Strategic Evidence Disclosure in Networks and Equilibrium Discrimination

Leonie Baumann* and Rohan Dutta‡

May 2, 2023

Abstract

A group of agents with ex-ante independent and identically uncertain quality compete for a prize, awarded by a principal. Agents may possess evidence about the quality of those they share a social connection with (*neighbours*), and themselves. In one equilibrium, adversarial disclosure of evidence leads the principal to statistically discriminate between agents based on their number of neighbours (*degree*). We identify parameter values for which an agent's ex-ante winning probability is monotone in degree. All equilibria that satisfy some robustness criteria lie between this adversarial disclosure equilibrium and a less informative one that features no snitching and no discrimination.

*McGill University, Department of Economics, leonie.baumann@mcgill.ca. Corresponding author.

†McGill University, Department of Economics, rohan.dutta@mcgill.ca.

‡We gratefully acknowledge support under the FRQSC grant 2023-NP-310915 and the SSHRC grant 430-2022-00601. We thank Francis Bloch, Gabriel Carroll, Rahul Deb, Federico Echenique, Sanjeev Goyal, Matthew Jackson, Frederic Koessler, Rachel Kranton, Rohit Lamba, David Levine, Elliot Lipnowski, Mihai Manea, Paula Onuchic, Ariel Rubinstein and seminar participants at the Barcelona GSE Summer Forum (2022) and BiNoMa workshop (2022) for helpful comments. Special thanks to Yingni Guo for a discussion that inspired this paper.

1 Introduction

In a group of agents with indistinguishable productive ability, economic rewards often flow to those with more social connections. A common explanation for such outcomes, especially those that make news, is taste-based discrimination such as cronyism or nepotism. Social network theory highlights other rationales in the labour market context. In Calvo-Armengol and Jackson (2004), connections are valuable sources of information about job vacancies. In Montgomery (1991), connections permit hiring via referrals.

In this paper, we present a novel equilibrium process through which benefits get distributed unequally as a function of an agent's number of connections (*degree*), in an otherwise homogeneous group of agents. A principal crowdsources information from and about a group of ex-ante identical agents, in order to assign a prize. The agents' strategic revelation of information, in equilibrium, leads the wholly unbiased and rational principal to reward the agents unequally based on their degree. Key to this finding are the different meanings rationally inferred by the principal from the lack of information about two agents who differ in degree alone.

Our model involves a group of agents, each of whom could be good or bad (*own-type*). An agent's own-type is his private information and all agents are ex-ante identical, in that they share the same, commonly known, i.i.d. probability, γ, of being good. A principal wishes to award a prize to a good agent only, but every agent wants the prize. Examples of such an environment include a supervisor deciding which employee to promote, or a political party selecting a member to run in an election.

The agents have social or professional connections among themselves. These are described by a commonly known network, which limits the information

agents may come to possess. In addition to knowing their own own-type, agents may obtain hard information about themselves or any agent they are connected to (*neighbour*). So, in particular, an agent can possess fully verifiable evidence about his own own-type and those of his neighbours. The former may be in the form of client reviews of past work. Evidence about neighbours can take the form of emails, shared work, documentation of performance in a joint project, or photos. A key feature is that if the principal observes evidence about an agent, she is certain of the latter's own-type. An agent obtains evidence about the own-type of any given neighbour of his, or himself, with i.i.d. probability q, and this is his private information. This stochastic evidence structure is essentially the same as in Dye (1985), and more recently, Hart, Kremer and Perry (2017) and Ben-Porath, Dekel and Lipman (2019).

Following the arrival (or not) of evidence, the agents simultaneously decide which evidence, if any, to reveal to the principal. The principal then uses this revealed information to determine the prize winner. We study equilibrium behaviour of this game in which an agent's prize winning chances is determined solely by the principal's equilibrium belief about his own-type. So, for instance, if the principal knew for sure that two different agents were both good, then they would receive the prize with equal probability.[1]

We identify two distinct equilibria (Proposition 1). In one, the agents behave adversarially, revealing good evidence about themselves, bad evidence about their neighbours and suppressing the rest. In the other equilibrium, agents only reveal good evidence about themselves. In the latter, the principal interprets bad evidence from an agent about his neighbour as proof that the agent himself is bad. In this "no-snitching" equilibrium, the ex-ante probability of receiving the prize is the same for all agents and independent of their network position (Proposition 2). This is not true in the adversarial disclosure equilibrium. We show that irrespective of the network, in the adversarial disclosure equilibrium agents with more connections receive the prize with higher ex-ante probability if the probability of being good is sufficiently high and/or

[1]In particular, the discrimination we identify in our model is not due to the use of a discriminatory tie-breaking rule by the principal.

the probability of obtaining evidence is sufficiently low. This result does not rely on extreme parameter values. For instance, an easy to parse sufficient condition under which more connections lead to higher ex-ante reward probability is

$$\frac{\gamma}{1-\gamma}(1-q) > 1.$$

We also show that for any given level of γ, agents with fewer connections do better ex-ante for sufficiently large q. In summary, the adversarial disclosure equilibrium leads to discrimination based on degree (Proposition 3).

The direction of discrimination in the adversarial disclosure equilibrium is decided by the interplay of two key forces. In equilibrium, when the principal receives no evidence about two agents with different degrees, she rationally infers that the one with the higher degree has a higher probability of being good. This is simply a case of statistical discrimination that arises endogenously in equilibrium due to the specific disclosure strategies used by the agents.[2] This force, therefore, favours agents with higher degrees. Conditional on being bad, however, a higher degree increases the chances of being snitched upon. This force favours agents with lower degrees. Which of these forces wins out at the ex-ante stage depends on the two parameters of the model, and is the content of our main result (Proposition 3).

The richness of the strategy space compared to the number of possible outcomes makes agents indifferent across a number of strategy profiles. This can be exploited to generate a multiplicity of equilibria that rely on fragile constructions. To avoid artificial equilibria we focus on sequential equilibria that satisfy a robustness property introduced in Ben-Porath, Dekel and Lipman (2019), and survive a further perturbation. In the latter, the agents believe, with vanishing probability, that the principal follows a coarse strategy. The "no-snitching" and adversarial disclosure equilibria, both satisfy these robustness criteria (Proposition 5). Importantly, any sequential equilibrium that satisfies these properties lies between these two equilibria, in that an agent's strategy never reveals less evidence than in the "no-snitching" equilibrium

[2]Therefore our results fall under the heading of equilibrium discrimination as opposed to statistical discrimination.

and never reveals more evidence than in the adversarial disclosure equilibrium (Proposition 4 and 5).

The central finding of our paper contributes to the economics literature on discrimination. Most of the latter builds on the theory of taste-based discrimination, following Becker (1971) or statistical discrimination, following Phelps (1972), in the context of race and the labour market. In taste-based discrimination, interaction across different groups (races) generates disutility. In statistical discrimination, employers obtain a noisy signal of an agent's productivity, which is correlated with the agent's race, an observable attribute. By contrast, in our model, degree is independent of own-type and therefore payoff irrelevant. It is only in equilibrium that the distribution of evidence about an agent anticipated by the principal ends up as a function of degree. For this reason, our paper belongs in the equilibrium discrimination literature that follows the lead of Arrow (1973). Typically in this literature the observable payoff-irrelevant dimension along which discrimination occurs is skin colour or race.[3] To the best of our knowledge, ours is the first paper wherein discrimination occurs on the basis of degree. We position our work in this literature in greater detail in section 6.

This paper also contributes to the literature that analyzes the strategic disclosure of evidence to an uninformed principal. The vast majority of this literature examines settings where individuals can disclose evidence about own but not about competitors' characteristics. In the seminal papers of Grossman and Hart (1980), Grossman (1981), and Milgrom (1981), an unravelling result obtains. Agents voluntarily disclose all relevant evidence in equilibrium. Dye (1985) and Jung and Kwon (1988) consider a framework of stochastic evidence in which the agent obtains, with some probability, fully revealing evidence, and no evidence at all with the remaining probability. The unravelling result no longer holds in this setting. In particular, the agents only disclose favourable evidence. Recent work such as Kremer, Hart and Perry (2017), Ben-Porath, Dekel and Lipman (2018, 2019) examine the import of this *Dye-evidence* in a number of settings. A major concern in these studies is whether and when

[3]See, for instance, Eeckhout (2006) and Peski and Szentes (2013).

the principal's ability to commit is redundant. The disclosure game we study also features Dye-evidence, but we are interested in the possibility and nature of equilibrium discrimination, and to the best of our knowledge, the first to do so. The novel feature in our setting is the ability of agents to disclose evidence about others, which is constrained by their network position. This structure is at the heart of the strategic considerations that lead to equilibrium discrimination in our study.

The literature wherein individuals cannot only disclose evidence about own but also others' characteristics is sparse. Baumann (2018) is the first to analyse a setting in which individuals disclose partially conclusive evidence about themselves and their acquaintances. Ben-Porath, Dekel and Lipman (2018) consider a competition between an incumbent and a challenger where each might have evidence about the characteristics of the incumbent.

Recently, the literature on social and economic networks has started examining the design of peer evaluation mechanisms in which agents send information about each other. Baumann (2018) proposes a robust peer evaluation mechanism to identify the highest quality agent when agents disclose partially verifiable information. Bloch and Olckers (2021; 2022) investigate the design of an incentive-compatible mechanism to extract the entire quality ranking of agents in a network. In their framework, agents are fully informed about themselves and their neighbours but do not possess evidence, in that their messages are non-verifiable. Bloch and Olckers (2022) provide necessary and sufficient conditions for such a mechanism to exist.

The structure of the paper is as follows. We introduce the model in section 2. In section 3, we describe and prove the existence of the adversarial disclosure and no-snitching equilibrium. Section 4 contains the key analysis of the ex-ante award probabilities of agents in these two equilibria. The characterization of sequential equilibria that survive certain robustness criteria is in section 5. We discuss some implications of our findings in section 6.

2 Model

There is a principal and a set of agents $N = \{1, ..., n\}$ with $n \geq 2$. An agent is either good (G) or bad (B), and this is his private information. Formally, each agent i has a privately known own-type $\omega_i \in \{G, B\}$. The commonly known prior probability that agent i has good own-type is $Pr(\omega_i = G) = \gamma \in (0, 1)$, and is independent of the own-type realizations of other agents. The space of own-type profiles is $\Omega = \{G, B\}^n$ with generic profile $\omega \in \Omega$. Subsequently, for any variable x_i defined for agent i, the vector x will denote the corresponding profile, i.e. $x = (x_i)_{i \in N}$, unless otherwise stated.

The principal has a single prize to award to an agent of her choosing and may decide to withhold the award. She gets a payoff of $v_G > 0$ and $-v_B < 0$ from awarding the prize to a good own-type and a bad own-type, respectively. Withholding the award brings 0. Agent $i \in N$ receives utility $v > 0$ from getting the prize and 0 from not.

Connections between agents are commonly known and captured by an undirected network L. We write $ij \in L$ if agents i and j are linked (*neighbours*), and denote by N_i the set of all of agent i's neighbours. Let $d_i := |N_i|$ be the number of agent i's neighbours (*degree*). We assume that $d_i \geq 1$ for all i.

In addition to knowing his own-type, each agent may or may not obtain hard evidence about his own-type or that of his neighbours. Formally, given own-type profile realization ω and any $j \in \tilde{N}_i \equiv N_i \cup \{i\}$, agent i privately receives verifiable evidence about j's own-type, $e_{ij} = \omega_j$, with probability $0 < q < 1$ and no evidence, $e_{ij} = \emptyset$, with probability $1 - q$. The evidence realization e_{ij} is i.i.d. for all $i \in N$ and $j \in \tilde{N}_i$. The evidence vector obtained by i is $e_i = (e_{ij})_{j \in \tilde{N}_i}$ and the set of all feasible evidence vectors for agent i given own-type profile ω is $E_i(\omega) = \{e_i | e_{ij} \in \{\omega_j, \emptyset\}$ *for all* $j \in \tilde{N}_i\}$. Agent i's type, as in a standard Bayesian game framework, corresponds to an own-type-evidence realization and is denoted by $t_i = (\omega_i, e_i)$ and the set of all feasible types by $T_i = \{(\omega_i, e_i) | \omega = (\omega_i, \omega_{-i}) \in \Omega, e_i \in E_i(\omega)\}$. The set of all feasible type profiles is therefore $T = \{(\omega, e) | \omega \in \Omega$ *and* $\forall i \in N, e_i \in E_i(\omega)\}$ with generic element t. Let $Pr(t)$ denote the prior probability of type profile t.

Given a type profile realization, $t = (\omega, e)$, agents simultaneously send messages about themselves and their neighbours to the principal. Agent i sends message $m_i = (m_{ij})_{j \in \tilde{N}_i}$ with statement $m_{ij} \in \{e_{ij}, \emptyset\}$. Notice that upon obtaining evidence about j (i.e. $e_{ij} = \omega_j$), i can choose to disclose the evidence, $m_{ij} = \omega_j$, or withhold it, $m_{ij} = \emptyset$. With no evidence about j (i.e. $e_{ij} = \emptyset$), i must send $m_{ij} = \emptyset$. The set of feasible messages for agent i, given t_i, is $M_i(t_i) = \{m_i | m_{ij} \in \{e_{ij}, \emptyset\}\}$, with the set of feasible messages $M_i = \cup_{t \in T} M_i(t_i)$. Let $M = \cup_{t \in T}(\prod_{i \in N} M_i(t_i))$ denote the set of feasible message profiles, with generic profile m. A strategy for agent i, σ_i, assigns to any type t_i a probability distribution over the set of feasible messages, $M_i(t_i)$. Formally, $\sigma_i : T_i \to \Delta(M_i)$ such that $\sigma_i(t_i) \in \Delta M_i(t_i)$, where $\sigma_i(m_i|t_i)$ is the probability with which agent i sends message m_i when his type is t_i.[4] Denote by $\sigma_{ij}(m_{ij}|t_i) := \sum_{m_i = (m_{ij}, \cdot) \in M_i(t_i)} \sigma_i(m_i|t_i)$ the probability with which agent i with type t_i sends message m_{ij} about agent j. Abusing notation, we write $\sigma_{ij}(t_i) = m_{ij}$, $\sigma_i(t_i) = m_i$, and $\sigma(t) = m$ in place of $\sigma_{ij}(m_{ij}|t_i) = 1$, $\sigma_i(m_i|t_i) = 1$, and $\sigma_i(m_i|t_i) = 1, \forall i \in N$, respectively, to represent the use of pure strategies concisely. A strategy σ_i is *completely mixed* if $\sigma_i(m_i|t_i) > 0$ for all $m_i \in M_i(t_i)$ and all $t_i \in T_i$, and a profile σ is completely mixed if σ_i is completely mixed for all $i \in N$.

Upon receiving message profile m, the principal updates her belief about t (which includes ω). Let $\mu_i(t_i|m, \sigma)$ be the principal's belief that agent i's type is $t_i \in T_i$ after receiving message profile m, given agents' strategy profile σ. The corresponding belief profile is μ. The principal's belief that $\omega_i = G$ is then given by $\beta_i(m) := \sum_{t_i \in T_i | \omega_i = G} \mu_i(t_i|m, \sigma)$.[5]

Given her belief, the principal chooses an award rule, $r = (r_1, ..., r_n)$ with $r_i \geq 0$ and $\sum_{i \in N} r_i \leq 1$, where r_i is the probability with which agent i receives the prize. Let R be the set of all award rules. The principal's award strategy is a function $\hat{r} : M \to R$.

The agents' strategy profile σ, the principal's belief μ and award strategy $\hat{r}$ together constitute an assessment of the game. An assessment $(\sigma, \mu, \hat{r})$ is a

[4]For any finite set A, $\Delta(A)$ denotes the set of all probability distributions over A.

[5]The dependence of $\beta_i(m)$ on the belief μ is suppressed for notational convenience.

sequential equilibrium if

1. (*consistency*) there is a sequence of strategy profiles and beliefs $(\sigma^n, \mu^n)_{n=1}^{\infty}$ that converges to (σ, μ) such that each strategy profile σ^n is completely mixed and beliefs μ^n are derived from σ^n using Bayes' rule, and

2. (*sequential rationality*) for each $i \in N$, σ_i maximizes agent i's expected payoff given $(\sigma_{-i}, \hat{r})$ and $\hat{r}$ maximizes the principal's expected payoff, given her belief μ.

The goal of this paper is to show how discrimination may arise even when the principal cares only about an agent's own-type. To this end, we focus on sequential equilibria in which the principal's award strategy is *anonymous*. This requires $\hat{r}_i(m) = \hat{r}_j(m)$ if $\beta_i(m) = \beta_j(m)$.

The function β captures the only part of the belief function μ that is relevant for the principal's choice, and therefore also the agents' strategic considerations. It is also easy to describe, unlike the full belief function μ. Therefore, for expositional ease, we refer to $(\sigma, \beta, \hat{r})$ as an assessment, with the understanding that $(\sigma, \mu, \hat{r})$ is an assessment for some μ, and β is derived from μ.

The model concludes with the following assumption.

Assumption **1.**

$$v_G > v_B \frac{1-\gamma}{\gamma(1-q)^n}.$$

The assumption ensures that in all the scenarios relevant to our analysis, the principal wishes to award the prize to agents she believes most likely to be good, as long as that likelihood is positive. Notice that, independent of this assumption, she strictly prefers not to award an agent she believes to be bad with certainty.

3 Adversarial Disclosure and No Snitching

We start our analysis with some handy observations that hold true for all sequential equilibria. In any sequential equilibrium if the principal receives

evidence G about agent i, then she must be certain that agent i is good. Similarly, evidence B about agent i convinces the principal that the agent is bad. This simply follows from the evidence technology, wherein an evidence of G (B) about agent i can realize only if $\omega_i = G(B)$.

Given a sequential equilibrium $(\sigma, \beta, \hat{r})$, and some message profile $m \in M$, let $W(m) = \{i|\beta_i(m) = \max_k \beta_k(m)\}$ capture the set of agents that the principal believes most likely to be good. If the principal awards the prize at all following the message m, then it is only to members of $W(m)$, since that alone maximizes her expected payoff $\sum_{i\in N} r_i(v_G\beta_i(m) - v_B(1-\beta_i(m)))$. Anonymity ensures that in this case each member of $W(m)$ gets the prize with equal probability. It is also immediate that the principal never awards the prize to an agent she is certain is bad. We collect these observations in the following lemma.

LEMMA **1.** *If $(\sigma, \beta, \hat{r})$ is a sequential equilibrium, then*

$$\begin{aligned} m_{ki} = G &\Rightarrow \beta_i(m) = 1 \\ m_{ki} = B &\Rightarrow \beta_i(m) = 0 \\ \beta_i(m) = 0 &\Rightarrow \hat{r}_i(m) = 0 \end{aligned}$$

$$\left.\begin{array}{ll} \hat{r}_i(m) = 1/|W(m)|, & \text{if } i \in W(m) \\ \hat{r}_i(m) = 0 & \text{otherwise.} \end{array}\right\} \text{ if } \hat{r}_j(m) > 0 \text{ for some } j \in N.$$

The model allows for a rich set of strategic behaviour. Nevertheless, two specific assessments turn out to be particularly salient. In the first, which we label *adversarial disclosure*, the agents send good evidence about themselves, bad evidence about their neighbours and withhold all else.

CONSTRUCTION **1** (*adversarial disclosure*)**.**

$$\sigma_{ij}(t_i) = \begin{cases} G & \text{if } i = j \text{ and } e_{ij} = G \\ B & \text{if } i \neq j \text{ and } e_{ij} = B \\ \emptyset & \text{otherwise} \end{cases}$$

$$\beta_i(m) = \frac{\gamma(1-q)}{\gamma(1-q) + (1-\gamma)(1-q)^{d_i}} \qquad \text{if } m_{ki} = \emptyset, \forall k \in \tilde{N}_i.$$

A key feature of this assessment is the principal's belief about agent i in the absence of any evidence about him. Given the strategy profile, the lack of evidence suggests that either the agent is good and did not obtain evidence to prove it or the agent is bad and none of his neighbours obtained evidence to prove that. The expression is simply the conditional probability of the former being true. This belief depends on agent i's degree and therefore leads the principal to interpret the lack of evidence about two different agents in different ways, in turn leading to an award rule that discriminates in favour of agents with greater or fewer connections, based on the parameters of the model. This feature is at the heart of the main findings in this paper.

We label the second salient assessment *no snitching*. In it, agents send good evidence about themselves and withhold all else. In the absence of any evidence about agent i, the principal believes agent i to be bad with certainty, if he sends bad evidence about anyone else.

CONSTRUCTION **2** (*no snitching*)**.**

$$\sigma_{ij}(t_i) = \begin{cases} G & \text{if } i = j \text{ and } e_{ij} = G \\ \emptyset & \text{otherwise} \end{cases}$$

$$\beta_i(m) = \begin{cases} \frac{\gamma(1-q)}{\gamma(1-q)+(1-\gamma)} & \text{if } m_{ki} = \emptyset, m_{ik} \neq B, \forall k \in \tilde{N}_i \\ 0 & \text{if } m_{ki} = \emptyset, \forall k \in \tilde{N}_i \text{ and } m_{ij} = B \text{ for some } j \in N_i. \end{cases}$$

The assessments described in constructions 1 and 2 are completed with the specification in Lemma 1 and a reward rule $\hat{r}$ such that

$$\hat{r}_i(m) = \begin{cases} 1/|W(m)|, & \text{if } i \in W(m), \max_k \beta_k(m) \geq \frac{v_B}{v_G+v_B} \\ 0, & \text{otherwise.} \end{cases}$$

The next lemma establishes an important property of both assessments above. Following any message profile, the principal is either certain that all

agents are bad, or the highest probability she assigns to some agent being good exceeds a threshold. The proof is in the appendix.

LEMMA **2.** *Suppose $(\sigma, \beta, \hat{r})$ is either the no snitching or the adversarial disclosure assessment, and $m = \sigma(t)$ for some $t \in T$. Then,*

$$\max_k \beta_k(m) > 0 \Rightarrow \max_k \beta_k(m) \geq \frac{v_B}{v_G + v_B}.$$

The lemma implies that in both assessments, either the principal receives bad evidence about every agent and the prize is not awarded at all, or the prize is awarded with equal odds to all agents that are most likely to be good. We can now state our first result.

PROPOSITION **1.** *Adversarial disclosure and no snitching are sequential equilibria for all values of the parameters γ, q.*

Proof. For each of the two constructions, we need to show that there exists an assessment $(\sigma, \mu, \hat{r})$ where $\sigma, \hat{r}$ are as specified in the construction and β can be derived from μ. Further, $(\sigma, \mu, \hat{r})$ must satisfy the requirements of consistency and sequential rationality. We relegate the derivation of β from μ and the proof of consistency to the appendix. Here, we establish sequential rationality, for which the information in $(\sigma, \beta, \hat{r})$ is sufficient.

The reward rule is indeed sequentially rational. In particular, the principal would not award the prize at all, if doing so to the agent she believes most likely to be good still brings her a negative expected payoff,

$$v_G(\max_k \beta_k(m)) - v_B(1 - \max_k \beta_k(m)) < 0 \Leftrightarrow \max_k \beta_k(m) < \frac{v_B}{v_G + v_B}.$$

If instead $\max_k \beta_k(m) \geq \frac{v_B}{v_G + v_B}$, then she cannot do better than follow $\hat{r}$ by awarding the prize to members of the set $W(m)$ with equal probability.

Now consider the adversarial disclosure profile. Given $\hat{r}$ and σ_{-i}, no agent i has an incentive to deviate from σ_i: Suppose $\omega_i = G$. Then any message $m'_i \in M_i(t_i)$ where $m'_{ii} = \emptyset$ if $e_{ii} = G$ or $m'_{ij} = \emptyset$ if $e_{ij} = B$ or $m'_{ij} = G$ if $e_{ij} = G$ implies $\beta_i(m'_i, \sigma_{-i}(t_{-i})) \leq \beta_i(\sigma(t))$ and $\beta_j(m'_i, \sigma_{-i}(t_{-i})) \geq \beta_j(\sigma(t))$

for $j \neq i$ and hence $\hat{r}_i(m'_i, \sigma_{-i}(t_{-i})) \leq \hat{r}_i(\sigma(t))$. Suppose $\omega_i = B$. Then any message $m'_i \in M_i(t_i)$ where $m'_{ii} = B$ if $e_{ii} = B$ or $m'_{ij} = \emptyset$ if $e_{ij} = B$ or $m'_{ij} = G$ if $e_{ij} = G$ implies $\beta_i(m'_i, \sigma_{-i}(t_{-i})) \leq \beta_i(\sigma(t))$ and $\beta_j(m'_i, \sigma_{-i}(t_{-i})) \geq \beta_j(\sigma(t))$ for $j \neq i$ and hence $\hat{r}_i(m'_i, \sigma_{-i}(t_{-i})) \leq \hat{r}_i(\sigma(t))$.

Consider next the no snitching profile, σ. Fix some $t \in T$, and suppose $\sigma(t) = m$. It follows that $\max_k \beta_k(m) > 0$. Then by Lemma 2 we have $\max_k \beta_k(m) \geq v_B/(v_G + v_B)$. Now, agent i has only two ways to change his payoff by deviating to m'_i. Either $i \notin W(m)$ and $i \in W(m'_i, m_{-i})$ with $\max_k \beta_k(m'_i, m_{-i}) \geq v_B/(v_G + v_B)$ or $i \in W(m) \cap W(m')$ and $|W(m')| < |W(m)|$, again with $\max_k \beta_k(m'_i, m_{-i}) \geq v_B/(v_G + v_B)$. Notice if $i \notin W(m)$, then it must be that $m_{ji} \neq G$ for all $j \in \tilde{N}_i$. Deviating to sending good evidence about some neighbour ensures that $i \notin W(m')$. Sending bad evidence about a neighbour, in turn, ensures that $\beta_i(m) = 0$. So the first possibility of profitable deviation is ruled out. Now suppose $i \in W(m) \cap W(m')$. If $\beta_i(m) = 1$, then under σ it must be that $m_{ii} = G$. But then it must be that $m_{jj} = G$ for all $j \in W(m)$. In this case, no matter what $m'_i \in M_i(t)$ agent i sends instead, $j \in W(m)$ implies $j \in W(m'_i, m_{-i})$. Therefore $i \in W(m) \cap W(m')$ and $|W(m')| < |W(m)|$ cannot be true. If $\beta_i(m) \neq 1$, then it must be that $m_{ji} \neq G$ for all $j \in \tilde{N}_i$. In this case, the only message m'_i that can lead to $|W(m')| < |W(m)|$ must have $m'_{ij} = B$ for some $j \in W(m)$. But this results in $\beta_i(m'_i, m_{-i}) = 0$ and therefore rules out $i \in W(m) \cap W(m')$ and $\max_k \beta_k(m'_i, m_{-i}) \geq v_B/(v_G + v_B)$.

□

The evidence disclosure strategy in the adversarial disclosure equilibrium is more Blackwell-informative than that in the no snitching equilibrium, since the latter is a garbling of the former. Therefore, by Blackwell (1951, 1953), we obtain the following observation.

Observation **1.** *The principal has a higher ex-ante payoff in the adversarial disclosure equilibrium than in the no snitching equilibrium.*

Given a chance, therefore, the principal would opt for the adversarial disclosure equilibrium. This raises a legitimate concern as to whether, beyond being an

equilibrium, the no snitching profile has any compelling reason to prevail as a norm. The next section finds one such rationale; the no snitching equilibrium, unlike the adversarial disclosure equilibrium, does not permit discrimination.

4 Ex-ante award probabilities and discrimination

We now turn to the key object of our analysis, the probability with which an agent expects to win the prize before his type is realized. We call this the agent's *ex-ante award probability*. The next result rules out discrimination in the no snitching equilibrium.

PROPOSITION **2.** *In the no snitching equilibrium the ex-ante award probability is identical for all agents.*

Proof. An agent wins the prize in this equilibrium in two different ways. In the first, the principal receives no evidence about any agent. The probability of winning this way is $[(1-\gamma)+\gamma(1-q)]^n/n$. In the second way, the agent provides good evidence about himself and obtains the prize with equal odds with every other agent that also provides good evidence. The probability of winning this way as part of a winning group of size $k+1$ is simply

$$\frac{[(1-\gamma)+\gamma(1-q)]^{n-k-1}[\gamma q]^{k+1}}{k+1}$$

and there are $\binom{n-1}{k}$ different $k+1$ sized winning groups the agent could be a part of. Therefore, agent i's ex-ante award probability, P_i, satisfies

$$P_i = \frac{[(1-\gamma)+\gamma(1-q)]^n}{n} + \sum_{k=0}^{n-1} \binom{n-1}{k} \frac{[(1-\gamma)+\gamma(1-q)]^{n-k-1}[\gamma q]^{k+1}}{k+1},$$

which is independent of i. □

A similar direct computation of agents' ex-ante award probabilities for the adversarial disclosure equilibrium is unwieldy and therefore ill suited for our

purpose. Instead, we proceed by establishing some partial findings that, taken together, help us derive our main result.

LEMMA **3.** *Fix $i, j \in N$. Consider the adversarial disclosure equilibrium $(\sigma, \beta, \hat{r})$ and m such that $m = \sigma(t)$ for some t and $m_{ki} = m_{lj} = \emptyset$ for all k, l. Then, $\beta_i(m) > \beta_j(m)$ if and only if $d_i > d_j$.*

Lemma 3 follows immediately from Construction 1 by comparing the beliefs for agents with different degrees. The result implies that despite agents having the same prior probability of being good, the principal prefers to award an agent with a higher degree over one with a lower degree, conditional on not receiving any evidence about either agent.

The intuition is simple. Recall that the absence of evidence about an agent in the adversarial disclosure equilibrium arises either if the agent is good and he does not obtain evidence for it, or if the agent is bad and his neighbours obtain no evidence of the same. The probability of the former is independent of the agent's degree, but the latter is less likely the more neighbours the agent has. As a result, the absence of evidence about an agent is a stronger signal of the agent being good, the more neighbours he has.

The next lemma establishes that agents with the same degree have the same ex-ante award probabilities. The proof is relegated to the appendix. Partition the set of individuals in N according to their degree, $\{D^i\}_{i=1}^K$ such that $i, j \in D^k$ implies $d_i = d_j$. Let d^i be the degree of any agent in group D^i. Let these K sets be ordered such that $d^i > d^{i+1}$. Abusing notation we will use D^i both as a set and its corresponding cardinality.

LEMMA **4.** *In the adversarial disclosure equilibrium, if $i, j \in D^k$, then the ex-ante award probabilities for i and j are the same.*

We next compare ex-ante award probabilities across agents with different degrees. Let the event that every agent is either bad, or good without evidence about himself be denoted by $E = \{t \mid \omega_i = B \text{ or } (\omega_i = G, e_{ii} = \emptyset) \text{ for each } i \in N\}$. Let E^c denote the complement event, in that at least one agent is good and has evidence about it.

LEMMA **5.** *In the adversarial disclosure equilibrium, the probability of i winning the award and E^c being true is identical for all $i \in N$.*

The proof is in the appendix.

By contrast, in the event that no agent is proven good, E, the probability of winning the award depends on the agent's degree. This in turn yields our main result on discrimination.

PROPOSITION **3.** *In the adversarial disclosure equilibrium,*
(a) for a given γ, agents with higher degrees have higher ex-ante award probabilities for sufficiently small q,
(b) for a given q, agents with higher degrees have higher ex-ante award probabilities for sufficiently high γ,
(c) for a given γ, agents with lower degrees have higher ex-ante probabilities for sufficiently large q.

Proof. In the adversarial disclosure equilibrium, let P^i be the ex-ante probability that event E occurs and someone in D^i wins the prize. Then

$$P^i = [(1-\gamma)+\gamma(1-q)]^{N\setminus\bigcup_{j=1}^{i} D^j} \prod_{j=1}^{i-1} \left\{(1-\gamma)^{D^j}[1-(1-q)^{d^j}]^{D^j}\right\} \left\{[(1-\gamma)+\gamma(1-q)]^{D^i} - (1-\gamma)^{D^i}[1-(1-q)^{d^i}]^{D^i}\right\}. \quad (1)$$

The first in the product of three expressions above is the probability that all agents with strictly lower degree than i's do not have good evidence for their own selves. The second expression is the probability that each agent with strictly higher degree than i's is bad and at least one neighbour of theirs has evidence for it. The final expression is the probability that no agent in D^i has good evidence about himself and yet it is not true that each of them has a neighbour with bad evidence about him. In summary, the product is the probability of the event that all agents with higher degree than i are proven bad, those with lower degree than i are not proven good and not every agent in D^i is proven bad.

It follows that

$$P^i/D^i \geq P^{i+1}/D^{i+1}$$

$$\Leftrightarrow \frac{\left[\left\{\frac{[(1-\gamma)+\gamma(1-q)]}{(1-\gamma)[1-(1-q)^{d^i}]}\right\}^{D^i} - 1\right]}{D^i} \geq \frac{\left[1 - \left\{\frac{(1-\gamma)[1-(1-q)^{d^{i+1}}]}{[(1-\gamma)+\gamma(1-q)]}\right\}^{D^{i+1}}\right]}{D^{i+1}}. \qquad (2)$$

Observe that the right hand side is bounded above by 1. For any given γ, the left hand side goes to ∞ as $q \to 0$. Similarly, for any given q, the left hand side goes to ∞ as $\gamma \to 1$. Together with Lemma 4 and 5, this proves parts (a) and (b) of the proposition.

Let $a = \frac{[(1-\gamma)+\gamma(1-q)]}{(1-\gamma)[1-(1-q)^{d^i}]}$ and $b = \frac{(1-\gamma)[1-(1-q)^{d^{i+1}}]}{[(1-\gamma)+\gamma(1-q)]}$. The weak opposite of inequality 2 can be written as

$$\begin{aligned}
&P^i/D^i \leq P^{i+1}/D^{i+1} \\
&\Leftrightarrow \left[a^{D^i} - 1\right]/D^i \leq \left[1 - b^{D^{i+1}}\right]/D^{i+1} \\
&\Leftrightarrow \left[(a-1)(\sum_{j=1}^{i} a^{D^{i-j}})\right]/D^i \leq \left[(1-b)(\sum_{j=0}^{i} b^{D^{i-j}})\right]/D^{i+1} \\
&\Leftrightarrow \frac{a-1}{1-b} \leq \frac{\sum_{j=0}^{i} b^{D^{i-j}}}{\sum_{j=1}^{i} a^{D^{i-j}}} \frac{D^i}{D^{i+1}} \\
&\Leftrightarrow \frac{(1-\gamma)(1-q)^{d^i} + \gamma(1-q)}{(1-\gamma)(1-q)^{d^{i+1}} + \gamma(1-q)} \frac{(1-\gamma)+\gamma(1-q)}{(1-\gamma)[1-(1-q)^{d^i}]} \leq \frac{\sum_{j=0}^{i} b^{D^{i-j}}}{\sum_{j=1}^{i} a^{D^{i-j}}} \frac{D^i}{D^{i+1}} \\
&\Leftrightarrow z^l \equiv \frac{(1-\gamma)(1-q)^{d^i} + \gamma(1-q)}{(1-\gamma)(1-q)^{d^{i+1}} + \gamma(1-q)} \leq \frac{(1-\gamma)[1-(1-q)^{d^i}]}{(1-\gamma)+\gamma(1-q)} \frac{\sum_{j=0}^{i} b^{D^{i-j}}}{\sum_{j=1}^{i} a^{D^{i-j}}} \frac{D^i}{D^{i+1}} \equiv z^r
\end{aligned}$$

Now notice that $\lim_{q\to 1} z^r = \frac{i+1}{i}\frac{D^i}{D^{i+1}}$. On the other hand, $\lim_{q\to 1} z^l$ takes the indeterminate form $\frac{0}{0}$. Repeated use of l'Hospital's rule gives

$$\lim_{q\to 1} z^l = \lim_{q\to 1} \frac{(1-q)^{d^i - d^{i+1}} \prod_{j=0}^{d^{i+1}} (d^i - j)}{d^{i+1}!} = 0.$$

Together with Lemma 4 and 5 this proves part (c) of the proposition. □

To see the intuition for this result, consider two agents with different degrees. There are three salient and jointly exhaustive outcome scenarios. In the first, there is good evidence about some agent on the network. The absence of good evidence is split into the second and third scenarios. In the second, additionally there is no bad evidence about the two agents we consider, while in the third there is bad evidence about at least one of the two.

The odds of winning in the first scenario is identical across all agents, by Lemma 5. In the second scenario, it is better to have more neighbours, by Lemma 3. The third scenario benefits the agent with fewer neighbours since that lowers the risk of being snitched on.

Fixing γ, a lower chance of obtaining evidence (smaller q) increases the probability of the second scenario at the expense of the other two, benefitting the agent with more neighbours. Fixing q, increasing the prior probability of an agent being good (higher γ) increases the probability of the first scenario. Importantly, this also increases the odds of second scenario relative to the third, since only bad evidence is transmitted by neighbours which in turn requires the agent to be bad. This, too, benefits the agent with more neighbours.

Fixing γ, better evidence technology (higher q) increases the probability of the first scenario at the expense of the other two, but it also raises the relative odds of the third scenario over the second, benefitting the agent with fewer neighbours. The effect of lowering γ keeping q fixed is ambiguous, since it raises the odds of both the second and third scenarios relative to the first.

5 Robustness

Agents in our model only care about winning the prize. Upon losing out, they are indifferent about whether some other agent wins, and if so, which one. This indifference permits a multiplicity of equilibria held together by uncompelling strategies and beliefs. For instance, an agent with good evidence about a neighbour and bad evidence about himself may choose to reveal both. The equilibrium logic behind such behaviour is as follows. Suppose in the absence

of any evidence about either the agent or his neighbour, the principal infers the neighbour has a higher probability of being good (as can happen in the adverse disclosure equilibrium). Knowing this, the agent realizes that he simply cannot win. Since he is bad and his neighbour is good, the agent's best hope is that there is no evidence about either. But even then he loses out to his neighbour.

Consider a principal who responds to the agents' messages in a coarse manner. In particular, she continues to treat good or bad evidence about an agent as conclusive but treats all agents she lacks evidence on identically. In the absence of any good evidence, she awards the prize to members of the latter group with equal odds. The indifference described above is broken with even the smallest probability that the principal is of this coarser kind.

We now define this notion of coarse principal formally. For any feasible message profile $m \in M$, let $B(m) = \{i \in N | m_{ji} = B \text{ for some } j \in \tilde{N}_i\}$ denote the set of all agents proven bad. Likewise let $G(m) = \{i \in N | m_{ji} = G \text{ for some } j \in \tilde{N}_i\}$ be the set of all agents proven good. Finally, let $I(m) = N \setminus (B(m) \cup G(m))$ be the set of agents about whom there is no evidence in m.

DEFINITION **1** (*coarse principal*)**.** A principal is *coarse* if she uses the award rule r^c where

$$r_i^c(m) = \begin{cases} 0 & \text{if } i \in B(m) \\ 1/|G(m)| & \text{if } i \in G(m) \\ 1/|I(m)| & \text{if } i \in I(m) \text{ and } G(m) = \emptyset. \end{cases}$$

A principal is ϵ-coarse if with probability $1 - \epsilon$ she is a standard principal as described in section 2 and with probability ϵ she is coarse.

DEFINITION **2** (*sequential equilibrium with an ϵ-coarse principal*)**.** An assessment $(\sigma, \mu, \hat{r})$ is a *sequential equilibrium with an ϵ-coarse principal* if

1. (*consistency*) there is a sequence of strategy profiles and beliefs $(\sigma^n, \mu^n)_{n=1}^{\infty}$ that converges to (σ, μ) such that each strategy profile σ^n is completely mixed and beliefs μ^n are derived from σ^n using Bayes' rule,

2. (*principal's sequential rationality*) $\hat{r}$ maximizes the principal's expected payoff, given her belief μ,

3. (*agents' sequential rationality*) for each $i \in N$, σ_i maximizes agent i's expected payoff given σ_{-i} and the principal's award strategy, which is $\hat{r}$ with probability $1-\epsilon$ and r^c with probability ϵ,

for all small enough ϵ.

Notice that any sequential equilibrium would satisfy properties 1 and 2 above, by definition. The stricter requirement only applies to the agents' strategies, which must now remain optimal when facing a principal who is coarse with vanishing probability.

Finally, we import a notion of robustness used in Ben-Porath, Dekel, and Lipman (2019). Let $M_{-i}(t_i) = \cup_{(t_i,\cdot)\in T}(\prod_{j\in N\setminus\{i\}} M_j(t_j))$ denote the set of feasible message profiles of agents $N \setminus \{i\}$ given that agent i's type is t_i.

Definition **3** (*robustness*)**.** A sequential equilibrium with an ϵ-coarse principal, $(\sigma, \mu, \hat{r})$, is *robust* if, for all $i \in N$, $t_i \in T_i$, $\sigma_i(t_i)$ maximizes agent i's expected payoff for any $m_{-i} \in M_{-i}(t_i)$, given that the principal plays $\hat{r}$ with probability $1-\epsilon$ and r^c with probability ϵ.

In other words, an agent's strategy $\sigma_i(t_i)$ must be optimal for t_i, given *any* feasible message profile sent by other agents and the *equilibrium* strategy of the ϵ-coarse principal. Robust equilibria are less demanding in terms of how precisely an agent anticipates the strategies of other agents. Consequently, they survive as equilibria in a range of related games, such as where the agents send their messages sequentially with each agent observing the earlier reports. Agents' strategies in any such robust equilibrium share a number of features, as we show below.

Proposition **4.** *Suppose $(\sigma, \mu, \hat{r})$ is a robust sequential equilibrium with an ϵ-coarse principal. Then,*
(a) agents do not reveal bad evidence about themselves, $e_{ii} = B \Rightarrow \sigma_{ii}(t_i) = \emptyset$,
(b) agents do not reveal good evidence about others, $e_{ij} = G \Rightarrow \sigma_{ij}(t_i) = \emptyset$ for

$i \neq j$, *and*
(c) agents (almost) always reveal good evidence about themselves,
$e_{ii} = G \Rightarrow \sigma_{ii}(t_i) = G$, *unless* $\tilde{N}_i = N$ *and* $e_{ij} = B, \forall j \in N_i$.

Proof. (a) Fix some $i \in N$ and suppose $e_{ii} = B$. Take $m_{-i} \in M_{-i}(t_i)$ such that $m_{jk} = \emptyset$ for all k and $j \neq i$ and message m_i with $m_{ii} = B$. Then $\hat{r}_i(m_i, m_{-i}) = r_i^c(m_i, m_{-i}) = 0$, by Lemma 1 and the definition of a coarse principal. Now consider message m_i' with $m_{ij}' = \emptyset$ for all $j \in \tilde{N}_i$. Then $\hat{r}_i(m_i', m_{-i}) \geq 0$ and $r_i^c(m_i', m_{-i}) = 1/n > 0$. Message m_i' is strictly better for agent i than m_i given m_{-i}. Therefore $\sigma_{ii}(B|t_i) > 0$ fails robustness.

(b) Fix some $i \in N$. Suppose first that $\omega_i = G$ and $m_{ij} = G$ for some $j \in N_i$. Let m_{-i} be such that $m_{kk} \neq \emptyset$ for all $k \neq j$ and $m_{kj} = \emptyset$ for all $k \in \tilde{N}_j \setminus \{i\}$. Further, $m_{li} = G$ for some $l \in N_i$. Then $\hat{r}_i(m_i, m_{-i}) = r_i^c(m_i, m_{-i}) = 1/|G(m_i, m_{-i})|$. Now consider message m_i' with $m_{ij}' = \emptyset$ and $m_{ik}' = m_{ik}$ for all $k \neq j$. Then $\hat{r}_i(m_i', m_{-i}) \geq 1/|G(m_i, m_{-i})|$ and $r_i^c(m_i', m_{-i}) = 1/(|G(m_i, m_{-i})| - 1)$. Agent i is strictly better off sending message m_i' instead of m_i given m_{-i}. Thus, $\sigma_{ij}(G|t_i) > 0$ with $i \neq j$ cannot occur if $\omega_i = G$ in a robust sequential equilibrium with an ϵ-coarse principal.

Now suppose that $\omega_i = B$ and $m_{ij} = G$ for some $j \in N_i$. Consider m_{-i} such that $m_{kl} = \emptyset$ for all k, l. Notice that $m = (m_i, m_{-i})$ occurs on the equilibrium path if $\sigma_{ij}(G|t_i) > 0$. Since the argument above rules out agent i of own-type $\omega_i = G$ from sending such an m_i, it must be that $\beta_i(m_i, m_{-i}) = 0$. Thus $\hat{r}_i(m_i, m_{-i}) = r_i^c(m_i, m_{-i}) = 0$. Consider instead the message m_i' such that $m_{ij}' = \emptyset$ for all j. Then $\hat{r}_i(m_i', m_{-i}) \geq 0$ and $r_i^c(m_i', m_{-i}) > 0$. Since message m_i' is strictly better for agent i than m_i given m_{-i}, in a robust sequential equilibrium with an ϵ-coarse principal, $\sigma_{ij}(G|t_i) > 0$ with $i \neq j$ cannot occur if $\omega_i = B$.

(c) Fix some $i \in N$. Suppose it is not true that $\tilde{N}_i = N$ and $e_{ij} = B, \forall j \in N_i$. Let $e_{ii} = G$. Take $m_{-i} \in M_{-i}(t_i)$ such that $m_{kk} \neq \emptyset$ for all $k \neq i$, $m_{kk} = G$ for some $k \neq i$ and $m_{ki} = \emptyset$ for all $k \in N_i$. Note that such m_{-i} exists because $\tilde{N}_i \neq N$ or $e_{ij} \neq B$ for some $j \in N_i$, by assumption.

Consider any message m_i with $m_{ii} = \emptyset$. Then

$$\hat{r}_i(m_i, m_{-i}) \leq 1/(|G(m_i, m_{-i})| + 1)$$

because $\hat{r}$ randomizes uniformly over at least all agents about whom the principal receives evidence G. Moreover, $r_i^c(m_i, m_{-i}) = 0$.

Now consider message m_i' with $m_{ii}' = G$ and $m_{ij}' = m_{ij}$ for all $j \neq i$. Then $\hat{r}_i(m_i', m_{-i}) = 1/(|G(m_i, m_{-i})| + 1) = r_i^c(m_i', m_{-i})$ where the first equality follows from $m_{ii}' = G$ and $m_{kk} \neq \emptyset$ for all $k \neq i$. Message m_i' is strictly better for agent i than m_i given m_{-i}. Therefore for $(\sigma, \mu, \hat{r})$ to be a robust sequential equilibrium with an ϵ-coarse principal, it follows that $e_{ii} = G \Rightarrow \sigma_{ii}(t_i) = G$.

□

It is easy to verify that the agents' strategies in both the adversarial disclosure and no snitching equilibria satisfy the three properties in Proposition 4. The next result shows that both equilibria indeed satisfy the robustness criteria.

PROPOSITION **5.** *Adversarial disclosure and no snitching are both robust sequential equilibria with an ϵ-coarse principal.*

Proof. Consider the adversarial disclosure profile. The argument used to establish Proposition 1 can be generalized to show that given $\hat{r}$, no agent i has an incentive to deviate from σ_i, no matter the messages sent by the other agents. Suppose $\omega_i = G$. Then any message $m_i' \in M_i(t_i)$ where $m_{ii}' = \emptyset$ if $e_{ii} = G$ or $m_{ij}' = \emptyset$ if $e_{ij} = B$ or $m_{ij}' = G$ if $e_{ij} = G$ implies $\beta_i(m_i', m_{-i}) \leq \beta_i(\sigma_i(t_i), m_{-i})$ and $\beta_j(m_i', m_{-i}) \geq \beta_j(\sigma_i(t_i), m_{-i})$ for $j \neq i$ and hence $\hat{r}_i(m_i', m_{-i}) \leq \hat{r}_i(\sigma_i(t_i), m_{-i})$ for any $m_{-i} \in M_{-i}(t_i)$. Suppose $\omega_i = B$. Then any message $m_i' \in M_i(t_i)$ where $m_{ii}' = B$ if $e_{ii} = B$ or $m_{ij}' = \emptyset$ if $e_{ij} = B$ or $m_{ij}' = G$ if $e_{ij} = G$ implies $\beta_i(m_i', m_{-i}) \leq \beta_i(\sigma_i(t_i), m_{-i})$ and $\beta_j(m_i', m_{-i}) \geq \beta_j(\sigma_i(t_i), m_{-i})$ for $j \neq i$ and hence $\hat{r}_i(m_i', m_{-i}) \leq \hat{r}_i(\sigma_i(t_i), m_{-i})$ for any $m_{-i} \in M_{-i}(t_i)$. It is easy to verify that for any $t_i \in T_i$ and $m_i' \in M_i(t_i)$, $r_i^c(m_i', m_{-i}) \leq r_i^c(\sigma_i(t_i), m_{-i})$ for any $m_{-i} \in M_{-i}(t_i)$. Therefore the adversarial disclosure profile satisfies our robustness criteria.

Now consider the no-snitching profile. Fix $i \in N$, $t_i \in T_i$ and some $m_{-i} \in M_{-i}(t_i)$. Suppose $m_{ji} = B$ for some $j \in N_i$. Then $\hat{r}_i(\sigma_i(t_i), m_{-i}) = r_i^c(\sigma_i(t_i), m_{-i}) = \hat{r}_i(m_i', m_{-i}) = r_i^c(m_i', m_{-i}) = 0$ for any $m_i' \in M_i(t_i)$. Suppose instead that $m_{ji} = G$ for some $j \in \tilde{N}_i$. Then $\hat{r}_i(\sigma_i(t_i), m_{-i}) = r_i^c(\sigma_i(t_i), m_{-i}) = 1/|G(\sigma_i(t_i), m_{-i})|$. Further, for any $m_i' \in M_i(t_i)$, both

$$\hat{r}_i(m_i', m_{-i}) \leq 1/|G(\sigma_i(t_i), m_{-i})| \text{ and } r_i^c(m_i', m_{-i}) \leq 1/|G(\sigma_i(t_i), m_{-i})|$$

must be true since the set of other agents proven good to the principal cannot be shrunk by any feasible message from agent i.

Finally suppose $i \in I(\sigma_i(t_i), m_{-i})$. If $G(\sigma_i(t_i), m_{-i}) \neq \emptyset$, then again $\hat{r}_i(\sigma_i(t_i), m_{-i}) = r_i^c(\sigma_i(t_i), m_{-i}) = \hat{r}_i(m_i', m_{-i}) = r_i^c(m_i', m_{-i}) = 0$ for any $m_i' \in M_i(t_i)$. Suppose instead that $G(\sigma_i(t_i), m_{-i}) = \emptyset$. In this case $\hat{r}_i(\sigma_i(t_i), m_{-i}) = r_i^c(\sigma_i(t_i), m_{-i}) = 1/|I(\sigma_i(t_i), m_{-i})|$. Agent i stands to gain here with the coarse principal by revealing bad evidence, if he possessed it, about some neighbour of his in the set $I(\sigma_i(t_i), m_{-i})$. Such a deviation would deliver $r_i^c(m_i', m_{-i}) \leq 1$. Any such deviation would also mean $\hat{r}_i(m_i', m_{-i}) = 0$. So a maximum gain of $1 - 1/|I(\sigma_i(t_i), m_{-i})|$ which occurs with probability ϵ comes with a loss of $1/|I(\sigma_i(t_i), m_{-i})|$ with probability $1-\epsilon$. For small enough ϵ, agent i is worse off from such a deviation. This completes the argument for why the no snitching profile satisfies our robustness criteria. □

Proposition 4 and 5, taken together, deliver a useful characterization of all robust sequential equilibria with an ϵ-coarse principal. Any such equilibrium lies between the no snitching and adversarial disclosure equilibria in the following sense. Firstly, in equilibrium the evidence revealed by agents is never less than that in the no snitching equilibrium, in that good evidence about self is always revealed. Secondly, in equilibrium the revealed evidence never exceeds that in the adversarial disclosure equilibrium, in that bad evidence about self and good evidence about others are never revealed.

6 Discussion

A central feature in theories of equilibrium discrimination is that different groups behave differently in equilibrium. These disparate choices remain self enforcing due to the beliefs of the principal (e.g., employer), but it is nevertheless this very difference in action that permits discrimination. For instance, in Arrow (1973), black workers invest less in human capital than white workers.[6] By contrast, all agents in the adversarial disclosure equilibrium follow the same disclosure norm. This overt homogeneity in behaviour interacts with the heterogeneous network positions of agents to yield discrimination in equilibrium. In this sense, our results better fit the notion of institutional discrimination, as discussed in Small and Pager (2020). In the latter, institutional practices that seem neutral end up discriminating based on some payoff irrelevant dimension.

An institutional description of our model and specifically the adversarial disclosure equilibrium would include the principal's award rule. Our assumption of anonymity requires the award rule to solely depend on the principal's belief about an agent's own-type. On the face of it, this is a neutral institutional practice. Indeed, it rules out discrimination that may arise simply due to the principal breaking ties in a way that favours some agent along a payoff-irrelevant dimension.[7] There are often a variety of legal constraints that rule out such behaviour.[8] However, one may ask whether non-anonymous award rules could ameliorate the discrimination that arises in the more informative adversarial disclosure equilibrium. For parameter values that favour the better connected, the principal upon receiving good evidence about multiple agents could award the prize with a higher probability to those with lower degree. This is similar to a common recommendation by diversity, equity and inclusion practices in hiring, wherein employers rank candidates in coarse categories and

[6] See also Coate and Loury (1993), and more recently Onuchic ⓡ Ray (forthcoming). For a survey on recent theoretical contributions to the economics of discrimination see Onuchic (2022).

[7] Consider the modification of the no snitching equilibrium under which upon observing good evidence about multiple agents the principal awards the prize to those with higher degree with strictly higher probability.

[8] See, for instance, Canadian Human Rights Act, Part 1, 3(1).

then break ties in the highest category in favour of marginalized groups.

We have strived to keep our model simple in order to describe our main finding in a transparent manner. A particular feature is worth highlighting. The competition we impose in our model is of an extreme nature, with all agents competing for a single prize. While this does correspond to a number economic settings of interest, it would be valuable to understand the implication of multiple prizes. We leave that for future work.

Appendix

Proof for Lemma 2. The premise of $\max_k \beta_k(m) > 0$ implies that there exists $i \in N$ such that $m_{ji} \neq B$ for all $j \in \tilde{N}_i$. If $m_{ji} = G$ for some i, j, then $\max_k \beta_k(m) = 1$ and the implication follows. So it only remains to show that if there is no evidence about some agents and no good evidence at all, then in either assessment the highest probability of being good assigned to some agent is sufficiently high. This probability, $\max_k \beta_k(m)$, in the no snitching equilibrium is

$$\begin{aligned}\frac{\gamma(1-q)}{\gamma(1-q)+(1-\gamma)(1-q)} &\geq \frac{\gamma(1-q)^n}{\gamma(1-q)+(1-\gamma)(1-q)^n} \\ &\geq \frac{v_B}{v_G+v_B}\end{aligned}$$

where the second inequality follows from Assumption 1. In the adversarial disclosure equilibrium, we have

$$\max_k \beta_k(m) = \max_i \frac{\gamma(1-q)}{\gamma(1-q)+(1-\gamma)(1-q)^{d_i}} \geq \frac{\gamma(1-q)}{\gamma(1-q)+(1-\gamma)(1-q)}.$$

□

Proof for Lemma 4. For all type profiles $t \in T$ such that both i and j win the award with positive probability, this probability is identical given $\hat{r}$. Likewise for type profiles in which neither wins the award the probability of winning is identical (zero). So it suffices to show that in the event of only one of i and j being part of $W(m)$ and $\max_k \beta_k(m) > 0$, the probability of receiving the prize is the same for i and j.

Let T^{ij} be the set of all $t \in T$ such that $\max_k \beta_k(m) > 0$ and $|\{i, j\} \cap W(m)| = 1$ where $m = \sigma(t)$. Consider a one-to-one mapping $\rho : N \to N$.

$$\begin{aligned}\rho(k) &= k \qquad \forall k \notin \tilde{N}_i \cup \tilde{N}_j, \\ \rho(i) &= j, \\ \rho(j) &= i,\end{aligned}$$

$$\rho(k) \in N_i \qquad \text{if } k \in N_j,$$
$$\rho(k) = \rho^{-1}(k) \in N_j \qquad \text{if } k \in N_i.$$

Such a mapping exists since $d_i = d_j$. Next consider the mapping $\tilde{t} : T^{ij} \to T^{ij}$ where $\tilde{t}(\omega, e) = (\tilde{\omega}, \tilde{e})$ such that

$$\begin{aligned} \tilde{\omega}_{-ij} &= \omega_{-ij} \\ \tilde{e}_{kl} &= e_{kl} \qquad \forall l \neq i, j \\ \tilde{e}_{\rho(k)j} &= e_{ki} \\ \tilde{e}_{\rho(k)i} &= e_{kj}. \end{aligned}$$

Let T^i and T^j be the subsets of T^{ij} in which i and j win the prize with positive probability, respectively. Consider $t \in T^i$. Under the transformation $\tilde{t}(t)$, the own-types of i and j are interchanged as is the evidence profile about the two. All else remains the same. This implies that $\beta_i(m) = \beta_j(m')$ and $\beta_j(m) = \beta_i(m')$ where $m = \sigma(t)$ and $m' = \sigma(\tilde{t})$. But notice that messages about all agents $k \neq i, j$ remain unchanged between t and $\tilde{t}$, and so for such k, $\beta_k(m) = \beta_k(m')$. Therefore, it must be that $t \in T^i \Rightarrow \tilde{t}(t) \in T^j$. Likewise $t \in T^j \Rightarrow \tilde{t}(t) \in T^i$. We also have $|W(\sigma(t))| = |W(\sigma(\tilde{t}(t)))|$ and $Pr(t) = Pr(\tilde{t}(t))$. Finally, by construction $t = \tilde{t}(\tilde{t}(t))$, and so $\tilde{t}$ induces a one-to-one mapping from T^i to T^j. Therefore we have,

$$\sum_{t \in T^i} Pr(t) \frac{1}{|W(\sigma(t))|} = \sum_{t \in T^i} Pr(\tilde{t}(t)) \frac{1}{|W(\sigma(\tilde{t}(t)))|} = \sum_{t \in T^j} Pr(t) \frac{1}{|W(\sigma(t))|}.$$

□

Proof for Lemma 5. In the event E^c, agent i wins the award with positive probability if and only if he has good evidence about himself. Then too, he shares it with equal odds with every other agent who provides good evidence. The probability of winning this way in a winning group of size $k+1$ is

$$\frac{[(1-\gamma) + \gamma(1-q)]^{n-k-1}[\gamma q]^{k+1}}{k+1}$$

and there are $\binom{n-1}{k}$ different $k+1$ sized winning groups the agent could be a part of. Therefore agent i's probability of winning the award and E^c being true , $P_i^{E^c}$, satisfies

$$P_i^{E^c} = \sum_{k=0}^{n-1} \binom{n-1}{k} \frac{[(1-\gamma)+\gamma(1-q)]^{n-k-1}[\gamma q]^{k+1}}{k+1},$$

which is independent of i.

□

Construction of beliefs and proof of consistency for Proposition 1

For greater expositional ease in the rest of this section, we abuse notation in the following way. We use $\sigma(A|B)$ to denote the probability with which the event A occurs conditional on the event B under the strategy profile σ. So for instance $\sigma(m|t) = \prod_{i \in N} \sigma_i(m_i|t)$.

Adversarial Disclosure

Let $M^\sigma = \{m \in M | \sigma(m|t) > 0, t \in T\}$. Recall that M is the set of all feasible message profiles. Denote the set of type profiles that can generate message profile m as $T(m) \equiv \{(\omega, e) \in T | m_{ij} \neq \emptyset \Rightarrow e_{ij} = \omega_j = m_{ij}\}$. Consider the following fully mixed strategy profile σ^ϵ. For all $t_i \in T_i$ and $i \in N$,

$\sigma^\epsilon(m_{ij}\|t_i)$	$i=j, e_{ij}=G$	$i=j, e_{ij}=B$	$i\neq j, e_{ij}=G$	$i\neq j, e_{ij}=B$
$m_{ij}=G$	$1-\epsilon$		ϵ	
$m_{ij}=\emptyset$	ϵ	$1-\epsilon$	$1-\epsilon$	ϵ
$m_{ij}=B$		ϵ		$1-\epsilon$

Notice that $M^{\sigma^\epsilon} = M$, in that all feasible messages are realized with positive probability. The resulting conditional belief function, derived using Bayes'

rule, and defined for all $m \in M$ is denoted by

$$\mu(t|m,\sigma^\epsilon) = \frac{\sigma^\epsilon(m|t)Pr(t)}{\sum\limits_{t' \in T(m)} \sigma^\epsilon(m|t')Pr(t')}.$$

Next,

$$\begin{aligned}\mu_i(t_i|m,\sigma^\epsilon) &= \sum_{(t_i,\cdot)\in T(m)} \mu(t|m,\sigma^\epsilon) \\ &= \frac{1}{\sum\limits_{t' \in T(m)} \sigma^\epsilon(m|t')Pr(t')} \cdot \sum_{(t_i,t''_{-i})\in T(m)} \sigma^\epsilon(m|(t_i,t''_{-i}))Pr(t_i,t''_{-i}).\end{aligned}$$

The resulting belief about player i's own-type is

$$\begin{aligned}\beta_i^\epsilon(m) &= \sum_{t\in T(m)|\omega_i=G} \mu_i(t_i|m,\sigma^\epsilon) \\ &= \frac{\sum\limits_{t' \in T(m)|\omega'_i=G} \sigma^\epsilon(m|t')Pr(t')}{\sum\limits_{t' \in T(m)|\omega'_i=G} \sigma^\epsilon(m|t')Pr(t') + \sum\limits_{t' \in T(m)|\omega'_i=B} \sigma^\epsilon(m|t')Pr(t')}.\end{aligned}$$

For any $t \in T$ and $i \in N$, define b_i^t and b_{-i}^t such that $b_i^t = (\omega_i, (e_{ji})_{j\in\tilde{N}_i})$ and $(b_i^t, b_{-i}^t) = t$. Under the strategy σ^ϵ, conditional on type profile t the probability with which a message m_{ij} is sent is independent of any other message. Further the prior probability of t can be decomposed as a product of the prior probabilities of b_i^t and b_{-i}^t. As a result,

$$\begin{aligned}\sigma^\epsilon(m|t)Pr(t) &= \prod_{j\in\tilde{N}_i} \sigma^\epsilon(m_{ji}|t) \prod_{l\neq i} \sigma^\epsilon(m_{kl}|t)Pr(t) \\ &= \prod_{j\in\tilde{N}_i} \sigma^\epsilon(m_{ji}|b_i^t)Pr(b_i^t) \prod_{l\neq i} \sigma^\epsilon(m_{kl}|b_{-i}^t)Pr(b_{-i}^t)\end{aligned}$$

Next note that if $(b_i^t, b_{-i}^t), (b_i^{\tilde{t}}, b_{-i}^{\tilde{t}}) \in T(m)$, then $(b_i^t, b_{-i}^{\tilde{t}}), (b_i^{\tilde{t}}, b_{-i}^t) \in T(m)$. As

a result, we get

$$\sum_{t' \in T(m)|\omega_i'=G} \sigma^\epsilon(m|t')Pr(t') = \sum_{b_i^t|t\in T(m),\omega_i=G} \prod_{j\in\tilde{N}_i} \sigma^\epsilon(m_{ji}|b_i^t)Pr(b_i^t) \cdot \sum_{b_{-i}^{\tilde{t}}|\tilde{t}\in T(m),\omega_i=G} \prod_{l\neq i} \sigma^\epsilon(m_{kl}|b_{-i}^{\tilde{t}})Pr(b_{-i}^{\tilde{t}}).$$

Further, for any $t', t'' \in T(m)$, we have $\{b_{-i}^t|(b_i^{t'}, b_{-i}^t) \in T(m)\} = \{b_{-i}^t|(b_i^{t''}, b_{-i}^t) \in T(m)\}$. Therefore we obtain

$$\beta_i^\epsilon(m) = \frac{\sum_{b_i^t|t\in T(m),\omega_i=G} \prod_{j\in\tilde{N}_i} \sigma^\epsilon(m_{ji}|b_i^t)Pr(b_i^t)}{\sum_{b_i^t|t\in T(m),\omega_i=G} \prod_{j\in\tilde{N}_i} \sigma^\epsilon(m_{ji}|b_i^t)Pr(b_i^t) + \sum_{b_i^t|t\in T(m)\&\omega_i=B} \prod_{j\in\tilde{N}_i} \sigma^\epsilon(m_{ji}|b_i^t)Pr(b_i^t)}.$$

Hence, for any m where $m_{ki} = \emptyset$ for all k we have

$$\beta_i^\epsilon(m) = \frac{\gamma[(1-q)+q\epsilon][(1-q)+q(1-\epsilon)]^{d_i}}{\gamma[(1-q)+q\epsilon][(1-q)+q(1-\epsilon)]^{d_i} + (1-\gamma)[(1-q)+q(1-\epsilon)][(1-q)+q\epsilon]^{d_i}}.$$

No-snitching

Consider the following fully mixed strategy profile, σ^ϵ. For all $t_i \in T_i$ and $i \in N$,

				$i \neq j, e_{ij} = B$	
$\sigma^\epsilon(m_{ij}\|t_i)$	$i = j, e_{ij} = G$	$i = j, e_{ij} = B$	$i \neq j, e_{ij} = G$	$\omega_i = G$	$\omega_i = B$
$m_{ij} = G$	$1-\epsilon$		ϵ		
$m_{ij} = \emptyset$	ϵ	$1-\epsilon$	$1-\epsilon$	$1-\epsilon^2$	$1-\epsilon$
$m_{ij} = B$		ϵ		ϵ^2	ϵ

Since the strategy profile is fully mixed we have $M^{\sigma^\epsilon} = M$. Again, set the conditional belief function derived using Bayes' rule as

$$\mu(t|m, \sigma^\epsilon) = \frac{\sigma^\epsilon(m|t)Pr(t)}{\sum_{t'\in T(m)} \sigma^\epsilon(m|t')Pr(t')}.$$

As shown earlier this leads to

$$\beta_i^{\epsilon}(m) = \frac{\sum\limits_{t' \in T(m)|\omega_i'=G} \sigma^{\epsilon}(m|t')Pr(t')}{\sum\limits_{t' \in T(m)|\omega_i'=G} \sigma^{\epsilon}(m|t')Pr(t') + \sum\limits_{t' \in T(m)|\omega_i'=B} \sigma^{\epsilon}(m|t')Pr(t')}.$$

For a given $i \in N$, consider the following mapping $\tilde{t}^i : T \to T$, with $\tilde{t}^i(\omega, e) = (\tilde{\omega}, \tilde{e})$ such that

$$\begin{aligned}
\tilde{\omega}_{-i} &= \omega_{-i}, \\
\tilde{e}_{jk} &= e_{jk}, \qquad \textit{for all } k \neq i, \\
\tilde{e}_{ji} &= B \qquad \Leftrightarrow e_{ji} = G, \\
\tilde{e}_{ji} &= G \qquad \Leftrightarrow e_{ji} = B, \\
\tilde{\omega}_i &= G \qquad \Leftrightarrow \omega_i = B, \\
\tilde{\omega}_i &= B \qquad \Leftrightarrow \omega_i = G.
\end{aligned}$$

Effectively, given a type profile t, $\tilde{t}^i(t)$ corresponds to the type profile that switches agent i's own-type, but leaves all other own-type and evidence realizations (modulo i's own-type) the same. It follows from the definition that $\tilde{t}^i$ is a bijection. Let $\tilde{t}^i_{-e_{ii}} : T_{-e_{ii}} \to T_{-e_{ii}}$ be defined as $\tilde{t}^i_{-e_{ii}}(t_{-e_{ii}}) = (\tilde{t}^i(t_{-e_{ii}}, e_{ii}))_{-e_{ii}}$. The function $\tilde{t}^i_{-e_{ii}}$ is also a bijection that is well defined since for any given t the vector $(\tilde{t}^i(t))_{-e_{ii}}$ does not depend on the realized value of e_{ii}. Let $M^{i\emptyset} = \{m \in M | m_{ji} = \emptyset, \forall j \in \tilde{N}_i\}$. For any $m \in M^{i\emptyset}$, $\tilde{t}^i$and $\tilde{t}^i_{-e_{ii}}$ remain bijections on $T(m)$ and $T_{-e_{ii}}(m)$. Importantly, $Pr(t_{-e_{ii}}|\omega_i = G) = Pr(\tilde{t}^i_{-e_{ii}}(t_{-e_{ii}})|\omega_i = B)$.

Since own-type realizations are independent and conditional on an own-type realization, evidence realizations are independent too, the specification of σ^{ϵ} yields the following. For any $m \in M^{i\emptyset}$,

$$\begin{aligned}
\sum_{t \in T(m)|\omega_i=G} \sigma^{\epsilon}(m|t)Pr(t) = &\sum_{e_{ii} \in \{\emptyset, G\}} \sigma^{\epsilon}(m_{ii} = \emptyset | e_{ii}) Pr(e_{ii}|\omega_i = G) \\
&Pr(\omega_i = G) \sum_{t_{-e_{ii}} | t \in T(m), \omega_i = G} \sigma^{\epsilon}(m_{-ii}|t_{-e_{ii}}) Pr(t_{-e_{ii}}|\omega_i = G).
\end{aligned}$$

We then obtain

$$\beta_i^\epsilon(m) = \frac{\gamma[(1-q)+q\epsilon]A^\epsilon(m)}{\gamma[(1-q)+q\epsilon]A^\epsilon(m) + (1-\gamma)(1-q+q(1-\epsilon))B^\epsilon(m)}.$$

where

$$A^\epsilon(m) = \sum_{t_{-e_{ii}}|t\in T(m),\omega_i=G} \sigma^\epsilon(m_{-ii}|t_{-e_{ii}})Pr(t_{-e_{ii}}|\omega_i = G)$$

and

$$B^\epsilon(m) = \sum_{t_{-e_{ii}}|t\in T(m),\omega_i=G} \sigma^\epsilon(m_{-ii}|\tilde{t}^i_{-e_{ii}}(t_{-e_{ii}}))Pr(\tilde{t}^i_{-e_{ii}}(t_{-e_{ii}})|\omega_i = B).$$

Fix $i \in N$, $m \in M^{i\emptyset}$ and some $t \in T(m)$ such that $\omega_i = G$. Given the specification of σ^ϵ, the following holds.

$$\sigma^\epsilon(m_{kl}|t_{-e_{ii}}) = \sigma^\epsilon(m_{kl}|\tilde{t}^i_{-e_{ii}}(t_{-e_{ii}})) \qquad \forall k,l \neq i$$

$$\left.\begin{array}{c} \sigma^\epsilon(m_{kl}|t_{-e_{ii}}) = 1 \Rightarrow \sigma^\epsilon(m_{kl}|\tilde{t}^i_{-e_{ii}}(t_{-e_{ii}})) = 1 \\ \sigma^\epsilon(m_{kl}|t_{-e_{ii}}) \in \{1-\epsilon^2, 1-\epsilon\} \Rightarrow \sigma^\epsilon(m_{kl}|\tilde{t}^i_{-e_{ii}}(t_{-e_{ii}})) \in \{1-\epsilon^2, 1-\epsilon\} \\ \sigma^\epsilon(m_{kl}|t_{-e_{ii}}) \in \{\epsilon^2, \epsilon\} \Rightarrow \sigma^\epsilon(m_{kl}|\tilde{t}^i_{-e_{ii}}(t_{-e_{ii}})) \in \{\epsilon^2, \epsilon\} \end{array}\right\} \textit{ for } k = i, l \in N_i \textit{ or } l = i, k$$

Therefore for all $k,l \in N$, $\lim_{\epsilon\to 0}\sigma^\epsilon(m_{kl}|t_{-e_{ii}}) = \lim_{\epsilon\to 0}\sigma^\epsilon(m_{kl}|\tilde{t}^i_{-e_{ii}}(t_{-e_{ii}}))$. As a result, for such an $m \in M^{i\emptyset}$ we obtain $\lim_{\epsilon\to 0} A^\epsilon(m) = \lim_{\epsilon\to 0} B^\epsilon(m)$. This common limit is a non-zero finite number if m is additionally such that $m_{kl} \neq B$ for all $k,l \in N$ and $m_{kl} \neq G$ for all $k \neq l$. For all such $m \in M^{i\emptyset}$ we get

$$\lim_{\epsilon\to 0}\beta_i^\epsilon(m) = \frac{\gamma(1-q)}{\gamma(1-q)+(1-\gamma)}.$$

Let $M^{i\emptyset\emptyset} = \{m \in M^{i\emptyset} | m_{kk} \neq B, m_{kl} = \emptyset, \forall k \neq i \textit{ and } l \in \tilde{N}_k\}$ be the set of messages in which the only non-empty messages, if any, are those from player i about some neighbour of his, or agents, other than i, revealing good evidence about their own selves. For any $m \in M^{i\emptyset\emptyset}$, let $\hat{N}_i^G(m) = \{j \in N_i | m_{ij} = G\}$ and $\hat{N}_i^B(m) = \{j \in N_i | m_{ij} = B\}$. Let $e_{i*} = (e_{ii}, (e_{ij})_{j\in\hat{N}_i^B\cup\hat{N}_i^G})$. Then, for any

such $m \in M^{i\emptyset\emptyset}$ we obtain

$$\sum_{t\in T(m)|\omega_i=G} \sigma^\epsilon(m|t)Pr(t) = \left(\sum_{e_{ii}\in\{\emptyset,G\}} \sigma^\epsilon(m_{ii}=\emptyset|e_{ii})Pr(e_{ii}|\omega_i=G)\right)\epsilon^{|\hat{N}_i^G(m)|}\epsilon^{2|\hat{N}_i^B(m)|}$$
$$Pr(\omega_i=G)\sum_{t_{-e_{ii}}|t\in T(m),\omega_i=G} \sigma^\epsilon(m_{-e_{i*}}|t_{-e_{ii}})Pr(t_{-e_{ii}}|\omega_i=G)$$

and

$$\sum_{t\in T(m)|\omega_i=B} \sigma^\epsilon(m|t)Pr(t) = \left(\sum_{e_{ii}\in\{\emptyset,B\}} \sigma^\epsilon(m_{ii}=\emptyset|e_{ii})Pr(e_{ii}|\omega_i=B)\right)\epsilon^{|\hat{N}_i^G(m)|}\epsilon^{|\hat{N}_i^B(m)|}$$
$$Pr(\omega_i=B)\sum_{t_{-e_{ii}}|t\in T(m)\omega_i=G} \sigma^\epsilon(m_{-e_{i*}}|\tilde{t}^i_{-e_{ii}}(t_{-e_{ii}}))Pr(\tilde{t}^i_{-e_{ii}}(t_{-e_{ii}})|\omega_i=B).$$

Therefore

$$\beta_i^\epsilon(m) = \frac{\epsilon^{|\hat{N}_i^G(m)|}\epsilon^{2|\hat{N}_i^B(m)|}C^\epsilon(m)}{\epsilon^{|\hat{N}_i^G(m)|}\epsilon^{2|\hat{N}_i^B(m)|}C^\epsilon(m)+\epsilon^{|\hat{N}_i^G(m)|}\epsilon^{|\hat{N}_i^B(m)|}D^\epsilon(m)}$$

where

$$C^\epsilon(m) = \sum_{e_{ii}\in\{\emptyset,G\}} \sigma^\epsilon(m_{ii}=\emptyset|e_{ii})Pr(e_{ii}|\omega_i=G)$$
$$Pr(\omega_i=G)\sum_{t_{-e_{ii}}|t\in T(m),\omega_i=G} \sigma^\epsilon(m_{-e_{i*}}|t_{-e_{ii}})Pr(t_{-e_{ii}}|\omega_i=G)$$

and

$$D^\epsilon(m) = \sum_{e_{ii}\in\{\emptyset,B\}} \sigma^\epsilon(m_{ii}=\emptyset|e_{ii})Pr(e_{ii}|\omega_i=B)$$
$$Pr(\omega_i=B)\sum_{t_{-e_{ii}}|t\in T(m)\omega_i=G} \sigma^\epsilon(m_{-e_{i*}}|\tilde{t}^i_{-e_{ii}}(t_{-e_{ii}}))Pr(\tilde{t}^i_{-e_{ii}}(t_{-e_{ii}})|\omega_i=B).$$

Since $C^\epsilon(m)$ and $D^\epsilon(m)$ are both bounded away from 0 we obtain

$$\beta_i^\epsilon(m) = \frac{\epsilon^{|\hat{N}_i^B(m)|} C^\epsilon(m)}{\epsilon^{|\hat{N}_i^B(m)|} C^\epsilon(m) + D^\epsilon(m)}.$$

It follows that for all such $m \in M^{i\emptyset\emptyset}$,

$$\lim_{\epsilon \to 0} \beta_i^\epsilon(m) = 0.$$

References

Arrow, K. J. (1973). The theory of discrimination. In *Discrimination in Labor Markets*, edited by Orley C. Ashenfelter and Albert Everett Rees, 3-33. Princeton, NJ: Princeton University Press.

Baumann, L. (2018). Self-ratings and peer review. *Working paper*.

Becker, G.S. (1971). *The Economics of Discrimination*. Chicago: University of Chicago Press.

Ben-Porath, E., Dekel, E., & Lipman, B. L. (2018). Disclosure and choice. *The Review of Economic Studies*, 85(3), 1471-1501.

Ben-Porath, E., Dekel, E., & Lipman, B. L. (2019). Mechanisms with evidence: Commitment and robustness. *Econometrica*, 87(2), 529-566.

Blackwell, D. (1951). Comparisons of experiments. *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, 1, 93-102.

Blackwell, D. (1953). Equivalent comparisons of experiments. *The Annals of Mathematical Statistics*, 265-272.

Bloch, F., & Olckers, M. (2022). Friend-based rankings. *American Economic Journal: Microeconomics*, 14(2), 176-214.

Bloch, F., & Olckers, M. (2021). Friend-based ranking in practice. *AEA Papers and Proceedings*, 111, 567-71.

Calvo-Armengol, A., & Jackson, M. O. (2004). The effects of social networks on employment and inequality. *American Economic Review*, 94(3), 426-454.

Coate, S., & Loury, G.C. (1993). Will affirmative-action policies eliminate negative stereotypes? *American Economic Review*, 83(5), 1220-1240.

Dye, R. A. (1985). Disclosure of nonproprietary information. *Journal of Accounting Research*, 23, 123-145.

Eeckhout, J. (2006). Minorities and endogenous segregation. *Review of Economic Studies*, 73(1), 31-53.

Grossman, S. J. (1981). The informational role of warranties and private disclosure about product quality. *The Journal of Law and Economics*, 24(3), 461-483.

Grossman, S., & Hart, O. (1980). Disclosure laws and takeover bids. *The Journal of Finance*, 35(2), 323-334.

Hart, S., Kremer, I., & Perry, M. (2017). Evidence games: Truth and commitment. *American Economic Review*, 107(3), 690-713.

Jung, W. O., & Kwon, Y. K. (1988). Disclosure when the market is unsure of information endowment of managers. *Journal of Accounting Research*, 146-153.

Milgrom, P. R. (1981). Good news and bad news: Representation theorems and applications. *The Bell Journal of Economics*, 380-391.

Montgomery, J. D. (1991). Social networks and labor-market outcomes: Toward an economic analysis. *American Economic Review*, 81(5), 1408-1418.

National Legislative Bodies / National Authorities, Canadian Human Rights Act, 1985, R.S.C., 1985, c. H-6, available at: https://www.refworld.org/docid/5417ff844.html

Onuchic, P., ⓡ Ray, D. (forthcoming). Signaling and discrimination in collaborative projects. *American Economic Review*.

Onuchic, P. (2022). Recent contributions to theories of discrimination. *working paper*.

Peski, M., & Szentes, B. (2013). Spontaneous discrimination. *American Economic Review*, 103(6), 2412-2436.

Small, L. M., & Pager, D. (2020). Sociological perspectives on racial discrimination. *Journal of Economic Perspectives*, 34(2), 49-67.