Labor Associations: The Blue Wall of Silence☆

David K. Levine[1], Andrea Mattozzi[2], Salvatore Modica[3]

Abstract

We develop a model showing that when labor demand is inelastic and individual behavior is easily monitored a firm's employees may prefer to protect its shirkers. By optimally reducing overall effort and increasing wages for all, a labor association rationally uses its monopoly power as described in the left wing labor slogan "work less so that all may work." In addition, employees have a strong incentive to conceal information about peers' performance from firms, what has been infamously known as the *blue wall of silence* in the case of the police. We argue that a number of recently proposed remedies to this problem are unlikely to succeed and suggest a more promising alternative: increase competition.

Keywords: Labor Associations, Monitoring Costs, Self Organizing Groups.

☆First Version: March 7, 2021. We are especially grateful to Jean-Pierre Benoit, Alvaro Delgado, Juan Dubra, Gilat Levy, Gerd Muehlheusser, Pietro Ortoleva, Rocco Schiavone and participants at the Madrid Political Economy Workshop, the Beijing International Workshop on Microeconomics, the Essex Theory Workshop, the UCL Theory Workshop, and the Di Tella Economics Seminar. We gratefully acknowledge support from the MIUR PRIN 2017 n. prj-0317_d1.

Email addresses: `david@dklevine.com` (David K. Levine), `andrea.mattozzi@eui.eu` (Andrea Mattozzi), `salvatore.modica@unipa.it` (Salvatore Modica)

[1]Department of Economics and RSCAS European University Institute and Department of Economics WUSTL

[2]Department of Economics European University Institute and Università di Bologna

[3]Università di Palermo, Italy, Dipartimento di Matematica e Informatica

1. Introduction

> *Baltimore cop, stripped of police powers after fatally shooting unarmed teen, kept on payroll for 28 years.*
>
> Baltimore Brew August 12, 2021

Why do employees often protect shirkers even when they clearly are not a majority, and what can be done about it? An obvious context evoked by the news article mentioned in the incipit is that of the police in the USA, but teachers are often blamed for similar practices.[4] Our explanation for this apparent puzzle is that when demand is inelastic effort reduction arises from the rational use of monopoly power by employees.[5] From this directly follows that shirkers should be protected and information should not be shared with employers - what has been infamously known as the *blue wall of silence* in the case of the police.

We examine a setting in which there are two types of employees: workers and shirkers. These types are private information to the employees. Workers prefer to provide effort, perhaps because they get satisfaction out of a job well done; shirkers prefer not to provide effort and only employees observe signals about their peers' effort. This is consistent with a long literature showing that peer evaluation is more accurate than supervisor evaluation,[6] and is particularly important in jobs such as teaching or policing where supervisors do not observe performance in the classroom or field. To rule out the much studied issue of screening by firms we assume that types are determined *ex post* after employment. This corresponds to a situation in which it takes time on the job to learn how satisfying the work is (which seems appropriate both for policing and teaching). If labor demand is inelastic we show that it is advantageous to

[4] The Mollen Commission (1994) report documents police covering up for the misbehavior of other police. Concerning teachers Moe (2011) reports *"[New York] city's Rubber Rooms – Temporary Reassignment Centers – where teachers were housed when they were considered so unsuited to teaching that they needed to be kept out of the classroom, away from the city's children. [...] They got paid a full salary. They received full benefits, as well as all the usual vacation days, and they had their summers off. Just like real teachers. Except they didn't teach.*"

[5] In the case of the police, as we discuss subsequently, increased misbehavior can be viewed as a negative effort that lowers the average.

[6] See, for example, Kraut (1975).

employees to allow shirkers to shirk, deny the firm information about employee effort, and if it is inelastic enough it is advantageous to restrain workers from working too hard. In the opposite case in which labor demand is elastic it is advantageous for employees to encourage effort and share information with the firm. Our results follow from the fact that all employees agree it would be best to maximize overall employment. Then intuitively, when labor demand is inelastic firms need a certain amount of labor input "no matter what," and by reducing individual effort and hence by tolerating shirking, a labor association forces firms to hire more employees. Conversely, if demand is elastic reducing effort simply reduces the market wage.[7]

Our theory says that labor associations will protect their weakest members when demand for labor is inelastic. Given that the protection for weakest members is common we must ask if it is the case that the demand for labor is often inelastic. Indeed, we argue that the empirical evidence favors inelasticity across industries and countries. This may help explain why we so often see the protection of weak members.

In this case, where the optimal labor association plan calls for relatively low overall effort, we shed light on a specific aspect of the trade-off workers face in participating in the association. Specifically, we compare how much utility a worker gets under the optimal association plan versus how much they would get if there were no labor association and the firm observes effort. In the latter case workers receive a premium because shirkers must be compensated for their effort. When the required compensation is relatively low, however, the premium will be less than the increased wage they will receive if the association's effort reduction plan is in place.

We next ask what happens if employees cannot prevent the leakage of in-

[7]We should indicate that most occupations do have a limited form of performance-based pay in the form of a probationary period during which or at the end of which the employee can be laid off without cost to the firm. Our focus is on post probationary incentives: if shirkers learn their type during probation we would expect them to conceal it so that those laid off would most likely be types incapable of providing effort. It is also the case that wages can be the outcome of a bargaining process between firms and unions. Although wekeep the model simple by abstracting from this we later discuss the role of bargaining.

formation to the firm. Here we show that never-the-less improved information will help employees but harm rather than benefit consumers. To further highlight the role of information we will also consider the case where it is the firm rather than labor association that receives information and provides incentives for effort. Although better information in this case will help consumers, the association will refuse to provide any additional information to the firm and, if it is able to do so, attempt to degrade information flowing to the firm.

In order to understand under what conditions labor associations can successfully restrict effort of their members in the interest of the group we need a theory of how they provide incentives. We know from the work of Ostrom (1990) and her successors how this can be achieved: groups can self-organize to overcome the free rider problem and provide public goods (such as restraining effort) through peer monitoring and social punishments such as ostracism. Formal theories of this type originate in the work of Kandori (1992) on repeated games with many players and have been specialized to the study of organizations. The basic idea is that groups choose norms consisting of a target behavior for the group members and individual penalties for failing to meet the target; these norms are endogenously chosen in order to advance group interests. Specifically the group designs a mechanism to promote group interests subject to incentive constraints for individual group members, and it provides incentives in the form of punishments for group members who fail to adhere to the norm.[8]

In this paper we build on this theory and show that the optimal target level of average effort in an industry crucially depends on the elasticity of labor demand and on how difficult it is for the association to monitor individual behavior. While elasticity of demand determines whether it is optimal to restrain or incentivize effort, monitoring difficulty, which in turn depends on the social network structure of employees, determines whether it is possible to do so or not. Both elements are therefore necessary for effort quotas to emerge in equilibrium. We show, moreover, that similar considerations apply not only to labor associations but to individual proprietors who sell into the market at

[8] See for example Levine and Modica (2016) and Dutta, Levine and Modica (2021).

a piece rate: we argue that country squires should be "lazy" because they face inelastic output demand and industrialists "energetic" because they face elastic demand.

We are not the first to ask why labor associations protect their weakest members. Our explanation complements existing theories which focus on particular details of the punishment or production process. Benoit and Dubra (2004) focus on testimony before an imperfect court and show that this can lead employees to vote for a wall of silence. However, this does not explain as does our model why the blue wall of silence also applies outside of a court setting: for example to reporting non-criminal wrong-doing to supervisors. A second theory is that of Muehlheusser and Roider (2008) who focus on team production and emphasize the need for cooperation in such a setting.[9] This also does not apply to all the settings we consider: for example, teaching is not generally considered to be team production.

2. The Model

We consider the derived demand for labor in a particular industry. We denote by n the size of the labor force and let $\overline{e} \in [0, 1]$ represents average employee effort: total labor input given as $x = n\overline{e}$. We suppose that the social value of labor input is $U(x)$ (consumer surplus[10]), assumed to be smooth and strictly differentiably concave up to a satiation level X. Hence the marginal value $U'(x)$ is positive and declining with input for $x < X$. We assume that the industry takes average effort $\overline{e}$ and the expected wage W as given. The average effort level is taken as given because it is due to a prior commitment by a labor association as indicated below. There are two reasons the industry may take the wage as given. First, the industry may be competitive. Second, the industry may be a benevolent government monopsonist. Because it is benevolent it cares

[9]There is also a literature about the related problem of whistle-blowing. A recent experimental study Mechtenberg, Muehleusser and Roider (2020) finds troubling experimental evidence that legal protection does not improve organizational incentives. We have similar theoretical findings in our model for monitoring.

[10]This is a partial equilibrium model in which we assume that there are no income effects, or a general equilibrium model with transferable utility.

about worker welfare as well as consumer welfare, so behaves "as if" it were a competitive industry.

As the marginal value of an additional worker to the industry is $\overline{e}U'(n\overline{e})$, from the demand side the expected wage paid is $W = \overline{e}U'(n\overline{e})$. We define revenue to be quantity times price along the labor demand curve: $R(x) = xU'(x)$. We assume this is concave, that is, marginal revenue is declining with labor input.

On the supply side, we assume that the size of the labor force is determined by the opportunity cost of the marginal employee $v(n)$. This we assume to be smooth and strictly differentiably increasing so that $v'(n) > 0$. After employment each employee i provides effort $e^i \in [0, 1]$.

There are two types of employees, *workers* w and *shirkers* s. As indicated, types are realized *ex-post* after employment and before effort provision and they are private information. The exogenous probability of an employee being of type w is γ. For a shirker there is disutility from effort $\mu^s e^i$ while a worker receives disutility from lack of effort $\mu^w(1 - e^i)$ with μ^s and μ^w both positive numbers. In other words workers prefer to provide effort, perhaps because they get satisfaction out of a job well done, while for shirkers effort is costly.[11]

If workers work and shirkers shirk the average level of effort is γ. We refer to γ as the *natural level of effort.* If effort is not at the natural level there will be costs of providing or not providing effort, as well as costs related to the incentives needed to get workers to provide that level of effort. Denote those costs measured in worker utility by C. On the supply side since type is only realized after employment it must be $W - C \leq v(n)$ with equality if $n > 0$. In other words, market clearing is given by $\overline{e}U'(n\overline{e}) - C \leq v(n)$, with equality if $n > 0$. Because $U(x)$ is strictly concave and $v(n)$ is strictly increasing this has a unique solution, denoted by $\hat{n}(\overline{e}, C)$. If average effort is γ no incentives are needed that is $C = 0$. We assume that $\gamma U'(0) > v(0)$ so that $\hat{n}(\gamma, 0) > 0$ is uniquely defined by $\gamma U'(\gamma\hat{n}(\gamma, 0)) = v(\hat{n}(\gamma, 0))$. It will be convenient to define $x^\gamma = \gamma\hat{n}(\gamma, 0)$, which is the market clearing labor input when workers work and

[11]As subsequently discussed we regard misbehavior, such as police committing crimes rather than preventing them, as a kind of negative effort. We assume, however, that even shirkers do not on average wish to provide negative effort.

shirkers shirk; we call x^γ the *natural level of labor input.*

We now consider specifically how incentives are provided. Prior to the employment and the realization of types there is a collective *association* of employees. It can commit to setting an effort quota and will then be able to observe a noisy signal of whether the quota was adhered to by individual members. We consider the two alternative cases of a minimum and a maximum quota on effort - subsequently we show that no mechanism can do better than these. Denote by ϕ^- a minimum quota on effort meaning that only effort levels $e \in [\phi^-, 1]$ are acceptable, and by ϕ^+ a maximum quota meaning that only effort levels $e \in [0, \phi^+]$ are acceptable. While individual efforts are not observable, each employee - conditional on her effort choice - produces a signal observed by the association but not the industry $z^i \in \{0, 1\}$ where 0 is a signal of lack of effort and 1 is a signal of effort. If the quota was adhered to the probability of getting a signal that indicates it was violated is π while if the quota was violated the probability is $\pi' > \pi$. Hence if a minimum quota is adhered to the probability of getting a signal indicating effort (the quota is adhered to) is π, while if a maximum quota is adhered to the probability of getting a signal indicating effort (the quota was violated) is π'. If an employee has a signal indicating a quota violation an endogenous utility punishment P is issued.[12]

In addition to providing their own incentives the association may reveal signals to the industry so that firms may then offer a wage schedule $W(0), W(1)$. We assume that firms cannot impose unlimited wage penalties, but face a *feasibility constraint* in how much they can penalize signals of lack of effort. Specifically if the *ex ante* expected industry wage is W we assume[13] an upper bound $W(1) - W(0) \leq \overline{\Delta}(W)$ where $\overline{\Delta}(W) > 0$ is smooth and non-decreasing in W. For example, if firms are limited to non-negative wages so that $W(0) \geq 0$ and if $W = \pi W(0) + (1 - \pi)W(1)$then $\overline{\Delta}(W) = W/(1 - \pi)$.

We can now formulate the problem of the association as a mechanism design

[12] Note that as types are private information the punishment depends only on the signal, not on the type.

[13] The upper bound could be different for a minimum and maximum quota, but as it matters only for the minimum quota we do not introduce extra notation.

problem. It chooses a quota ϕ , a punishment P and we may think of it also as choosing the number of employees n and wage schedule $W(0), W(1)$. It must do so, however, in an incentive compatible way for both employees and firms and so that the market clears.

First, P must be chosen sufficiently high that employees are willing to honor the quotas. In this case for a minimum quota workers provide full effort and shirkers provide effort to quota, while with a maximum quota workers provide effort to quota and shirkers provide no effort. This determines the expected effort $\overline{e}$ and the expected wage W. It also determines the cost of providing incentives, which has two components: a direct cost D and a monitoring cost M, with overall cost $C = D + M$. The direct cost for a minimum quota is $D = (1-\gamma)\mu^s\phi^-$ and for a maximum quota $D = \gamma\mu^w(1-\phi^+)$, that is the disutility of effort provided or not provided. The monitoring cost for a quota is $M = \pi P$, that is, the expected cost of punishment "on the equilibrium path" when everyone adheres to the quota. Second, markets must clear: $n = \hat{n}(\overline{e}, C)$.

Third, if the association chooses not to reveal the signals to the firm then $W = W(0) = W(1)$. If it does choose to reveal the signals to the firm then the wage schedule must be feasible, and it must not be the case a firm can choose an alternative feasible wage schedule and incentive compatible effort levels for the two types that promises at least as much utility to workers *ex ante* as $W - C$ and yields a strictly positive profit.[14]

Finally, we say an incentive compatible mechanism is *employee dominated* if there is another incentive compatible mechanism that makes no potential employee[15] worse off and makes some potential employees better off. Our solution concept is that the mechanism chosen by the association must be incentive compatible and not employee dominated, that is, *employee efficient* in the class of

[14]There is one technical proviso when the association reveals the signal. When there is a maximum quota a high wage $W(1)$ for a signal of high effort provides an incentive for shirkers to exceed the quota by ϵ so as to increase the probability of getting that wage from π to the higher probability π'. This leads to an uninteresting existence problem. Following Simon and Zame (1990) we introduce an endogenous tie-breaking rule, and assert that if a type can violate a quota by ϵ then, if they exactly meet the quota, they can choose between π and π'.

[15]That is, regardless of whether they are employed in the industry.

incentive compatible mechanisms.

3. The Blue Wall of Silence

Define monitoring difficulty as $\theta = \pi/(\pi' - \pi)$. Recall that γ is the probability of a worker and define marginal incentive costs $c^s \equiv \mu^s(1-\gamma+\theta)/(1-\gamma)$ and $c^w \equiv \mu^w(\gamma+\theta)/\gamma$. Recall that $R(x) = xU'(x)$ and that x^γ is the natural level of labor input. As it is useful to think of the results in terms of elasticity, recall that marginal revenue is

$$R'(x) = U'(x) + xU''(x) = \frac{1}{xU''(x)}\left(\frac{U'(x)}{xU''(x)} + 1\right) = \frac{1}{xU''(x)}\left(\frac{d\log x}{d\log U'(x)} + 1\right)$$

so that $R'(x) < 0$ when the absolute elasticity $|d\log x/d\log U'(x)|$ is less than 1 (inelastic) and conversely. In addition we have assumed that marginal revenue $R'(x)$ declines along the demand curve. If, as we ordinarily expect, absolute elasticity rises along the demand curve then marginal revenue falls as absolute elasticity rises. Our main result from Propositions 1 and 2 in the Appendix is then the following.

Theorem 1. *An employee efficient mechanism exists; there is a unique average effort $\hat{e}$ and labor input $\hat{x}$ across all incentive compatible employee efficient mechanisms; and there are four regimes:*

- *(inelastic) If $R'(x^\gamma) < -c^w$ then the association does not reveal the signal, a maximum quota is optimal, $\hat{e} < \gamma$, and $\hat{x} > x^\gamma$.*
- *(slightly inelastic) If $-c^w < R'(x^\gamma) < 0$ then there is no quota and $\hat{e} = \gamma$.*
- *(slightly elastic) If $0 \leq R'(x^\gamma) < c^s$ then there is an incentive compatible employee efficient mechanism in which the association uses a minimum quota and reveals the signal: $\hat{e} \geq \gamma$, and $\hat{x} \geq x^\gamma$.*
- *(elastic) If $R'(x^\gamma) > c^s$ then the association reveals the signal, a minimum quota is optimal, and $\hat{e} > \gamma$, $\hat{x} > x^\gamma$.*

The intuition behind the above theorem can be understood from two ideas (a formal proof is in the Appendix). First, there is no conflict of interest among employees. Precisely: an incentive compatible mechanism A employee dominates another B, that is, it makes all employees weakly better off, if and only if the level of employment n_A at A is strictly greater than the level of employment n_B at B. This is because the amount received by an employee $W - C$ is given by the labor supply curve $v(n)$. Since $n_A > n_B$ and the supply curve is upwards sloping $W_A - C_A = v(n_A) > v(n_B) = W_B - C_B$, so everyone employed under A is better off than under B, and those employed under neither mechanism are indifferent. This means that an employee efficient mechanism maximizes employment.

Second, given that the goal of the labor association is to maximize employment, the association should ask: how does increased effort translate into employment? The inverse demand curve for employment is given by $\overline{e}U'(n\overline{e})$. This is increasing in $\overline{e}$ for $R'(n\overline{e}) > 0$ and decreasing in $\overline{e}$ for $R'(n\overline{e}) < 0$. To see this simply observe that $R(n\overline{e}) = n\left[\overline{e}U'(n\overline{e})\right]$ so that the partial derivatives of $\overline{e}U'(n\overline{e})$ and $R(n\overline{e})$ with respect to $\overline{e}$ have the same sign.

We see then that in the inelastic case $R'(x^\gamma) < 0$ the association prefers to lower effort below the natural level and in the elastic case $R'(x^\gamma) > 0$ the association would like to raise effort above the natural level: both increase the demand for employment. Firms by contrast always would like greater effort for any given wage.

In the inelastic case the incentives of the association and firm are opposed: the association want to provide incentives to reduce effort and the firm incentives to increase it. To reduce effort the association uses a maximum quota. By providing the signal to the firm the association would enable it to provide incentives to increase effort and this makes it more costly (higher C) for the association to reduce effort, so the association prefers not to reveal the signal.

In the elastic case the incentives of the association and firm are aligned: both want to increase effort. Hence the association uses a minimum quota. However, the firm technology for providing incentives is better than that of the association because it replaces costly punishments with efficient transfer payments (what is lost by those with bad signals is gained by those with good signals). Hence the association wants to provide the signal to the firm so it can implement an efficient incentive scheme, reducing the need for costly punishments and lowering C.

Note that as a matter of practice the labor association may effectively suppress the signal by requiring firms to set non-contingent wages as is often done in union contracts where wages must be based only on seniority and not on signals of job performance.

Discussion of the Assumptions

There are two key assumptions underlying the model: that the industry is competitive and that information on effort is available to peers but not to firms. Clearly in some industries information on effort is available to firms: for example, in manufacturing there are supervisors overseeing the assembly line, administrative assistants in law firms work directly for their employer, and so forth. In other industries employees are separated from their supervisors: police work in the field, teachers in their own classroom, miners are often dispersed throughout a mine and so forth. In these latter cases our model applies, in the former it does not - we discuss in more detail the consequences of "strong" firms that observe employee effort in Section 4 below.

With respect to competition, clearly not all private sector industries are competitive. We are familiar with shirking and the blue wall of silence in the public sector: whether by police or by teachers as discussed in the introduction. Some of us have been in the Italian post office where there is a large queue of customers and one employee working at a window while three others drink coffee and chat. We have also observed exactly the same at rental car counters in the US - which is the private rather than public sector. In supermarkets when we try to find help employees often avoid us. Cable TV employees are notorious for their lack of effort, and so forth. Indeed, the term featherbedding was coined to primarily to describe behavior in the private sector. Nor is there evidence that employees report each other for lack of effort in any of these industries.

Turning to the public sector: it may be argued that a large city such as New York has some monopsony power over the hiring of employees such as police - although we observe New York police frequently leave to other cities. Certainly the large national police forces in Europe have monopsony power. However, as we indicated, this does not matter if a benevolent government cares about employees: such a government would act "as if" competitive to maximize the combined welfare of consumers and employees. As employees are also voters, and indeed labor associations are politically active, they have a strong incentive to balance the interests of consumers with those of employees. Hence we think it reasonable to view government as choosing not to exploit their monopsony power.

There is also the issue of wage bargaining by trade unions. However, there is evidence that many unions are unsuccessful at wage bargaining. For example, if police and teachers were effective at getting above market wages we would expect queues of applicants for those jobs: instead we read about teacher shortages and police departments struggling to recruit and retain officers.[16] By contrast in other industries such as longshoremen we observe queues waiting for admission to the union and our model does not apply.

Trade Unions

Trade unions in our view are distinct from labor associations. Unions engage in collective bargaining. Labor associations engage in informal peer pressure to prevent effort and enforce a blue wall of silence. For example, while the New York police have been unionized for over a century we believe that the effort of police officers to set up Frank Serpico for assassination for having violated the blue wall of silence would have taken place independent of the existence of the union, and indeed there is no evidence the union played any role in the assassination attempt. Turning to the private sector, we observe that while rental car employees are unionized, cable television employees are not. In the supermarket sector Walmart is not unionized while Kroger and Albertsons are: yet surveys of consumer satisfaction do not indicate that Walmart employees are better serving customers.[17]

While we can have labor associations without a trade union, trade unions almost invariably contain a labor association. Indeed, while unions do not appear

[16] https://thehill.com/changing-america/enrichment/education/3276034-heres-whats-driving-the-nationwide-teacher-shortage/; https://edition.cnn.com/2022/02/02/us/police-departments-struggle-recruit-retain-officers/index.html

[17] https://www.supermarketnews.com/consumer-trends/supermarkets-hold-line-customer-satisfaction

that successful in increasing wage, they have had greater success bargaining over working conditions: and they bargain for terms that enhance the efficacy of their labor association role. In some cases unions successfully negotiate rules that explicitly allow for shirking (only electricians can change a light bulb, so everyone else has to stand idle waiting for the electrician), and that prevent incentive pay, such as allowing pay differences based only on occupation and seniority. We refer here particular to the work of Galdon-Sanchez and Schmitz Jr (2002) and Schmitz (2005) who document how work rules negotiated by unions lower effort and how greater elasticity of demand (in the form of competition) leads to their abandonment. This is well understood through the lenses of our theory.

Does the Model Apply to Police?

For the most part effort is fairly obvious: teachers either teach or they do not, rental counter employees either serve customers or they do not. In the case of the police sitting and drinking coffee and eating donuts constitutes lack of effort. However, we would like to argue that more serious matters such as corruption and violence against innocent individuals constitute lack of effort. Here we take the role of police to be to reduce crime, and consider the effort of an individual police officer to be measured by how much crime they decrease. Drinking coffee and eating donuts does not reduce crime. Taking payoffs is more complicated, since in some instances an agreement may be reached about allowable crimes ("I look the other way on your drug deals if you don't shoot people") so it is ambigous as to whether "crime" increases or decreases. Regardless, taking bribes is less costly for the police than pursuing criminals, so in this sense can be viewed as shirking.

Finally, consider police violence. If the police shoot an armed gunman they prevent a crime. If they shoot an innocent person this may not be legally a crime, but from the point of view of the individual and society an innocent person being shot by the police is no better than if they are shot by a criminal. We would all agree that a police force whose only activity was violence against the innocent, legal or not, was not doing its job. Hence, in assessing police effort we must balance the legitimate use of violence which represents a reduction of crime against the illegitimate use of violence which in a social if not legal sense adds to crime. In other words the illegitimate use of violence represents a decrease in effort, while the legitimate use of violence represents an increase in effort.

The bottom line here is that while the model is not a perfect fit with police violence, it is true that shirking in the form of police violence reduces the provision of police services, and it that reduction in services that is the key element of our analysis.

4. The Inelastic Case

Our theory says that labor associations will protect their shirkers exactly when demand for labor is inelastic. Given the ubiquity of protection for weakest members we must ask if it is indeed the case that the demand for labor is typically inelastic. Lichter, Peichl and Siegloch (2015) do a meta-study of labor demand.[18] The vast bulk of estimates are for absolute demand elasticities less then 1, that is, inelastic.

These studies measure elasticity at the equilibrium not at the natural level, which is what our results refer to. In the Appendix in Propositions 1 and 2 we show that sign of marginal revenue at the natural level and at the optimum are the same, so this does not matter. In addition these labor demand studies measure the elasticity of hours (n) with respect to wages not the elasticity of labor input (x) with respect to wages which is what our theory refers to: as our effort is endogenous, these two elasticities are not necessarily the same.

There are two cases to consider. If we observe either the short run or shocks that are temporary, effort is fixed and the two elasticities are the same. In the elastic case when effort adjusts to the optimal we show in the Appendix in Proposition 3 that for $\overline{\Delta}(W) = \overline{\Delta}$ either effort is constant so the elasticities are the same, or if effort is not constant then the absolute elasticity is 1. Hence

[18] Notice that the studies that underlie this data refer to elasticity at the equilibrium not at the natural level, while the theory does the opposite. However, Theorem 1 shows that if demand is elastic at x^γ it is at $\hat{x}$ and conversely.

we should not see measured absolute elasticity less than 1 if the absolute labor input elasticity is greater than 1.

In the remainder of this section we focus on the inelastic case. All the results of this section are proven in the Appendix.

Utility of Workers

To what extent are workers content with their colleagues shirking? Specifically we analyze the case in which $-c^w \leq R'(x^\gamma) < 0$ so that the labor association is passive and workers work and shirkers shirk. This is a tricky question to ask in the current context because we assume that *ex ante* employees do not know their own type. If they did we would need to consider the possibility that firms would introduce screening contracts in an effort to lure workers rather than shirkers. Never-the-less we can consider the following conceptual experiment. First, suppose that firms perfectly observes effort but are prohibited by a union contract from paying incentive wages. Second, suppose that after employment and after employees learn their type a vote is taken over whether to keep the labor association or to disband the labor association and allow the firm to pay incentive wages. Define n^* as the competitive equilibrium with full effort: $U'(n^*) = v(n^*)$.

Theorem 2. *Suppose that $-c^w \leq R'(x^\gamma) < 0$ and $\mu^s < U'(n^*)$. Then there exists an $m > 0$ such that for $0 < (1-\gamma)\mu^s < m$ workers are strictly better off with a labor association.*

Notice in particular that the condition $(1-\gamma)\mu^s$ is small will be satisfied if γ is large - that is, if there are many workers they will be strictly better off with the labor association. The point is that without the labor association if the firm observes effort then workers receive a premium because shirkers must be compensated for their effort. However, if shirkers do not require much compensation this premium will be less than the increased wage they will receive if instead shirkers do not provide effort.

An alternative way to analyze the issue of worker attitudes towards shirkers is this: if shirkers by reducing overall effort increase utility for workers then we imagine that workers are grateful to shirkers. Our next result shows that this is in fact the case.

Theorem 3. *Suppose that $-c^w \leq R'(x^\gamma) < 0$. Then worker utility is decreasing in γ.*

This theorem resolves a phenomenon that has long puzzled us. We have observed, for example, in the Italian Post Office, at Departments of Motor Vehicles, and in the private sector at rental car agencies, long queues and a number of windows for servicing customers. Behind most of these windows are employees shuffling papers or otherwise shirking.[19] Sometimes, however, behind one window there is an employee working like a demon trying to get the customers what they want. The question is why the worker puts up with the shirkers. As it appears that demand is generally inelastic, the answer we get from our theory is that the worker - who likes to work - receives a higher utility due to the presence of the shirkers.

Involuntary Disclosure

We now examine the situation in which the labor association cannot prevent the firm from observing the signal. We are particularly interested in how the comparative statics change in this case. For simplicity we will assume that the maximum wage differential $\overline{\Delta}(W)$ is a constant $\overline{\Delta}$ independent of W.

Theorem 4. *Suppose that $R'(x^\gamma) \leq 0$ and define $\tilde{v}(n) = v(n) + \pi\overline{\Delta}$. Then the equilibrium in which the labor association cannot prevent the firm from observing the signal is the same as that in which the firm does not observe the signal and opportunity cost is given by $\tilde{v}(n)$. Labor association utility is decreasing in $\pi\overline{\Delta}$ and consumer utility increasing.*

This implies in particular that in the inelastic case improved information (lower π) is better for the labor association but worse for consumers. In the case of the police improved monitoring technology such as body cams may in fact reduce consumer surplus.

[19] An interesting example is provided in the Walt Disney movie Zootropolis where DMV employees are sloths executing tasks extremely slowly, much to the frustration of customers.

Strong Firms

To further highlight the role of information provision by a labor association to the firm we will consider the case where the firm rather than the labor association sets the quota (which we may assume is a minimum quota) and receives the signal. For simplicity we continue to examine the case in which the maximum wage differential $\overline{\Delta}(W)$ is a constant $\overline{\Delta}$ independent of W. Our interest is in how *signal quality*, measured by $\sigma = \pi' - \pi$ impacts on consumers and the association. Even in the absence of provision of information by the association we may have $\sigma > 0$: for example in the case of the police, civilians with cell phone cameras and body cams may provide useful information about police behavior. While the association does not control the quota, cannot punish, and perhaps does not even see the signal, it can improve the quality of the signal by providing information to the firm. We will establish that greater σ increases labor input and hence consumer utility, but that in the inelastic case where $R'(x^\gamma) < 0$ it reduces employment and hence the association's utility $v(n)$. In this case the association will refuse to provide information to the firm and indeed, if it is able to do so - for example, in the case of the police, by harassing civilian photographers and sabotaging body cameras - the association will attempt to degrade that information. In conclusion, a better firm signal improves consumer surplus but does not break the blue wall of silence.

For concreteness we state the problem of a representative firm and the equilibrium conditions, and we focus on minimum quotas. To ease notation we will suppress the superscript and write simply ϕ to denote the quota ϕ^-. The firm pays $W(0)$ for a bad signal and $W(1) = W(0) + \Delta$ for a good signal and sets an incentive compatible minimum quota $1 \geq \phi \geq 0$. The wage differential constraint is $\Delta \leq \overline{\Delta}$. Workers work and the quota must be incentive compatible for shirkers $\mu^s \phi \leq (\pi' - \pi)\Delta$. The wage bill per worker for the firm is $W = \pi W(0) + (1-\pi)W(1) = W(0) + (1-\pi)\Delta$. The utility provided to a worker is the wage less the cost of effort $\nu = W - (1-\gamma)\mu^s\phi$. In the market the firm takes as given the output price, which we denote by Q, and the worker utility which we denote by V so that it maximizes per worker profits $Q((1-\gamma)\phi + \gamma) - W$ subject to $\Delta \leq \overline{\Delta}$, the incentive constraint and $\nu \geq V$.

In equilibrium $(1-\gamma)\phi + \gamma = \hat{e}$, $\nu = V = v(n)$, $Q = U'(n\hat{e})$ and there is zero profit per worker.

Define *high effort* by $\hat{\phi} = \min\{1, \sigma\overline{\Delta}/\mu^s\}$, $\overline{x}$ as the unique solution of $U'(\overline{x}) = \mu^s$, $\overline{n}$ as the unique solution to $v(n) = \gamma\mu^s$ and the *target effort* $\overline{\phi} = (\gamma/(1-\gamma))\overline{x}/\overline{n}$. Theorem 7 in the Appendix characterizes the unique equilibrium as a function of the target effort $\overline{\phi}$. If this is negative then the optimal effort is 0; if it is bigger than high effort $\hat{\phi}$ it is equal to high effort and if the target effort lies in between then the optimal effort is equal to the target effort.

Focus on the case where the high effort constraint binds $\overline{\phi} > \hat{\phi}$ and suppose that signal quality is not too high in the sense that $\sigma < \mu^s/\overline{\Delta}$. In this case signal quality σ increases high effort $\hat{\phi} = \sigma\overline{\Delta}/\mu^s$: this increases equilibrium effort. Intuitively the firm increases effort to save money by reducing employment and this makes employees worse off. This is proven in Theorem 7 in the Appendix. In othe words, employment and employees' utility are strictly decreasing in the signal quality: the labor association will not provide additional information to the firm, and if it is able to do so will degrade the information received by the firm.

5. Piece-rate Payments

We do not mean to pick on workers as being especially lazy as compared to, for example, proprietors. Proprietors unlike workers cannot contract to be paid regardless of effort - they (as do some workers such as garment workers) are paid a piece-rate proportional to effort. None-the-less similar considerations of elasticity and lack of effort apply. We turn here to proprietors who are paid a piece rate and for simplicity take the neutral assumption that all are identical and that effort has neither cost nor benefit.

We consider a fixed force N of identical proprietors who costlessly provide effort $e^i \in [0,1]$. As before if average effort is $\overline{e}$ total output is $x = N\overline{e}$. We continue to assume the value of output $U(x)$ is strictly differentiably concave up to a satiation level $X > N$ and that the revenue function $R(x) = xU'(x)$ is concave. Now, however, proprietors face a constant marginal cost ξ of other inputs used in producing output and are paid individually for the output they produce, so that the profit of a proprietor is $U'(x)e - \xi e$. We assume that $U'(N) > \xi$ so that the market clearing effort level absent any incentives is N with corresponding market price $U'(N)$. In this context $U'(x)/xU''(x)$ is the elasticity of demand for output.

The group of proprietors also faces a mechanism design problem: they can set an effort quota and observe a noisy signal of whether the quota is adhered to by a member. As before there can be either a minimum quota ϕ^- or a maximum quota ϕ^+. Although the market implicitly measures the effort of each individual proprietor we assume that this information is not so easy for other proprietors to observe. Hence we continue to assume that individual efforts are not observable so that other proprietors observe only a noisy public signal $z^i \in \{0, 1\}$ of adherence to the quota where again 1 is good and 0 is bad. The probabilities of the signal remain $\pi' > \pi$ as the quota is not or is adhered to. Proprietors can impose costly social punishments on each other of P. Roughly speaking we assume that proprietors, whether butchers, country squires, or industrialists like to socialize with people in the same line of business so that ostracism from the association of proprietors is costly.

In formulating a precise result it will be convenient to work with the *inverse elasticity of the price cost margin*

$$\eta(x) \equiv \frac{U'(x) - \xi}{xU''(x)}.$$

The monopoly solution is at $R'(x^m) = \xi$ that is $\eta(x^m) = 1$, while the competitive solution x^c has $U'(x^c) = \xi$ so $\eta(x^c) = 0$; also observe that $\eta(0) = -\infty$. In place of assuming that marginal revenue is increasing with output we will use here the obvious regularity condition that $\eta(x)$ is increasing. When $\xi = 0$ we have $\eta(x)$ is simply the elasticity of demand, and this is the usual assumption that demand elasticity is increasing with output. We can now state our main result in the case of piece-rate payments, proven in the Appendix:

Theorem 5. *A minimum quota is never used. A binding maximum quota is used if and only if*

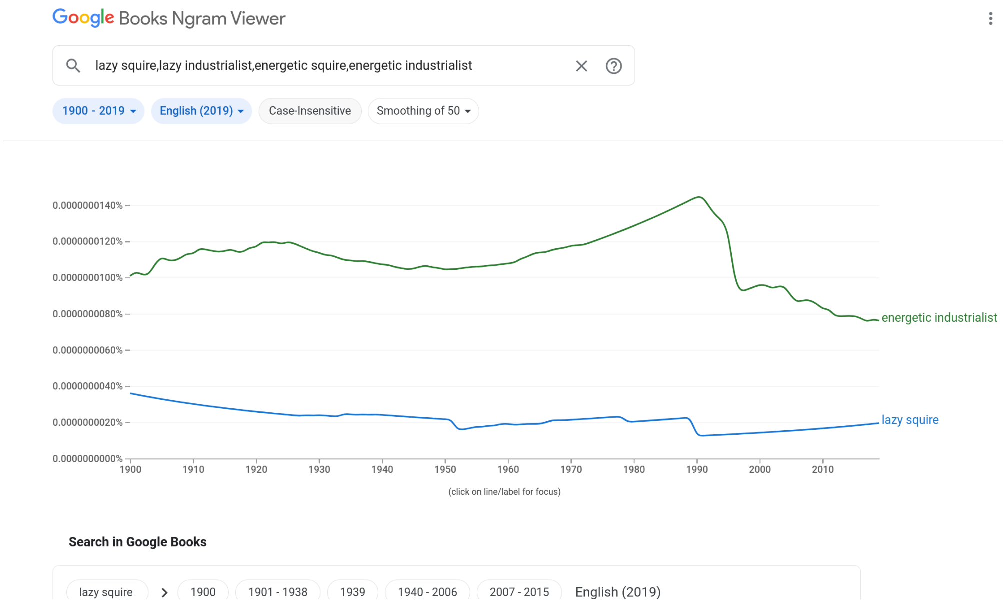
$$-\eta(N) < \frac{\pi' - \pi}{\pi'}.$$

Squires versus Industrialists

Are there cases where proprietors use social incentive to restrict effort? We argue that this was exactly what the British land-owning nobility - the "country squires" engaged in during the 18th and 19th Centuries. The country squire is infamous in British literature for their drunken lazy ways and their engagement in social activities such as throwing parties and fox hunting: Fielding (1742) is scathing in his description of the country squire. The Sicilian aristocracy is equally well known for the same kind of lifestyle (and in fact their British peers were not infrequently among the guests at their lavish parties).

The country squires produced mostly staple agricultural products, mostly grain and primarily for domestic consumption; and demand for these products is known to be inelastic - see, for example, Andreyeva, Long, and Brownell (2010). Since inelastic demand implies an inelastic inverse elasticity of the price cost margin, Theorem 5 implies that a social norm of "spend all your time having parties and fox-hunts rather than running your farm" makes sense - and has relatively low monitoring costs since it is easy to see if your colleagues are inviting you to parties and fox-hunts. In this view, then, the "laziness" of country squires was simply a rational way to restrict output and exercise monopoly power.

In contrast to country squires industrialists were not famed for their laziness. Our Ngram reported below examines the 20th Century English language literature for lazy squire, lazy industrialist, energetic squire, and energetic industrialist. As can be seen squires are frequently described as lazy and industrialists as energetic, but pretty much never the other way around. If indeed demand for industrial products is sufficiently elastic our Theorem 5 makes sense of this.



Intuitively, manufacturers exporting goods face fairly elastic demand due to the presence of many substitutes. From Stokey (2001) we find that indeed during the early industrial revolution output and revenue increased hugely, indicating a high elasticity. Specifically Stokey (2001) reports that from 1780 to 1850 GDP grew by a factor of 3.65 and industrial output by a factor of 6.07 so that there was a large increase in the relative share of industrial output. On the other hand capital's share of GDP rose from .35 to .44. A large relative increase in output share with an increased profit share indicates that indeed demand must have been highly elastic.

6. General Mechanisms

Our analysis has been of a special class of mechanisms: a quota with a bad punishment for a bad signal. Could the labor association do better with a more general mechanism? Roughly speaking the answer is no, but to make this precise we need to consider carefully what a general mechanism would look like.

We should indicate first that the "quota plus signal" is a special case of the type of flexible information system studied by Yang (2015). That is, we may define a flexible class of effort dependent *information systems* each defined by a threshold φ and a direction φ^+, φ^-. If effort lies at or below φ^+ the signal 0 is emitted with probability π and if it lies above the signal 0 is emitted with probability π'. Similarly the information systems φ^- emit 0 for effort at or above φ^- with probability π and below with probability π'. Here the information systems exhibit high sensitivity near the threshold. In Yang (2020) information systems for designing a bond should be sensitive near the default boundary: here they should be sensitive near the target effort level. Indeed, due to the discontinuity at the threshold the only incentive compatible effort targets are either at the discontinuity - equivalent to our quota model - or at 0 or 1 - which is also equivalent to our quota model.[20]

Next, in addition to choosing an information system, the labor association can ask members to reveal their types. It can then issue punishments based on the combination of type statements and signals. This would be the "general mechanism." An important issue is whether in addition to type contingent punishments the association can choose a type contingent information system. For example, it may be that the information system has to be chosen before types are realized. Implicitly we have assumed that this is the case; we indicate below what happens if the association can choose type-contingent information systems.

In the case of non-type contingent information systems, what extra leverage does the labor association gain from punishments that are type and signal contingent rather than merely punishing based on a bad signal? In this discussion, bear in mind that the relevant consideration is how the incentive constraints impact on the cost of achieving an effort target $\bar{e}$. The answer depends on whether or not $\pi < 1 - \pi'$. The reason is that (allowing employees to choose which signal probability to use when they are on the effort boundary $e = \varphi$) if $\pi > 1 - \pi'$ off path punishment costs could be reduced by reversing the role of the two information systems so that the on-path punishment probability would be $1 - \pi'$ rather than π. On the one hand this is really notational, since we can

[20]More general information systems including continuous ones are studied in Dutta, Levine and Modica (2022) who show that if the sensitivity is large enough equilibrium choice of effort resembles that in the discontinuous case.

just redefine the probabilities accordingly. Moreover, an issue not yet studied in the flexible information system literature is that of evasion: the signal that receives punishment creates incentives for the employee to obscure the signal. Hence it might be that $1-\pi'$ is relatively large because employees try to conceal their bad signals.

Assuming either that $\pi < 1 - \pi'$ or that reversal is impossible, basing punishments on types will simply cause employees to lie about their type to receive the lesser punishment. Similarly it makes no sense to punish on both signals since this reduces incentive compatibility while increasing cost. Hence we conclude that the mechanism studied here is indeed the best in the class of general mechanisms.

If it is possible to base the information system on type revelation then the model changes to one with a type-contingent quota. Each quota can have its own punishment which we may denote by P^τ where τ is the type. Let P denote the cost minimizing punishment in the original model with a type independent information system. We have

Theorem 6. *Cost minimization implies $P^s = P^w = P$, and in particular $C(\bar{e})$ does not depend on whether or not type dependent information systems are available.*

Proof. In the inelastic case take $\tau = w$ and in the elastic case take $\tau = s$. For the given maximum or minimum quota we must still minimize cost and have incentive compatibility, meaning that $P^\tau = P$. Certainly $P^{-\tau} = P$ is feasible and $P^{-\tau} > P$ raises costs, so we only need to show that $P^{-\tau} < P^t$ is not incentive compatible. To see why, notice that type τ must be indifferent to their favorite effort level using their own information system and punishment, so receives a utility of $V^\tau \geq -\pi' P^\tau$. If they lie about their type and choose their favorite effort level they would instead get $-\pi' P^{-\tau}$. Hence truth-telling requires $P^{-\tau} \geq V^\tau/\pi' \geq P^\tau$. □

7. Conclusion

Discouraging workers from working and imposing a blue wall of silence is inefficient, and, of course, particularly harmful to consumers. What can be done about it? In the context of the police three strategies have been suggested. The first is to abolish police unions. The second is the increased provision of information - laws that prevent police from interfering with civilians recording encounters and mandating the use of body cams. The third is to "defund the police." Based on our model each of these strategies is problematic; the more traditional solution to monopoly - competition - seems more promising.

Clearly police unions are not a problem per-se. Rather it is the social network of police officers enabling monitoring and peer punishment that leads to effort reduction and the blue wall of silence. Whether there is a formal structure - a union - or not, the police can engage in informal discouragement of effort, and indeed Ostrom (1990) clearly documents how formal institutions are not needed for collective action.

Consider, second, the increased provision of information to firms. This is a double-edged sword. It enables firms to provide better incentives to work - but it also reduces the cost to labor associations of discouraging work. When the labor association controls the quotas we showed that better information makes consumers strictly worse off.[21] In the case of strong firms it does improve consumer surplus but it does not give the labor association any incentive to provide additional information. Hence, it cannot crack the blue wall of silence and we may expect the labor association to fight back by trying to reduce the flow of information.

The policy of "defund the police" is not always clearly described. One thing it may mean is replacing the police with a different type of police with different or additional training, perhaps mental health workers. Since the incentives of whatever association provides "policing" services are the same, it is hard to see how this helps. Alternatively "defund the police" may mean conditioning wages on some sort of measure of average performance: if the police force as a whole fails to live up to some standard they are all fired or their wages are reduced. In the USA police forces are relatively competitive: even within a

[21] This may explain why some police unions have favored body cams: see, for example, https://eu.statesmanjournal.com/story/news/crime/2020/08/03/police-reform-salem-oregon-brutality-body-camera-budget/5453765002/.

single jurisdiction there are typically many police forces, and of course different suburbs often have their own police forces. In the USA as a whole there are roughly 18,000 different police forces.[22] While in reality - unlike the model - it is not costless for police to get a job with another force, it is never-the-less hard to see how such a threat of collective punishment by a single jurisdiction would have much effect on the behavior of the labor association.

Such a policy might have greater success with European style national police forces. Never-the-less we have difficulty in seeing how this would be practical. In terms of our model the government should make a "take-it-or-leave-it" offer of a particular wage per unit of labor input effectively making demand perfectly elastic. There are two issues which give us doubt over whether this would work. First, labor input would have to be measured: the incentive for the labor association to suppress information about wrong-doing would be enormous if their wages were to be reduced based on that information. Second, such a commitment would have to be credible. We wonder as well if this would be the case. Democratic governments have a difficult time with commitment: witness Donald Trump withdrawing the USA from the Paris accords and nuclear agreement with Iran upon becoming president. Perhaps even more striking is that the same British government that agreed to provisions they did not like concerning trade with Northern Ireland in exchange for provisions that they did like on lower tariffs intends to simply ignore the treaty provisions they do not like. Is it so hard to imagine that if a large increase in police wages was called for a government might renege on it?

In our view the heart of the problem is one of monopoly power: the protection of shirkers and the blue wall of silence are rational ways to exploit it. In general the "cure" for monopoly is competition.[23] In the model this corresponds to increasing the elasticity of demand. Not only does this increase labor input and consumer surplus but it can potentially break the blue wall of silence entirely - as we showed, with enough demand elasticity the labor association

[22] https://cops.usdoj.gov/pdf/taskforce/taskforce_finalreport.pdf

[23] There is evidence that increased competition improves labor productivity: see for example Galdon-Sanchez and Schmitz Jr (2002).

prefers to provide information to the firm.

How can increased competition come about? Consider breaking the "defund the police" scheme into specific police services: investigation, patrolling, response to domestic incidents and so forth. Narrower product categories generally have greater elasticity of demand. Hence, rather than a single "one-size fits all" police force, each of these services could be contracted to a different provider. For example, in the USA, the FBI could be hired to investigate, a private security service to patrol, and a mental health provider to respond to domestic incidents. If the social networks of these different providers are different, competition is induced between labor associations. Hence traditional police forces might bid against mental health firms for the contract to respond to domestic incidents - and police social networks are rather different than those of mental health firms. This increased competition also increases the elasticity of demand. While this solution has somewhat the flavor of "defund the police" it is better described as "make them compete."

Appendix

Lemma 1. *The natural level of labor input x^γ and corresponding consumer surplus $U(x^\gamma)$ are strictly increasing in γ.*

Proof. Since consumer surplus is $U(x^\gamma)$ it suffices to show $\partial x^\gamma / \partial \gamma > 0$. Applying the implicit function theorem to $\gamma U'(\gamma \hat{n}(\gamma, 0)) - v(\hat{n}(\gamma, 0)) = 0$ we have that

$$\frac{\partial \gamma \hat{n}}{\partial \gamma} = -\frac{U'(\hat{n}\gamma) + \gamma \hat{n} U''(\hat{n}\gamma)}{\gamma^2 U''(\hat{n}\gamma) - v'(\hat{n})}\gamma + \hat{n}$$

$$= -\frac{\gamma U'(x^\gamma)}{\gamma^2 U''(x^\gamma) - v'(\hat{n})} - \frac{\gamma x^\gamma U''(x^\gamma)}{\gamma^2 U''(x^\gamma) - v'(\hat{n})} + \hat{n}$$

$$= -\frac{\gamma U'(x^\gamma)}{\gamma^2 U''(x^\gamma) - v'(\hat{n})} - \frac{(x^\gamma)^2 U''(x^\gamma)}{(x^\gamma)^2 U''(x^\gamma) - v'(\hat{n})}\hat{n} + \hat{n} > 0.$$

□

How much does it cost to optimally implement a target level of average effort $\overline{e}$? Intuitively, it should depend on the ex-ante likelihood γ of an employee being of the w type. To implement a target level $\overline{e} > \gamma$, incentives for shirkers to provide effort must be used. Otherwise, to implement $\overline{e} < \gamma$, incentives for workers not to provide effort must be used. Defining monitoring difficulty as $\theta = \pi/(\pi' - \pi)$ we have the following result.

The Optimal Mechanism Without Revelation

Lemma 2. *If the association does not reveal the signals the cost of implementing a target level of effort $\overline{e}$ is given by*

$$C(\overline{e}) = \begin{cases} -\left(\mu^w(\gamma+\theta)/\gamma\right)(\overline{e}-\gamma) \equiv -c^w(\overline{e}-\gamma) & \overline{e} < \gamma \quad \text{(low)} \\ 0 & \overline{e} = \gamma \quad \text{(intermediate)} \\ \left(\mu^s(1-\gamma+\theta)/(1-\gamma)\right)(\overline{e}-\gamma) \equiv c^s(\overline{e}-\gamma) & \overline{e} > \gamma \quad \text{(high)} \end{cases}$$

Proof. Incentive compatibility for a minimum quota is that shirkers must prefer providing ϕ^- to not providing effort. This is $-\mu^s\phi^- - \pi P \geq -\pi' P$. The optimal mechanism must minimize C hence P, therefore this constraint must hold with equality. This gives $M = \theta\mu^s\phi^-$ and $C = \mu^s(1-\gamma+\theta)\phi^-$. For the maximum quota we have for workers $-\mu^w(1-\phi^+) - \pi P \geq -\pi' P$. This gives $M = \theta\mu^w(1-\phi^+)$ and $C = \mu^w(\gamma+\theta)(1-\phi^+)$. For $\overline{e} = \gamma$ no incentives are needed, $P = 0$ and maximal effort by workers $e = 1$ and minimal effort by shirkers $e = 0$ are incentive compatible and have associated cost $C = 0$. If $\overline{e} > \gamma$, a minimum quota must be established so that $\overline{e} = (1-\gamma)\phi^- + \gamma$, with corresponding cost $C = \mu^s(1-\gamma+\theta)\phi^- = \mu^s(1-\gamma+\theta)(\overline{e}-\gamma)/(1-\gamma)$. If $\overline{e} < \gamma$, a maximum quota must be established so that $\overline{e} = \gamma\phi^+$, with corresponding cost $C = \mu^w(\gamma+\theta)(1-\phi^+) = \mu^w(\gamma+\theta)(1-\overline{e}/\gamma) = \mu^w(\gamma+\theta)(\gamma-\overline{e})/\gamma$. □

Slightly abusing notation we let $\hat{n}(\overline{e}) \equiv \hat{n}(\overline{e}, C(\overline{e}))$. An optimal target level of average effort is a choice $\overline{e} = \hat{e}$ which maximizes $\hat{n}(\overline{e})$. We let $\hat{x} = \hat{n}(\hat{e})\hat{e}$ so that $\hat{x}$ denotes the *optimal* level of labor input as opposed to x^γ, which is the *natural* level of labor input. Finally, recall that $R(x) = xU'(x)$ is the (concave) revenue function. We are now ready to state our first main result. Recall that since consumer surplus is $U(\hat{x})$, it has the same comparative static as $\hat{x}$. Bearing in mind that x^γ and c^w, c^s are exogenous, we have that:

Proposition 1. *Without revelation there is an optimal mechanism, the optimal target level of average effort $\hat{e}$ is unique and*

- (low) If $R'(x^\gamma) < -c^w$ then a maximum quota is optimal, $\hat{e} < \gamma$, $\hat{x} < x^\gamma$, and $R'(\hat{x}) = -c^w$. Furthermore, optimal labor input and consumer surplus are increasing in c^w, while employee utility is strictly decreasing in μ^w and θ;

- (natural zone) If $-c^w \leq R'(x^\gamma) \leq c^s$ then $\hat{e} = \gamma$ that is no quota is optimal;

- (high) If $R'(x^\gamma) > c^s$ then a minimum quota is optimal ,$\hat{e} > \gamma$, $\hat{x} > x^\gamma$ and $R'(\hat{x}) \geq c^s$ with equality if $\hat{e} < 1$.

Proof. We prove existence constructively.

By Theorem 2 since $C(\gamma) = 0$ and the assumption that $\gamma U'(0) > v(0)$ we have $\hat{n}(\gamma) \equiv \hat{n}(\gamma, 0) > 0$ hence $\hat{n}(\hat{e}) \geq \hat{n}(\gamma) > 0$. When $\hat{n}(\overline{e}) > 0$ we must have $\overline{e}U'(\hat{n}(\overline{e})\overline{e}) - C(\overline{e}) - v(\hat{n}(\overline{e})) = 0$ and since v is positive, this requires $U'(\hat{n}(\overline{e})\overline{e}) > 0$; this implies that $\hat{n}(\hat{e})\hat{e} < X$ (X being the satiation level of utility). Consider two domains: in the higher domain $\overline{e} \geq \gamma$ and $\hat{n}\overline{e} \leq X$; in the lower domain $\overline{e} \leq \gamma$ and $\hat{n}\overline{e} \leq X$. Then $h(\overline{e}, \hat{n}) \equiv \overline{e}U'(\hat{n}\overline{e}) - C(\overline{e}) - v(\hat{n}) = 0$ is smooth in each of these domains and in either one by using the implicit function theorem we obtain

$$\frac{d\hat{n}}{d\overline{e}} = -\frac{U'(\hat{n}\overline{e}) + \hat{n}\overline{e}U''(\hat{n}\overline{e}) - C'(\overline{e})}{\overline{e}^2 U''(\hat{n}\overline{e}) - v'(\hat{n})} = -\frac{R'(\hat{n}\overline{e}) - C'(\overline{e})}{\overline{e}^2 U''(\hat{n}\overline{e}) - v'(\hat{n})}.$$

Since the denominator $\overline{e}^2 U''(\hat{n}\overline{e}) - v'(\hat{n}) < 0$, at an interior local maximum of $\hat{n}(\overline{e})$ where $\overline{e} \neq \gamma$ it must be that $R'(\hat{n}\overline{e}) - C'(\overline{e}) = 0$. Computing the second derivative where $R'(\hat{n}\overline{e}) - C'(\overline{e}) = 0$ and $\overline{e} \neq \gamma$ yields

$$\frac{d^2\hat{n}}{d^2\overline{e}} = -\frac{\hat{n}R''(\hat{n}\overline{e})}{\overline{e}^2 U''(\hat{n}\overline{e}) - v'(\hat{n})} < 0,$$

where we used that $C''(\overline{e}) = 0$ from Theorem 2. This implies that $R'(\hat{n}\overline{e}) - C'(\overline{e}) = 0$ is always a local maximum and not a local minimum. From Theorem 2, we know that in the lower domain $R'(\hat{n}\overline{e}) - C'(\overline{e}) = R'(\hat{n}\overline{e}) + c^w$ while in the higher domain $R'(\hat{n}\overline{e}) - C'(\overline{e}) = R'(\hat{n}\overline{e}) - c^s$. Hence in the lower domain if $R'(x^\gamma) + c^w \geq 0$ there can be no local maximum with $\overline{e} < \gamma$, while in the higher domain if $R'(x^\gamma) - c^s \leq 0$ there can be no local maximum with $\overline{e} > \gamma$. Hence if both these conditions hold we are in the natural zone. Moreover, if the first condition fails we must have $R'(x^\gamma) < 0$ while if the second fails we must have $R'(x^\gamma) > 0$ so at most one of them fails. If the first fails - that is $R'(x^\gamma) + c^w < 0$ - then there must be a unique local maximum in the strict lower domain which, since the second condition holds (that is, $R'(x^\gamma) - c^s < 0$), is a global maximum. Similarly if the second fails - so that $R'(x^\gamma) - c^s > 0$ - there must be a unique global maximum in the strict higher domain. In the lower domain the first order condition $R'(\hat{n}\hat{e}) - c^s = 0$ uniquely determines the maximum, while in the higher domain the constraint $\hat{e} \leq 1$ may bind, so the condition is that given in the theorem. Finally, since $R'(x^\gamma) > c^s$ implies the higher domain, from the first order condition $R'(\hat{n}\hat{e}) - c^s \geq 0$ we see that $R'(\hat{n}\hat{e}) \geq c^s > 0$, and similarly since $R'(x^\gamma) < -c^w$ implies the lower domain, from the first order condition $R'(\hat{n}\hat{e}) + c^w = 0$ we see that $R'(\hat{n}\hat{e}) = -c^w < 0$.

The comparative static results about $\hat{x}$ follow directly from the first order conditions and the fact that marginal revenue is assumed to be decreasing. The comparative statics about employee utility follows from the fact that the increases strictly lower the objective function. □

Proposition 2. *Consider whether there is an optimal mechanism in which the association reveals the signals to firms:*

(Inelastic demand) If $R'(x^\gamma) \leq 0$ there is an optimal mechanism in which the labor association does not to reveal the signal and if $R'(x^\gamma) < -c^w$ there is no optimal mechanism in which it does so.

(Elastic demand) If $R'(x^\gamma) \geq 0$ there is an optimal mechanism in which the labor association does reveal the signal and if $R'(x^\gamma) > c^s$ in every optimal mechanism it does so. A minimum quota is optimal ,$\hat{e} > \gamma$, $\hat{x} > x^\gamma$ and $R'(\hat{x}) \geq 0$.

Proof. In an optimal mechanism with revelation for a given market wage W the labor association must choose a ϕ, P such that there exists a firm optimal choice with expected wage W that is incentive compatible for both types. The premium for a high effort signal is $\Delta = W(1) - W(0)$. In the case of a minimum quota the binding constraint is that shirkers must prefer providing ϕ^- to not providing effort. This is $\mu^s\phi^- + \pi P - \pi W(0) - (1-\pi)W(1) \leq \pi' P - \pi' W(0) - (1-\pi')W(1)$, or

$$\mu^s \phi^- \leq (\pi' - \pi)(P + \Delta).$$

For the maximum quota analogously we have for workers

$$\mu^w (1 - \phi^+) \leq (\pi' - \pi)(P - \Delta).$$

Observe that a deviation from Δ to Δ' by the firm only impacts its profits if it is chosen so that the constraint is violated.

Consider first the inelastic case. If there is revelation it is still the case that no quota is at least as good as a minimum quota. In fact, while no quota has no cost, a minimum quota can only raise output over no quota and raising output in an incentive compatible way has a non-negative punishment cost. If the optimum with revelation is no quota then either that was the optimum without revelation in which case no revelation is weakly preferred, or it was not in which case no revelation is strictly preferred.

Suppose then that there is an optimal mechanism $\hat{\phi}^+, \hat{P}, \hat{W}, \hat{\Delta}$ with revelation and a maximum constraint. Recall that in the original model the punishment value is $\mu^w(1 - \hat{\phi}^+)/(\pi' - \pi)$. Suppose that

$$\mu^w (1 - \hat{\phi}^+) \geq (\pi' - \pi)\hat{P}.$$

Consider $\Delta > 0$, then

$$\mu^w (1 - \hat{\phi}^+) > (\pi' - \pi)(\hat{P} - \Delta)$$

violating the constraint and inducing the workers to work. Per employee this costs the firm $(\pi' - \pi)(W(0) - W) = (\pi' - \pi)(1 - \pi)\Delta$ and the gain is $U'(n\bar{e})\gamma(1 - \hat{\phi}^+)$. In other words, by choosing Δ close to zero the firm could increase its profit. Hence if the maximum constraint is to be incentive compatible for the firm it must be that

$$\mu^w (1 - \hat{\phi}^+) < (\pi' - \pi)\hat{P}.$$

In this case the labor association can get the same outcome by not revealing and choosing $P = \mu^w(1 - \hat{\phi}^+)/(\pi' - \pi)$ strictly reducing cost and is therefore better off.

Consider then the elastic case. Let $\hat{\phi}^-, \hat{P}, \hat{W}$ be the optimum without revealing the signal. If this is a no quota then the firm cannot provide incentives either, so revealing makes no difference and the labor association weakly prefers to reveal. The greatest incentive that can be provided by firms is $\overline{\Delta}(\hat{W})$ and if it is employer optimal this will minimize punishment costs. If the firm can increase Δ it must do so by raising W and this raises firm costs without changing worker behavior.[24] If it decreases Δ this will induce the shirkers to violate the constraint decreasing output. Let $x = n\left(\gamma + (1-\gamma)\hat{\phi}^-\right)$. The reduction in output costs the firm $U'(x)(1-\gamma)\hat{\phi}^-$ but enables it to reduce wages by $\mu^s(1-\gamma)\hat{\phi}^-$. We know that

$$U'(x) \geq R'(x) > c^s = \mu^s \frac{1-\gamma+\theta}{1-\gamma} \geq \mu^s$$

so this is unprofitable.

Hence in an optimal mechanism $P = \max\{0, \mu^s\hat{\phi}^-/(\pi'-\pi) - \overline{\Delta}(\hat{W})$. Recall that the punishment value in the original model is $\mu^s\hat{\phi}^-/(\pi'-\pi)$. Hence the labor association can obtain the same result in terms of effort and wage by revealing while incurring strictly less punishment cost, so it strictly prefers to reveal.

It remains to show that an optimal mechanism with revelation actually exists. We have $C(W,\overline{e}) = \mu^s(\overline{e}-\gamma) + \max\{0, \theta(\overline{e}-\gamma) - \overline{\Delta}(W)$, $W = \overline{e}U'(n\overline{e})$, $W - C(W,\overline{e}) = v(n)$. Fixing $\overline{e}$ we see that there are two equations in W, n: $W = \overline{e}U'(n\overline{e})$ and $W - C(W,\overline{e}) = v(n)$. Since $\overline{\Delta}(W)$ is non-decreasing $C(W,\overline{e})$ is non-increasing in W so $W - C(W,\overline{e})$ is strictly increasing in W so has an inverse $\hat{W}(v,\overline{e})$ strictly increasing in v. Hence $W = \hat{W}(v(n),\overline{e})$ is strictly increasing in n. Since $W = \overline{e}U'(n\overline{e})$ is weakly decreasing in n there is a unique $\hat{n}(\overline{e})$. Moreover, if $\overline{e}^m \to \overline{e}$ and $n^m \to n$, $W^m \to W$ with $C(W^m,\overline{e}^m) = \mu^s(\overline{e}^m - \gamma) + \max\{0, \theta(\overline{e}^m - \gamma) - \overline{\Delta}(W^m)$, $W^m = \overline{e}^m U'(n^m\overline{e}^m)$, $W^m - C(W^m,\overline{e}^m) = v(n^m)$

[24]Note that the firm cannot decrease W since then it will lose all its workers.

then since everything is continuous $C(W, \overline{e}) = \mu^s(\overline{e}-\gamma)+\max\{0, \theta(\overline{e}-\gamma)-\overline{\Delta}(W)$, $W = \overline{e}U'(n\overline{e})$, $W - C(W, \overline{e}) = v(n)$, so $\hat{n}(\overline{e})$ is continuous, so has a unique maximum defining the optimal mechanism.

Since $\hat{n}(\overline{e})$ with revelation is at least $\hat{n}$ without revelation, the properties ,$\hat{e} > \gamma$, $\hat{x} > x^\gamma$ are inherited from the case without revelation. Finally, $R'(\hat{x}) \geq 0$ since $R'(\hat{x}) = 0$ is strictly better for the association and they can limit x through choice of the quota. □

Proposition 3. *In the elastic case for labor supply* $v(n) + \epsilon$ *and* $\overline{\Delta}(W) = \overline{\Delta}$ *either effort is constant with respect to* ϵ *or* nW *is constant with respect to* ϵ.

Proof. At a corner solution for $\overline{e}$ effort does not respond to ϵ. At an interior optimum $R'(\hat{n}\overline{e}) = c^s$ so that $\hat{x}$ is fixed and $nW = R(\hat{x})$. □

The Inelastic Case

Theorem. *(Theorem 2 in the text) Suppose that* $-c^w \leq R'(x^\gamma) < 0$ *and* $\mu^s < U'(n^*)$. *Then there exists an* $m > 0$ *such that for* $0 < (1-\gamma)\mu^s < m$ *workers are strictly better off with a labor association.*

Proof. Suppose there is no labor association and the firm can observe effort. For $\mu^s < U'(n^*)$, as in any equilibrium $n \leq n^*$ we have $\mu^s < U'(n)$ and it is efficient for shirkers to provide full effort. Hence the equilibrium wage (with full effort) is $W^{NA} = U'(n^{NA}) = v(n^{NA}) + \mu^s(1-\gamma)$ where n^{NA} is equilibrium employment. That is, workers get a premium because shirkers must be compensated for their effort. With a labor association, since $\overline{e} = \gamma$ and $C = 0$, the wage is given by $W^A = \gamma U'(\gamma n^A) = v(n^A)$. Let $W^* = U'(n^*)$. Because $R' < 0$ it must be that $n^A > n^*$ and $W^A > W^*$. As n^*, W^* corresponds to n^{NA}, W^{NA} with $\mu^s = 0$, by continuity for small enough μ^s we have $n^A > n^{NA}, W^A > W^{NA}$. □

Theorem. *(Theorem 3 in the text) Suppose that* $-c^w \leq R'(x^\gamma) < 0$. *Then worker utility is decreasing in* γ.

This theorem resolves a phenomenon that has long puzzled us. We have observed, for example, in the Italian Post Office, at Departments of Motor Vehicles, and in the private sector at rental car agencies, long queues and a number of windows for servicing customers. Behind most of these windows are employees shuffling papers or otherwise shirking.[25] Sometimes, however, behind one window there is an employee working like a demon trying to get the customers what they want. The question is why the worker puts up with the shirkers. As it appears that demand is generally inelastic, the answer we get from our theory is that the worker - who likes to work - receives a higher utility due to the presence of the shirkers.

Proof. Define $\hat{n} = \hat{n}(\gamma, 0)$.From the market clearing condition $\gamma U'(\gamma\hat{n}) = v(\hat{n})$ and the implicit function theorem we get

$$\frac{\partial \hat{n}}{\partial \gamma} = -\frac{U'(\hat{n}\gamma) + \gamma\hat{n}U''(\hat{n}\gamma)}{\gamma^2 U''(\hat{n}\gamma) - v'(\hat{n})} = -\frac{R'(x^\gamma)}{\gamma^2 U''(x^\gamma) - v'(\hat{n})},$$

which is negative when $R'(x^\gamma) < 0$. Since in the natural zone workers get to make full effort and there is no punishment, their utility is their wage, and this decreases as higher γ lowers employment. □

Theorem. *(Theorem 4 in the text) Suppose that $R'(x^\gamma) \leq 0$ and define $\tilde{v}(n) = v(n) + \pi\overline{\Delta}$. Then the equilibrium in which the labor association cannot prevent the firm from observing the signal is the same as that in which the firm does not observe the signal and opportunity cost is given by $\tilde{v}(n)$. Labor association utility is decreasing in $\pi\overline{\Delta}$ and consumer utility increasing.*

Proof. Note that since the labor association controls the quota they need not accept any output higher than the natural level x^γ since they can attain this at zero cost by setting a maximum quota of zero. The incentive constraint is $\mu^w(1-\phi^+) \leq (\pi' - \pi)(P - \Delta)$ and must hold for all $\Delta \leq \overline{\Delta}$ so is equivalent to $\mu^w(1-\phi^+) \leq (\pi' - \pi)(P - \overline{\Delta})$. Minimizing with respect to P gives

$$P = \frac{\mu^w(1-\phi^+)}{\pi' - \pi} + \overline{\Delta}$$

[25] An interesting example is provided in the Walt Disney movie Zootropolis where DMV employees are sloths executing tasks extremely slowly, much to the frustration of customers.

resulting in a cost of $C(\overline{e}) = -\left(\mu^w(\gamma+\theta)/\gamma\right)(\overline{e}-\gamma) + \pi\overline{\Delta}$. In the equilibrium condition this is equivalent to shifting the opportunity cost of labor up by $\pi\overline{\Delta}$.

In the natural zone utility is unchanged since the signal is being used. When the first order condition holds $R'(\hat{n}\hat{e}) + c^w = 0$ and the equilibrium condition is $U'(\hat{n}\hat{e}) = v(\hat{n}) + \pi\overline{\Delta}$. Since $\hat{x}$ does not depend on $\pi\underline{\Delta}$ we see that consumer surplus does not change, and from the implicit function theorem follows that

$$\frac{\partial \hat{n}}{\partial \pi\overline{\Delta}} = -\frac{1}{v'(\hat{n})} < 0$$

meaning that the labor association is worse off.

However: increasing $\pi\overline{\Delta}$ will cause the solution to $\gamma U'(n\gamma) = v(n) + \pi\overline{\Delta}$ to decline which can cause a jump from the low solution to the natural zone: this would raise consumer utility (and lower labor association utility). $\square$

Recall that *high effort* is $\hat{\phi} = \min\{1, \sigma\overline{\Delta}/\mu^s\}$, $\overline{x}$ is the unique solution of $U'(\overline{x}) = \mu^s$, $\overline{n}$ is the unique solution to $v(n) = \gamma\mu^s$ and $\overline{\phi} = (\gamma/(1-\gamma))\overline{x}/\overline{n}$, and recall that we have specialized to $\overline{\Delta}(W) = \overline{\Delta}$.

Theorem 7. *With strong firms there is a unique equilibrium given as follows:*

(low effort) If $\overline{\phi} < 0$ then $\phi = 0$, employment n^ℓ is the unique solution of $\gamma U'(n\gamma) - v(n) = 0$ with labor input $x^\ell = \gamma n^\ell$.

(intermediate effort) If $0 \leq \overline{\phi} \leq \hat{\phi}$ then $\phi = \overline{\phi}$, employment is $\overline{n}$ with labor input $x^m = \left((1-\gamma)\overline{\phi} + \gamma\right)\overline{n}$.

(high effort) If $\overline{\phi} > \hat{\phi}$ then $\phi = \hat{\phi}$, employment n^h is the unique solution of

$$U'\left(n((1-\gamma)\hat{\phi}+\gamma)\right)((1-\gamma)\hat{\phi}+\gamma) - v(n) - \mu^s(1-\gamma)\hat{\phi} = 0$$

with labor input $x^h = \left((1-\gamma)\hat{\phi}+\gamma\right)n^h$.

For fixed γ, μ^s employment is strictly increasing in ϕ for $R'(x) - \mu^s > 0$ and strictly decreasing for $R'(x) - \mu^s < 0$, while labor input and consumer surplus are strictly increasing in both cases.

Proof. We may write the per worker profit as $Q((1-\gamma)\phi+\gamma) - W(0) - (1-\pi)\Delta$ and this must be maximized subject to $\Delta \leq \overline{\Delta}$, $W(0) + (1-\pi)\Delta - (1-\gamma)\mu^s\phi \geq V$, $(\pi' - \pi)\Delta \geq \mu^s\phi$, and $0 \leq \phi \leq 1$.

Suppose we have a solution to this problem $W(0), \Delta, \phi$ and consider an alternative profile $\tilde{W}(0) = W(0) - (1-\pi)(\overline{\Delta} - \Delta)$, $\tilde{\Delta} = \overline{\Delta}$ and $\tilde{\phi} = \phi$: this is feasible and yields the same profit so it also optimal. Hence we may assume $\Delta = \overline{\Delta}$. Hence per worker profit is $Q((1-\gamma)\phi + \gamma) - W(0) - (1-\pi)\overline{\Delta}$ and this must be maximized subject to $W(0) + (1-\pi)\overline{\Delta} - (1-\gamma)\mu^s\phi \geq V$, $(\pi' - \pi)\overline{\Delta} \geq \mu^s\phi$, and $0 \leq \phi \leq 1$.

We immediately see that the constraint $W(0) + (1-\pi)\overline{\Delta} - (1-\gamma)\mu^s\phi \geq V$ must bind so $W(0) = V + (1-\gamma)\mu^s\phi - (1-\pi)\overline{\Delta}$ and profits are

$$Q((1-\gamma)\phi + \gamma) - V - (1-\gamma)\mu^s\phi.$$

The derivative with respect to ϕ is $Q(1-\gamma) - (1-\gamma)\mu^s$. If $Q < \mu^s$ then $\phi = 0$. If $Q = \mu^s$ then ϕ is limited by the second constraint $(\pi' - \pi)\overline{\Delta} \geq \mu^s\phi$. If $Q > \mu^s$ then the constraint on ϕ must bind. Hence we have that:

If $Q < \mu^s$ then $\phi = 0$ and in equilibrium $Q = U'(\gamma)$ and profits are $\gamma U'(\gamma) - v(n)$ which is positive for $n < n^\ell$ and negative for $n > n^\ell$.

If $Q = \mu^s$ then profits are $\mu^s((1-\gamma)\phi + \gamma) - v(n) - (1-\gamma)\mu^s\phi = \mu^s\gamma - v(n)$ which is is positive for $n < \overline{n}$ and negative for $n > \overline{n}$.

If $Q > \mu^s$ then the constraint on ϕ must bind so that $\phi = \hat{\phi}$ and $W(0) = V + (1-\gamma)\mu^s\hat{\phi} - (1-\pi)\overline{\Delta}$. Take $\hat{e} = \gamma + (1-\gamma)\hat{\phi}$. Per worker profit is

$$Q\hat{e} - \left[V + (1-\gamma)\mu^s(\hat{e} - \gamma) - (1-\pi)\overline{\Delta}\right] - (1-\pi)\overline{\Delta}$$

$$= Q\hat{e} - V - \mu^s(\hat{e} - \gamma).$$

The equilibrium condition is then

$$U'(n\hat{e})\hat{e} - v(n) - \mu^s(\hat{e} - \gamma) = 0.$$

In all cases profits are zero

$$U'(n((1-\gamma)\phi + \gamma))((1-\gamma)\phi + \gamma) - v(n) - (1-\gamma)\mu^s\phi = 0.$$

From the implicit function theorem

$$\frac{dn}{d\phi} = -(1-\gamma)\frac{((1-\gamma)\phi+\gamma)nU''(x) + U'(x) - \mu^s}{((1-\gamma)\phi+\gamma)^2U''(x) - v'(n)}$$

where $x = ((1-\gamma)\phi+\gamma)n$, which further simplify into

$$= -(1-\gamma)\frac{R'(x) - \mu^s}{((1-\gamma)\phi+\gamma)^2U''(x) - v'(n)}$$

Moreover,

$$\frac{dx}{d\phi} = (1-\gamma)n + ((1-\gamma)\phi+\gamma)\frac{dn}{d\phi}$$

$$= (1-\gamma)\left(-\frac{xU''(x) + U'(x) - \mu^s}{((1-\gamma)\phi+\gamma)^2U''(x) - v'(n)}((1-\gamma)\phi+\gamma) + n\right)$$

$$= (1-\gamma)\left(\frac{x^2U''(x)n + (U'(x) - \mu^s)nx}{-x^2U''(x) + n^2v'(n)} + n\right)$$

Since

$$\frac{x^2U''(x)n}{-x^2U''(x) + n^2v'(n)} > -n$$

we have $dx/d\phi > 0$ for $U'(x) \geq \mu^s$, and x, ϕ constant for $U'(x) < \mu^s$.

We conclude that $x^h > x^m > x^l$ hence $U'(x^h) < U'(x^m) < U'(x^l)$. If $\mu^s < U'(x^h)$ then it is the high case and $dx/d\phi > 0$ implies $d\phi/dx > 0$ so $\overline{\phi} > \hat{\phi}$. If $U'(x^h) \leq \mu^s \leq U'(x^l)$ then by definition of x^m we have $U'(x^m) = \mu^s$ putting us in the intermediate case and $d\phi/dx \geq 0$ implies that $0 \leq \overline{\phi} \leq \hat{\phi}$. Finally if $\mu^s > U'(x^l)$ then we are in the low case and clearly $\phi = 0$. □

Proprietors

Theorem. *(Theorem 5 in the text) With proprietors a minimum quota is never used. A binding maximum quota is used if and only if*

$$-\eta(N) < \frac{\pi' - \pi}{\pi'}.$$

Proof. Individual proprietor profit is given by $U'(N\overline{e})e - \xi e$. Since $\overline{e} \leq 1$, $U'(N) > \xi$ and decreasing, it follows that proprietors would like to increase effort over any quota so minimum quotas would be useless.

The maximum incentive constraint is $U'(N\phi^+)\phi^+ - c\phi^+ - \pi P \geq U'(N\phi^+) -$

$\xi - \pi' P$ or

$$P = \frac{(U'(N\phi^+) - \xi)(1 - \phi^+)}{\pi' - \pi}.$$

Hence profits of a member of the association is given by

$$\begin{aligned} & U'(N\phi^+)\phi^+ - \xi\phi^+ - \frac{\pi}{\pi' - \pi}\big(U'(N\phi^+) - \xi\big)(1 - \phi^+) \\ = & \big(U'(N\phi^+) - \xi\big)\left[\frac{\pi'}{\pi' - \pi}\phi^+ - \frac{\pi}{\pi' - \pi}\right] \end{aligned}$$

Differentiate with respect to ϕ^+ to get

$$\frac{\pi'}{\pi' - \pi}\big[N\phi^+ U''(N\phi^+) + U'(N\phi^+) - \xi\big] - \frac{\pi}{\pi' - \pi} N U''(N\phi^+).$$

This has the same sign as

$$-\frac{U'(N\phi^+) - \xi}{N\phi^+ U''(N\phi^+)} - \left[1 - \frac{\pi}{\pi'}\frac{1}{\phi^+}\right] = -\eta(N\phi^+) - \left[1 - \frac{\pi}{\pi'}\frac{1}{\phi^+}\right].$$

As this is decreasing in ϕ^+ we see that the condition for a binding maximum constraint is indeed

$$-\eta(N) - \left[1 - \frac{\pi}{\pi'}\right] > 0.$$

□

References

Andreyeva, T., M. W. Long and K. D. Brownell (2010): "The impact of food prices on consumption: a systematic review of research on the price elasticity of demand for food," *American Journal of Public Health* 100: 216-222.

Benoît, Jean-Pierre, and Juan Dubra (2004). "Why do good cops defend bad cops?" *International Economic Review* 45: 787-809.

Chin, G. J., and S. C. Wells, S. C. (1997): "The blue wall of silence as evidence of bias and motive to lie: A new approach to police perjury," *U. Pitt. L. Rev.* 59, 233.

Dutta, R., D. K. Levine and S. Modica (2022): "Interventions with Sticky Social Norms: A Critique," forthcoming *Journal of the European Economic Association* 20: 39-78.

Dutta, R., D. K. Levine and S. Modica (2021): "The Whip and the Bible: Punishment Versus Internalization," forthcoming *Journal of Public Economic Theory* 1-37.

Fielding, Henry (1742): *Joseph Andrews*.

Galdon-Sanchez, J. E., and J. A. Schmitz Jr (2002): "Competitive pressure and labor productivity: World iron-ore markets in the 1980's," *American Economic Review* 92: 1222-1235.

Iris, Mark (1998): "Police Discipline in Chicago: Arbitration or Arbitrary," *J. Crim. L. & Criminology* 89: 215.

Kandori, M. (1992): "Social Norms and Community Enforcement," *The Review of Economic Studies* 59: 63-80.

Khaliq, Amir A., Hani Dimassi, Chiung-Yu Huang, Lutchmie Narine, and Raymond A. Smego Jr. (2005): "Disciplinary action against physicians: who is likely to get disciplined?: *The American journal of medicine* 118: 773-777.

Kraut, A. I. (1975): "Prediction of managerial success by peer and training-staff ratings," *Journal of Applied Psychology* 60: 14.

Levine, D. K., A. Mattozzi and S. Modica (2020): "Trade Associations: Why Not Cartels?" *International Economic Review* 62: 47-64.

Levine, D. K.and S. Modica (2016): "Peer Discipline and Incentives within Groups," *Journal of Economic Behavior and Organization* 123: 19-30

Levine, D. K. and W. Pesendorfer (1995): "When are Agents Negligible?" *American Economic Review* 85: 1169-1179.

Levine, D. K.and S. Modica (2017): "Size, Fungibility, and the Strength of Lobbying Organizations", *European Journal of Political Economy* 49: 71-83

Lichter, A., A. Peichl, A. and S. Siegloch. (2015): "The own-wage elasticity of labor demand: A meta-regression analysis," *European Economic Review* 80:

94-119.
Mechtenberg, L., G. Muehlheusser and A. Roider (2020): "Whistleblower protection: Theory and experimental evidence," *European Economic Review* 126.
Moe, T. (2011): *Special Interest: Teachers Unions and America's Public Schools*, Brookings Institution Press.
Commision to Investigate Allegations of Police Corruption and the Anti-corruption Procedures of the Police Department Report. The City of New York, New York.
Muehlheusser, Gerd, and Andreas Roider (2008): "Black sheep and walls of silence," *Journal of Economic Behavior & Organization* 65: 387-408.
Ostrom, E. (1990): *Governing the commons: the evolution of institutions for collective action*, Cambridge university press.
Schmitz Jr, James A. (2005): "What determines productivity? Lessons from the dramatic recovery of the US and Canadian iron ore industries following their early 1980s crisis." *Journal of political Economy* 113: 582-625.
Simon, L. K. and W. R. Zame (1990): "Discontinuous Games and Endogenous Sharing Rules," *Econometrica* 58: 861-872.
Stokey, Nancy L. (2001): "A quantitative model of the British industrial revolution, 1780–1850," *Carnegie-Rochester conference series on public policy* 55: 55-109.
Yang, M. (2015): "Coordination with flexible information acquisition," *Journal of Economic Theory* 158: 721-738.
Yang, M. (2020): "Optimality of Debt under Flexible Information Acquisition," *Review of Economic Studies* 87: 487-536
Zieger, R. H. (1994): *American workers, American unions*, Johns Hopkins Univ. Press.