

# Supplementary Information for Can editors protect peer review from bad reviewers?

Rafael D'Andrea<sup>1,\*</sup> & James P O'Dwyer<sup>1</sup>

<sup>1</sup>Dept. Plant Biology, University of Illinois at Urbana-Champaign, IL, USA

\*rdandrea@illinois.edu

## Blacklisting selfish reviewers

Editors will blacklist any referee whose reviewing record indicates that they are likely to be selfish. Impartial referees never disagree on their reviews, so a disagreement between two reviewers implies that at least one of them is selfish. A referee with a high record of disagreements is therefore likely to be selfish. Suppose editors blacklist referees whose probability of being selfish is equal or greater than  $p_0$ . We can calculate  $p_0$  in terms of a referee's record of reviews and disagreements:

$$p_0 = P(S|k; n) = \frac{P(k|S; n)P(S)}{P(k|S; n) + P(k|D; n)} = \frac{f_s}{1 + \frac{P(k|D; n)}{P(k|S; n)}} \quad (1)$$

where  $P(S|k; n)$  is the probability that a referee who has reviewed  $n$  papers and has a record of  $k$  disagreements is selfish,  $P(k|S; n)$  is the probability that a selfish referee who has reviewed  $n$  papers will have a record of  $k$  disagreements (similarly for an impartial referee and  $P(k|D; n)$ ), and  $P(S)$  is the prior probability that the referee is selfish, which is simply the frequency of selfish referees in the pool,  $f_s$ .

The number of disagreements given the number of reviews is binomially distributed:

$$P(k|S; n) = \binom{n}{k} q_s^k (1 - q_s)^{n-k} \quad (2)$$

$$P(k|D; n) = \binom{n}{k} q_d^k (1 - q_d)^{n-k} \quad (3)$$

where  $q_s$  ( $q_d$ ) is the probability that a selfish (impartial) referee will disagree with the other referee in a given review. Those in turn depend on the probability of disagreement between two selfish referees,  $q_{ss}$ , and the probability of disagreement between a selfish and an impartial referee,  $q_{sd}$  (two impartial referees never disagree).

$$q_s = f_s q_{ss} + f_d q_{sd} \quad (4)$$

$$q_d = f_s q_{sd} \quad (5)$$

where the terms are weighted by the probability that the other referee is selfish or impartial.

Two selfish referees will only disagree on a paper if the quality of that paper falls between the upper limits of their acceptance windows:

$$q_{ss} = \int_{Q_{\min}}^{\infty} \left( P_A(b_1) \int_{b_1}^{\infty} \left( P_A(b_2) \int_{b_1}^{b_2} P_P(x) dx \right) db_2 \right) db_1 \quad (6)$$

where  $Q_{\min}$  is the lower limit of the acceptance window of all selfish referees,  $P_P(x)dx$  is the probability that the quality of a given submitted paper will fall between  $x$  and  $x + dx$ , and  $P_A(b)db$  is the probability that the quality of a given referee as an author will be between  $b$  and  $b + db$ . Equation 6 expresses the probability that the quality of the submitted paper falls between the upper limits of the acceptance windows of each selfish referee, integrated over all possible combinations of these upper limits (referee 2 is assumed, without loss of generality, to have the higher upper limit).

In contrast, a selfish referee and an impartial referee will agree if the paper falls either below the lower limit of acceptance of the selfish referee (in which case both will reject) or above the minimum imposed by the impartial referee but still below the upper limit imposed by the selfish referee (in which case both will accept). Otherwise they will disagree:

$$q_{sd} = 1 - \left( \int_{-\infty}^{Q_{\min}} P_P(x)dx + \int_{\bar{Q}_a}^{\infty} \left( P_A(b) \int_{\bar{Q}_a}^b P_P(x)dx \right) db \right) \quad (7)$$

where  $\bar{Q}_a$ , the average quality of accepted papers in stationary equilibrium, is the minimum quality accepted by impartial referees. This quantity is a function of the proportion of selfish referees in the pool, and we calculate it numerically. (Strictly speaking,  $\bar{Q}_a(f_s, t)$  varies in time and is affected by blacklisting. However,  $q_{sd}$  is reasonably insensitive to the exact value of  $\bar{Q}_a$ , so for simplicity we use the stationary value  $\bar{Q}_a(f_s, \infty)$  observed under no blacklisting.) As described in the main text, the distribution of authors is  $N(100, \sigma_{\text{author}}^2)$ , and the distribution of submitted papers is  $N(100, \sigma_{\text{author}}^2 + \sigma_{\text{quality}}^2)$ :

$$P_P(x) = \frac{1}{\sqrt{2(\sigma_{\text{author}}^2 + \sigma_{\text{quality}}^2)}\pi} \exp\left(-\frac{(x - 100)^2}{2(\sigma_{\text{author}}^2 + \sigma_{\text{quality}}^2)}\right) \quad (8)$$

$$P_A(b) = \frac{1}{\sqrt{2\sigma_{\text{author}}^2}\pi} \exp\left(-\frac{(b - 100)^2}{2\sigma_{\text{author}}^2}\right) \quad (9)$$

Supplementary Figure 1 shows the minimum number of disagreements in a referee's record required for blacklisting as a function of the number of papers the referee has reviewed, obtained by numerically evaluating Equation 1. Notice that the number of disagreements needed for blacklisting is roughly proportional to the number of reviews, and is typically close to half of the number of reviews whether selfish referees are as rare as 10% or as common as 40% of the referee pool. We use a cutoff  $p_0 = 0.9$ , but results are insensitive to that choice; in fact,  $p_0$  can be as low as 0.5 without changes to the figure. This is because  $P(S|k; n)$  rises abruptly with  $k$ , from close to 0 below the threshold to close to 1 above it.

**Figure 1:** Minimum number of disagreements on a referee's reviewing record given the number of papers reviewed that indicate higher than 90% probability that the referee is selfish. Referees remain in the pool for 35 years, or 70 reviewing cycles, and then are replaced by new referees.

