good voter $G_i$ independently chooses action $V_{G_i} = 1$, or 'upvote', with probability $1 - \epsilon$ and action $V_{G_i} = 0$, or 'downvote' with probability $\epsilon$. Each Bad voter $B_j$ independently chooses to upvote, $V_{B_j} = 1$ with probability $1 - p$, and downvote, $V_{B_j} = 0$, with probability $p$. The "noisy vote" probability $\epsilon$ is fixed; however, the Bad voters can collectively choose $p \in [0, 1]$ before each voting round. Let $S$ be the sum of the upvotes,

$$S = V_{G_1} + \cdots + V_{G_{N_G}} + V_{B_1} + \cdots + V_{B_{N_G}}$$

In the judging round, the judge $J$ guesses the identities of the Bad voters. The judge is required to pick $N_B$ players to identify as Bad players; let $C$ be the number of those players who are correctly identified.

The judge and the Good voters are on the same team, and their payoff $P(S, C)$ is a function of the number of upvotes $S$ and the number of Bad voters correctly identified by the judge, $C$. The game is zero-sum, so the payoff for the Bad votes is $-P(S, C)$.

The form of the payoff function $P(S, C)$ and the parameters $N_B$ and $\epsilon$ determine the set of optimal choices of downvoting probability $p$ for the Bad players. A particular optimal choice of $p$ determines an equilibrium strategy for the game. We categorize the possible equilibria as follows:

1. A unique equilibrium with $p = 0$
2. A unique equilibrium with $p = 1$
3. A unique equilibrium with $p \in (0, 1)$
4. Every $p \in [0, 1]$ results in the same expected payoff, hence there is no unique eqilibrium.