Achieving Cooperation under Privacy Concerns

By Wioletta Dziuda and Ronen Gradwohl*

Two players choose whether to cooperate on a project. Each of them is endowed with some evidence, and if both possess a sufficient amount, then cooperation is profitable. In order to facilitate cooperation, the players reveal evidence to one another. However, some players are concerned about privacy, and so revelation of evidence that does not result in cooperation is costly.
We show that in equilibrium evidence can be exchanged both incrementally and all at once, and identify conditions under which the different rates of evidence exchange are optimal.

When two parties communicate in an attempt to undertake a joint venture, the conventions and protocols that structure their communication may be formed by a variety of factors. In this paper we analyze the interplay of two such factors: On the one hand, in order to cooperate successfully a party must communicate some proprietary information that is necessary for the venture. On the other hand, parties may have privacy concerns: If the joint venture fails to materialize, a party may be adversely affected by the other's use of the revealed information.

Consider the following examples of communication with such privacy concerns: Two firms with complementary expertise wish to cooperate on a project. The execution and success of the project depend on firms sharing their expertise and ideas. However, some firms' level of expertise and novelty of ideas may be too poor to permit successful cooperation. If this fact becomes clear in the process of communication, the project is abandoned. In such an event, however, a firm that revealed promising ideas prior to the abandonment may regret doing so. Anticipating this, firms may want to structure their communication in ways that minimize the harm sustained in case cooperation fails.

Next, consider two shady characters who wish to engage in a less-than-legal venture. They exchange plans for a potential criminal scheme, references to criminal connections that may be useful in their venture, and descriptions of other activities to which only the criminal underworld is privy. However, some characters may be undercover cops—they are uninterested in a criminal venture, but rather in obtaining information from the criminal that may lead to an arrest. So while the potential profits from the venture may render information exchange appealing, the characters' concerns for privacy drive them to structure communication in a way that minimizes the amount of incriminating information revealed

* Dziuda: Kellogg School of Management, Northwestern University, Evanston, IL 60208, USA. E-mail: wdziuda@kellogg.northwestern.edu. Gradwohl: Kellogg School of Management, Northwestern University, Evanston, IL 60208, USA. E-mail: r-gradwohl@kellogg.northwestern.edu. We gratefully acknowledge NSF award #1216006. We would also like to thank participants of seminars at Northwestern, UCSD, and Penn State, as well as the Midwest Economic Theory Meetings in St. Louis.

to undercover agents.

As a third example, consider two researchers, a theoretician and an empiricist, who are interested in joining forces on a project. They engage in communication to convey the content of their research and write a joint paper. At the same time, they face uncertainty about the viability of their potential research partner. Does the theoretician actually have a sound and reasonable model, and does she have theorems with correct proofs that fit the project's aims? Does the empiricist have sufficient data, and do those data support the project's aims? If the answer is no, each researcher may be unwilling to reveal her own ideas, as those might potentially be exploited by the other.

In this paper we analyze the tradeoff between the two conflicting forces demonstrated in the examples—the necessity of information exchange versus the concern for privacy—and examine its effect on the structure of communication. In particular, what is the optimal rate of information exchange? Is optimal communication incremental, with parties revealing little bits of information in alternating fashion? Or is optimal communication simple, with one party revealing all her information at once?

Our main result is that both modes of communication—incremental and simple—may be optimal in equilibrium. Which of the two obtains depends on the order in which information must be revealed. Suppose that the initial pieces of information to be exchanged are unlikely to reveal the viability of the opponent or that they tend to inflict a relatively high cost due to privacy loss. In this case, the optimal mode of communication is simple, with one player revealing all evidence first. By a similar intuition, incremental information exchange is optimal when the initial pieces of information to be exchanged are relatively likely to reveal the viability of the opponent or when they inflict little harm due to privacy loss.

In this paper, we take the order of information exchange as given. Such an assumption is plausible in many applications in which information must be revealed in a predetermined order. For example, a theoretician describing a proof must reveal lemmas that build one on top of another. However, in other applications parties may have more flexibility as to the order in which information is presented: A firm can reveal its financial statements to potential partners before or after allowing them to visit its factories, and criminals may agree on what proofs of viability to present first. For such cases, our analysis sheds light on the optimal order in which information should be revealed. We show that privacy leakage is smaller when incremental exchange is optimal than when simple exchange is optimal. Hence, when they have the flexibility to do so, parties should order information in such a way that optimal communication is incremental: Information that is less valuable or more likely to demonstrate viability should be revealed first.

Organization. — Immediately following is a brief survey of the related literature. Sections I and II contain the model and its analysis, the latter of which includes

our main results about the optimality of incremental and simple communication. In Section III we then discuss the robustness of the results to our assumptions. Finally, most proofs are deferred to the Appendix.

A. Related Literature

This paper is part of a large literature on strategic information exchange that was pioneered by Crawford and Sobel (1982) (cheap-talk), Milgrom (1981), and Milgrom and Roberts (1986) (verifiable information). Our notion of information is modeled after Shin (1994) and Dziuda (2011), in that players must reveal their information truthfully, but can obstruct their type by withholding some information. Unlike most of the communication literature (exceptions include Li, Rosen and Suen (2001)), in our paper both players have private information about the relevant state variable (their type) and both communicate this information. Moreover, in our paper the messages sent affect players' utilities. This is similar to Kartik (2009), who assumes that lying is costly, with the difference that in our paper revealing (by assumption truthfully) information is costly. It is also similar to Jovanovic (1982), who assumes communication is verifiable, but has a fixed cost.

The paper that is closest to ours is Augenblick and Bodoh-Creed (2014). In that paper, developed concurrently with and independently of ours, each party has a privately observed type and wishes to find a matching partner, but prefers to confuse non-matching partners about her type. The authors assume that pieces of information are heterogenous and focus on the order in which they should be revealed. We analyze a somewhat orthogonal problem by taking the order of information exchange as given and focusing on whether information exchange should be simple or incremental. This question is minor in Augenblick and Bodoh-Creed (2014), because in their setting communication is essentially one-sided: One player reveals information about her type, the other confirms whether her type matches or not, and communication stops when the receiver does not confirm the match. Therefore in the sender-optimal equilibrium, incremental communication is optimal. In our model communication must be two-sided as each of the viable types possesses different information. As a result, simple communication can be optimal.

Information in our model can be interpreted as money or effort, and hence communication can be viewed as contributions to a common project. In most models on this topic, contributions of players are substitutes in the sense that a contribution by one player decreases the amount that another player needs to contribute. This generates a free-riding problem. The literature finds that gradualism—splitting contributions into smaller pieces and contributing over time—may ameliorate this problem. There are two main mechanisms through which this happens. Marx and Matthews (2000) (and also Compte and Jehiel (2004), Lockwood and Thomas (2002) in the context of a cooperation game, and Pitchford and Snyder (2004) in the context of the hold-up problem) construct an equilibrium in which

contributions are supposed to follow certain gradual pattern, and any deviation from it is punished by halting all further contributions. By making contributions gradual, the incentives of the early contributors are maximized: Any deviation results in a drastic decrease in the expected benefit. Yildirim (2006), Kessing (2007), and Georgiadis (2013) restrict the contributions to take place over time (either as a physical restriction or by assuming a quadratic cost of per-period contributions) and show that then players' contributions become complements: The discounted benefit of the project may be insufficient for a single player to contribute. But a small contribution by one player brings forward the date of completion of the project, which in turn increases the incentives of other players to contribute.[1]

Our model can be reinterpreted as a contribution game with two major changes: (i) Each player can contribute at most half the cost of the public good, so the contributions are complements and no free-riding occurs; (ii) There is uncertainty about whether the opponent can contribute, which can be resolved only in the process of contributing. We assume that this uncertainty is sufficiently low that the viable players are willing to cooperate even if contributions are not gradual. Hence, in contrast with the literature mentioned above, gradualism in our model is not a means to achieve cooperation.[2] However, the resolution of the uncertainty depends on the amount contributed by the players. If the probability that this uncertainty is resolved after an additional contribution decreases with the level of contributions, gradualism minimizes the expected contributions at the time of resolution of uncertainty.

Watson (1999, 2002) obtains that gradualism may be optimal in partnerships if asymmetric information is present. In these papers, high types want to stay in the partnership forever, while low types have an incentive to exit unless the partnership level increases quickly. Watson (1999, 2002) shows that starting at a low level of partnership and increasing it slowly encourages the low types to exit early. In our main model, the unviable players cannot be incentivized to drop out early. Instead, the probability of an unviable type dropping out is increasing in the amount of "investment" she made. Additionally, unlike in Watson (1999, 2002), players in our model can choose the level of "investment" in an asymmetric way: One player can reveal a lot of evidence before the other starts speaking. Revealing only a small piece first and asking the opponent to reciprocate is clearly beneficial for the player in question, but is harmful to the opponent. As a result, whether socially optimal exchange is incremental or not depends on how these benefits of one player compare to the costs to the other. In an extension, we analyze a model in which the unviable types also have privacy concerns, and hence like in Watson

[1] In Admati and Perry (1991) gradualism in contributions comes from the convexity of the cost function and would be optimal even with a single player. Moreover, as Compte and Jehiel (2003) argue, the insight of Admati and Perry (1991) about gradualism is sensitive to the symmetry assumption.

[2] Clearly, if gradualism increases players' payoffs, that increases their incentives to participate in the game, which in turn increases the set of parameters for which cooperation is possible. This, however, is a secondary aim.

(1999, 2002), they can be incentivized to drop out. The equilibria we construct in this section rely on a similar mechanism as in Watson (1999, 2002): If evidence is exchanged in small pieces, then the benefit of staying in the game and obtaining an extra piece of information outweighs the cost of pretending to be viable for one more round.

Hörner and Skrzypacz (2011) analyze a problem in which an uninformed principal wishes to acquire information from a possibly informed agent. The principal cares about money (which translates to a form of privacy concerns), but the agent does not have privacy concerns. Hörner and Skrzypacz (2011) show that in the equilibrium that maximizes the surplus of the principal and the informed agent, information and payments are exchanged gradually.

In Chen and Olszewski (forthcoming), two discussants each wish to persuade an audience, and the order in which they exchange evidence affects the audience's posterior beliefs. Also related is the literature on sustaining conversations, including Stein (2008) and Ganglmair and Tarantino (2014). While these papers also have an element of privacy concerns, the driving force behind their models is that conversation generates new ideas.

Within the computer science literature on exchange protocols, Blum (1983), Damgård (1995), and Bardsley, Clausen and Teague (2008) show that incremental communication can facilitate the exchange of secrets in settings in which players can, for some cost, discover the opponent's secret even in the absence of communication. Finally, in the cryptography literature on zero-knowledge proofs (pioneered by Goldwasser, Micali and Rackoff (1989)) and secure 2-party computation (introduced by Yao (1982)) computationally-bounded players jointly compute a function of their respective private information, while maintaining the privacy of this information. However, in many economic applications information cannot be generically encoded and must be seen by the opponent to be verified, or it is not feasible to run a cryptographic protocol, which renders these tools inapplicable to our setting.

I. The Model

PLAYERS AND THEIR TYPES. — There are two players $\{1,2\}$, which we typically denote by $i \in \{1,2\}$ and $j \stackrel{\text{def}}{=} 3-i$, each of which has a type τ_i. Each player possesses a unit of evidence. This evidence can potentially lead to a successful project, in which case we call the player *viable* ($\tau_i = V$), or not, in which case we call the player *unviable* ($\tau_i = U$). The type of a player is her private information. The prior probability that a player is viable is p and is independent across players.

Evidence is to be interpreted as a code, a recipe, or a proof that takes a fixed amount of time (or space) to transmit, but can be divided into smaller pieces, each of which can be transmitted in correspondingly less time (or space).

GAME. — There are possibly infinitely many rounds of communication. In each round, one player is called upon to speak, and this happens in an alternating fashion with player 1 moving first. In each round t, the speaking player i chooses the amount of new evidence to disclose in that round. Formally, let N_t^i be the amount of evidence disclosed by player i up to round t (including t). In each round t, the speaking player i chooses $N_t^i \in [N_{t-2}^i, 1]$, where $N_{-1}^i = N_0^i \stackrel{\text{def}}{=} 0$. That is, we are assuming that a player can withhold evidence, but cannot withdraw evidence already disclosed.

We want the model to capture the idea that a viable type needs to reveal the entire proof, code, or recipe in order to prove its viability. Unviable types are those whose proofs or recipes are incomplete or contain a fatal flaw. Hence, after an unviable player reveals a sufficient amount of the evidence, her type becomes known to the opponent. To this end, we assume that the unviable type of player i is characterized by a number $K^i \in (0,1)$, which is her private information. If in some round t, the unviable type with K^i reveals $N_t^i > K^i$, player j receives a signal that the opponent is unviable: $s_t^i = U$. We assume that K^i is distributed with a strictly increasing, continuous distribution function $F(K^i)$ and is independent of K^j. Note that we are implicitly assuming that an unviable player cannot fabricate the evidence of a viable one.[3]

A history of play up to and including round t is $H_t = \{\{N_1^1, N_2^2, ..., N_t^i\}, s_{t-1}^j, s_t^i\}$, where $s_t^i \in \{\emptyset, U\}$ and $s_t^i = U$ means that player j received a signal that i is unviable and $s_t^i = \emptyset$ means that player j did not receive this signal. A pure strategy of player i is a function that for each t in which i speaks maps H_{t-1} into $N_t^i \in \left[N_{t-2}^i, 1\right]$.

We allow the players to split evidence into arbitrarily small pieces, but we want the number of such pieces to be finite. Hence, we place the following assumption on the set of strategies available to the players:

ASSUMPTION I.1: *For any pair of strategies (σ_1, σ_2), there exists some $T(\sigma_1, \sigma_2)$ such that for every $t > T(\sigma_1, \sigma_2)$, the history H_t generated by strategies (σ_1, σ_2) has the property that $N_t^i = N_{t-2}^i$ for $i \in \{1, 2\}$.*

This assumption states that no matter how the game unravels, players will stop revealing new evidence after a finite number of rounds.[4]

For fixed strategies of the players, denote by $N^i = \max_t N_t^i$ the largest amount of evidence revealed in the game by i. By Assumption I.1, N^1 and N^2 exist.

[3]Note that K^i can be also interpreted as the highest amount of evidence that the unviable type possesses; for example, the number of lemmas proved. In this case after revealing K^i amount of evidence, player i would be unable to reveal any further evidence. We find it convenient to assume that i can continue revealing evidence, but since this evidence does not enter the payoffs, our assumption is without loss of generality.

[4]Without this assumption, to complete the specification of the game we would have to specify the payoffs from communicating in infinitely many rounds. These payoff would have to be sufficiently low to guarantee that the uninteresting behavior of never exchanging all evidence is not an equilibrium. We use Assumption I.1 in Propositions II.1 and II.3.

PAYOFFS. — There exists a project that pays $v > 0$ to each player if and only if both players are viable and share all their respective evidence: that is, if $\tau_1 = \tau_2 = V$ and $N^1 = N^2 = 1$. In this case, we say that players *cooperate* on the project.

In addition to the payoff from the project, players obtain payoffs from the evidence exchanged in the game. We assume that only the evidence provided by the viable types is valuable, as we believe that this assumption is consistent with our motivating examples. We discuss the robustness of our results to this assumption in Section III. A viable player i who reveals N^i suffers a disutility $h\left(N^i\right)$, and a player i who receives N^j from a viable opponent benefits $g\left(N^j\right)$. Formally, the utilities of the viable and unviable types, respectively, are:

$$u_i^V = v\mathbf{1}_{(\tau_j=V,\ N^1=N^2=1)} + g\left(N^j\right)\mathbf{1}_{(\tau_j=V)} - h\left(N^i\right), \tag{1}$$

and

$$u_i^U = g\left(N^j\right)\mathbf{1}_{(\tau_j=V)}, \tag{2}$$

where the symbol $\mathbf{1}_{(.)}$ denotes the indicator function. Both g and h are strictly increasing, continuous, and take value 0 at 0. We normalize $h(1) = 1$.

EQUILIBRIUM. — The solution concept is a pure-strategy weak Perfect Bayesian Equilibrium (PBE).[5]

Throughout the paper, we will primarily be interested in optimal modes of communication; in particular, in whether it is optimal for each player to reveal all evidence at once, or whether splitting evidence into finer pieces and revealing them in an alternating fashion can improve welfare. Our optimality criterion will be the joint payoff of the viable types. If our game is part of a larger game in which players make costly investments to become viable, then maximizing the payoff of the viable types can be consistent with providing the largest incentives for such investments. However, in Section III.A we discuss how our results would extend if the goal were to maximize the payoff of the unviable types or the total payoff of all types.

COMMENTS. — In our model, the evidence exchanged is instrumental to the project: The project cannot be undertaken unless all evidence is exchanged. Moreover, once information is exchanged, cooperation happens automatically. We find this assumption reasonable in a variety of settings. For example, the research project cannot be completed unless all ideas are put down in the paper, and as soon as they are written down, the decision on whether to submit the project for

[5]Restricting attention to pure strategies is with little loss of generality. We explain why this is the case after we state Proposition II.3.

publication or shelve it is trivial. However, in a variety of settings the decision about cooperation may be undertaken even before all information is exchanged. For example, consider two firms entertaining a merger. They exchange proprietary information that includes financial statements, contracts with other firms, and revenues; they visit each other's factories; and most generally, they "open the books" to each other. However, they can commit to the merger even before they exchange all information, that is, before they are certain that the opponent is viable and the merger will result in synergies. Similarly, the information that criminals exchange may only serve to signal that they are not undercover cops, and they may engage in a successful criminal venture even without exchanging this information. Situations like these can be easily mapped into our model with one modification: One has to endogenize the decision of cooperation and hence the amount of information exchanged. It is straightforward to show, however, that if the privacy concerns are smaller than the disutility from engaging in a venture with an unviable opponent, players will exchange all information before they decide to cooperate. In such cases all our results will continue to hold.

A few comments on our assumptions are in place. First, if players had no privacy concerns, all projects could be undertaken with only a two-round evidence exchange—player 1 would reveal all her evidence in round 1, and then player 2 would reveal all her evidence in round 2. All equilibria leading to the project would deliver the same welfare.

Second, in the current model only the amount of evidence revealed, and not its order, matters. Such a modeling assumption is clearly appropriate if there is only one feasible order of evidence exchange (e.g., subsequent lemmas feed on the previous ones) or if a player cannot distinguish ex ante between different pieces of the opponent's evidence, and hence is unable to require them in any particular order. However, our model is more general: If each player can require the opponent to reveal her evidence in a particular order, then once this order is agreed upon, it results in some F, h, and g, and our analysis follows. We will discuss this further in Section II.B.

And finally, the problem of privacy could be easily solved if an expert benevolent mediator were available. However, in many circumstances a mediator with expertise sufficient to judge the viability of the presented evidence is unlikely to be benevolent.

II. Analysis

A. Preliminaries

Our first proposition states that we can divide all equilibria into two categories.

PROPOSITION II.1: *In any equilibrium, either*

a. the players cooperate on the project with probability 1 *when* $\tau_1 = \tau_2 = V$*, or*

b. viable types reveal no evidence.

We delegate most proofs to the Appendix, but provide the proof of Proposition II.1 here as it is short and conveys the intuition.

PROOF: Since we consider pure strategy equilibria, if the project is undertaken with some probability when players are viable, this probability must be 1. Suppose then that there exists an equilibrium in which the project is never undertaken. By Assumption I.1, there is a last round $T(\sigma_i, \sigma_j)$ in which a viable type reveals new evidence. Suppose without loss of generality that it is the viable type of player i. Then in $T(\sigma_i, \sigma_j)$, player i knows that if she reveals the piece of evidence prescribed by the equilibrium, she suffers a disutility from that, and in return she can at most receive evidence from the unviable type, which is not valuable. Hence, not revealing any evidence at $T(\sigma_i, \sigma_j)$ and beyond is a profitable deviation.

We will call the equilibria from part (a) *cooperating* and the equilibria from part (b) *non-cooperating*.

LEMMA II.2: *Any cooperating equilibrium strictly Pareto dominates any non-cooperating equilibrium.*

The intuition for Lemma II.2 is simple: Since viable players always have the option to reveal no evidence, they must be better off in any equilibrium in which they willingly reveal evidence, and since information revelation imposes positive externalities on the other player, both players must be better off.

Given Lemma II.2, we will henceforth focus on cooperating equilibria. The following proposition outlines the most important aspects of all such equilibria.

PROPOSITION II.3: *In any cooperating pure strategy PBE there exists $T > 0$ and a sequence $\left\{\bar{N}_t\right\}_{t=1}^{T}$ with $\bar{N}_{T-1} = \bar{N}_T = 1$ such that in each t,*

a. after any H_{t-1} with $\{N_1^1, N_2^2, ..., N_{t-1}^i\} = \{\bar{N}_1, \bar{N}_2, ..., \bar{N}_{t-1}\}$ and $s_{t-2}^j = s_{t-1}^i = \emptyset$, all viable types of j and all unviable types of j with $K^j > \bar{N}_t$ reveal $\bar{N}_t$;

b. if $N_{t-1}^i < \bar{N}_{t-1}$, then the viable type j reveals $N_t^j = \bar{N}_{t-2}$;

c. if $s_{t-1}^i = U$, then the viable type j reveals no new evidence in the game.

If $pv - h(1) \geq 0$, then a cooperating equilibrium exists, and any sequence $\left\{\bar{N}_t\right\}_{t=1}^{T}$ with $\bar{N}_{T-1} = \bar{N}_T = 1$ can be supported as a cooperating equilibrium.

Proposition II.3 fully characterizes behavior on the equilibrium path: Players adhere to the prescribed sequence of evidence revelation as long as no one has deviated (part *a*). As soon as one player deviates (unless she deviates to revealing

everything and proves that she is viable), the exchange of valuable evidence stops (parts b and c).[6] Note that the proposition does not characterize the strategies of the players off the equilibrium path, as those may vary across equilibria.

It should not come as a surprise that many sequences $\left\{\bar{N}_t\right\}_{t=1}^{T}$ can be supported as cooperating equilibria. This is simply because for a given sequence, adherence to this sequence can be enforced by off-equilibrium beliefs that consider every deviation (except possibly to $N_t^i = 1$) as coming from the unviable type. The only constraint that $\boldsymbol{N}_T$ must satisfy is that every viable player prefers to adhere to the sequence instead of walking away after some round t with the evidence received from the opponent. The condition $pv - h(1) \geq 0$ assures that at the beginning of the game the expected benefit from cooperation is larger than the highest privacy loss possible. Since the posterior on opponent's viability only goes up over time (as long the opponent has not proved herself unviable), this condition assures that in any round, the expected benefit from adhering to any remaining sequence of evidence exchange will be bigger than the expected privacy loss.

Note that standard equilibrium refinements such as Intuitive Criterion, Divinity or Universal Divinity would not restrict the set of equilibria here. To see this note that for any possible deviation, there exist some unviable types with sufficiently high K^i for whom this deviation is costless. If there is a deviation that is profitable to a viable type, this deviation must provide her with useful evidence, which means that this deviation strictly benefits some unviable types as well.

From now on, we will identify each cooperating equilibrium with its corresponding sequence $\boldsymbol{N}_T \overset{\text{def}}{=} \left\{\bar{N}_t\right\}_{t=1}^{T}$ of evidence revelation. Since we are interested in whether more incremental—and hence taking place in more rounds—communication is beneficial, we will restrict our attention to sequences $\boldsymbol{N}_T$ for which $\bar{N}_t \neq \bar{N}_{t-2}$ for all $t \leq T$. Such a restriction is without loss of generality and allows us to treat T as a measure of how incremental different sequences of evidence exchange are.[7]

We conclude this section by deriving the payoffs of the viable players as a function of $\boldsymbol{N}_T$. Consider an equilibrium $\boldsymbol{N}_T$, and observe that if the viable type of player i faces another viable type, the two will exchange all evidence and receive $v + g(1) - h(1)$ each. If, however, a viable type of player i faces an unviable type, she will stop revealing evidence—and hence stop incurring disutility from privacy losses—as soon as $\bar{N}_t^j > K^j$. This is because in such a round t, either she will

[6] To be precise, part b does not say that the exchange of valuable evidence stops completely after any deviation from $\left\{\bar{N}_t\right\}_{t=1}^{T}$, but in the appendix we show that it stops after all deviations from this sequence that happen on the equilibrium path.

[7] At this point it should be relatively clear why allowing for mixed strategies might expand the set of optimal equilibria, but there are no mixed strategy equilibria that deliver a strictly higher welfare than all pure strategy equilibria. To see this, suppose that σ is a mixed strategy equilibrium profile that delivers the highest welfare and suppose that both players happen to be viable. Then σ delivers a distribution over a set of outcomes, where each outcome is associated with some sequence $\boldsymbol{N}_T = \left\{\bar{N}_1, \bar{N}_2, \bar{N}_3, ..., \bar{N}_{T-1} = 1, \bar{N}_T = 1\right\}$. Pick $\boldsymbol{N}_T$ that delivers the highest welfare. Under our assumption that $pv \geq 1$, one can construct a pure strategy equilibrium with $\boldsymbol{N}_T$.

receive the signal that her opponent is not viable, or her opponent will deviate from the prescribed sequence $\boldsymbol{N}_T$. The probability that the unviable opponent has $K^j \in (\bar{N}_{t-2}, \bar{N}_t)$ is equal to $F(\bar{N}_t) - F(\bar{N}_{t-2})$. Thus, the expected utility of a viable player 1 is

(3)
$$E[u_1^V(\boldsymbol{N}_T)] = p\left(v + g\left(1\right) - h\left(1\right)\right) - (1-p)\sum_{k=0}^{K}(F(\bar{N}_{2k+2}) - F(\bar{N}_{2k}))h\left(\bar{N}_{2k+1}\right),$$

where $K = \frac{T-2}{2}$ if T is even and $K = \frac{T-3}{2}$ if T is odd. The expected utility of a viable player 2 is derived similarly.

Equation 3 reveals that differences in payoffs across cooperating equilibria come only from differences in the expected disutilities from evidence revealed to unviable types. It will be convenient to denote this disutility

$$\Psi_1(\boldsymbol{N}_T) \stackrel{\text{def}}{=} \sum_{k=0}^{K}(F(\bar{N}_{2k+2}) - F(\bar{N}_{2k}))h\left(\bar{N}_{2k+1}\right), \tag{4}$$

and call it the *privacy leakage* of player 1. Similarly, denote by $\Psi_2(\boldsymbol{N}_T)$ the privacy leakage of player 2. Hence, the equilibria that maximize the joint payoff of the viable types $E[u_1^V(\boldsymbol{N}_T)] + E[u_2^V(\boldsymbol{N}_T)]$ are those that minimize the joint privacy leakage

$$\Psi_1(\boldsymbol{N}_T) + \Psi_2(\boldsymbol{N}_T) = \sum_{t=1}^{T-1}(F(\bar{N}_{t+1}) - F(\bar{N}_{t-1}))h\left(\bar{N}_t\right).^{8} \tag{5}$$

Note that in the shortest cooperating equilibrium, $\boldsymbol{N}_2 = \{1, 1\}$, a viable player 1 reveals all her evidence in the first round. Hence, in the second round, the viable type of player 2 knows which type she is facing. If she faces the unviable type she reveals nothing. If she faces the viable type, she reveals 1 unit of evidence and cooperation on the project is successful. Hence, $\Psi_1(\boldsymbol{N}_2) = h(1) = 1$ and $\Psi_2(\boldsymbol{N}_2) = 0$ independent of the particular shape of F and h. We call this equilibrium *simple*, and any other cooperating equilibrium *incremental*.

B. Simple Versus Incremental Evidence Exchange

Before we state the first result, it is useful to understand that, in our model, the value of evidence is two-fold. First, evidence has an *intrinsic* value—the

[8]Note that equation 3 reveals that it actually does not matter whether the privacy cost is born always or only if cooperation is not achieved. In the latter case, $g(1) - h(1)$ disappear from the first part of the equation, but the privacy leakage remains the same.

actual content that makes it relevant for the success of the project—such as the description of the project design, the relevant computer code, or a proof. The more evidence a viable player reveals, the higher its intrinsic value. The intrinsic value of evidence is measured by h, and it counts as a loss for the viable player who reveals it. However, the evidence revealed by a player also carries information about the type of this player: The more evidence a player reveals, the more likely the opponent is to believe that she is the viable type. This *extrinsic* value of evidence is measured by F. Whenever a player is called upon to reveal $\bar{N}_t$, the viable type of this player suffers from the intrinsic value lost, but the viable type of the other player benefits from the extrinsic value gained.

It turns out that what matters for the optimality of equilibrium exchange is the precise relationship between the intrinsic and extrinsic values of evidence. To summarize this relationship, it is convenient to use the following change of variables

$$M \overset{\text{def}}{=} h(N) \tag{6}$$

and define $\phi(M) \overset{\text{def}}{=} F(h^{-1}(M))$. M measures the units of intrinsic value contained in an amount N of evidence. The expression $\phi(M)$ then measures the extrinsic value associated with M units of intrinsic value.

We are now ready to state the conditions under which the optimal equilibria are simple.

PROPOSITION II.4: *Suppose that $pv > h(1)$.*

a. *If $M = \phi(M)$ for all $M \in (0,1)$, then all cooperating equilibria deliver the same joint payoff to the viable types.*

b. *If $M > \phi(M)$ for all M, then the unique cooperating equilibrium that maximizes the joint payoff to the viable types is the simple one.*[9]

For intuition, suppose that at a certain stage of communication player 1 has revealed less evidence than her opponent. Who should reveal new evidence next? In (a), $\phi(M)$ is assumed to be linear in M, and so revealing an additional unit of intrinsic value always delivers one additional unit of extrinsic value. Hence, it does not matter which player reveals new evidence next – the total gain/loss from the next piece of evidence will be the same. This implies that any mode of evidence exchange is optimal.

To understand the intuition behind part (b), it is easier to focus on the case in which $\phi(\cdot)$ is strictly convex, which is a sufficient condition for $M > \phi(M)$. Under strict convexity, any unit of intrinsic value revealed by a player delivers less extrinsic value than each additional unit. This implies that a unit of intrinsic value revealed by the player who lags in the exchange carries less extrinsic value

[9] In fact, any sequence in which one player reveals all her evidence first is an optimal equilibrium, but recall that we are assuming that there are no "silent" stages.

than a unit revealed by the player who is ahead. This means that as soon as player 1 reveals some evidence, it is optimal to ask her to reveal the rest of her evidence before player 2 speaks. This, in turn, implies simple evidence exchange.

When $M > \phi(M)$ but $\phi(\cdot)$ is locally concave, it is not longer true that at any stage of the exchange it is optimal to ask the player who is ahead to reveal the rest of her evidence. However, since the revelation of all evidence carries the same intrinsic and extrinsic value (by the normalization $\phi(1) = 1$), and the revelation of $N < 1$ units of evidence delivers less extrinsic value than its intrinsic value, the result still holds.

The intuition above immediately suggests that when $\phi(\cdot)$ is concave, incremental evidence exchange should be optimal. In this case, a unit of intrinsic value revealed by the player who lags in the exchange carries more extrinsic value than a unit revealed by the player who is ahead. Hence, optimality requires that in equilibrium the player who lags in evidence exchange is the one who reveals an additional unit of evidence. This implies that evidence should be exchanged in turns, and the pieces exchanged should be as small as possible. Proposition II.5 below formalizes this intuition.

PROPOSITION II.5: *Suppose that $pv > h(1)$.*

a. *If $M < \phi(M)$ for some M, then there exists an equilibrium that delivers a higher total payoff to the viable types than the simple equilibrium.*

b. *Suppose that $\phi(\cdot)$ is strictly concave, and consider a cooperating equilibrium $\boldsymbol{N}_T$. Then there exists another cooperating equilibrium with $T+1$ rounds of communication that delivers a strictly higher total payoff to the viable types.*

What can we say about optimal equilibria in setting where $M < \phi(M)$ for some but not all M? Let $c_\phi = \inf\{f \mid f \text{ is a concave function and } f \geq \phi\}$. That is, c_ϕ is the lowest concave function that lies above ϕ. Let $C = \{M \in [0,1] \text{ such that } \phi(M) = c_\phi(M)\}$. The following proposition characterizes optimal equilibria for any ϕ:

PROPOSITION II.6: *Let $\mathbf{M}_T$ be an optimal equilibrium. Then $M_t \in C$ for all t. Moreover, if $\phi(\cdot)$ is strictly concave for some subset of C, then for any equilibrium $\boldsymbol{M}_T$, there exists another cooperating equilibrium with $T+1$ that delivers a strictly higher total payoff to the viable types.*

We illustrate Proposition II.6 using an example in Figure 1.

In Figure 1, $\phi(\cdot)$ is the thin solid curve and $c_\phi(\cdot)$ is the thick dashed curve. As we see, only the initial evidence lies in C. Hence, only the evidence lying in C will be split in smaller amounts, and since $\phi(\cdot)$ is strictly concave there, this evidence will be split as finely as possible. Outside of C, it is optimal to reveal $1 - |C|$ at the end of the exchange in one round.

The general intuition for Proposition II.6 is as follows. In the proof of Proposition II.6 we show that we can restrict our attention to equilibria in which in

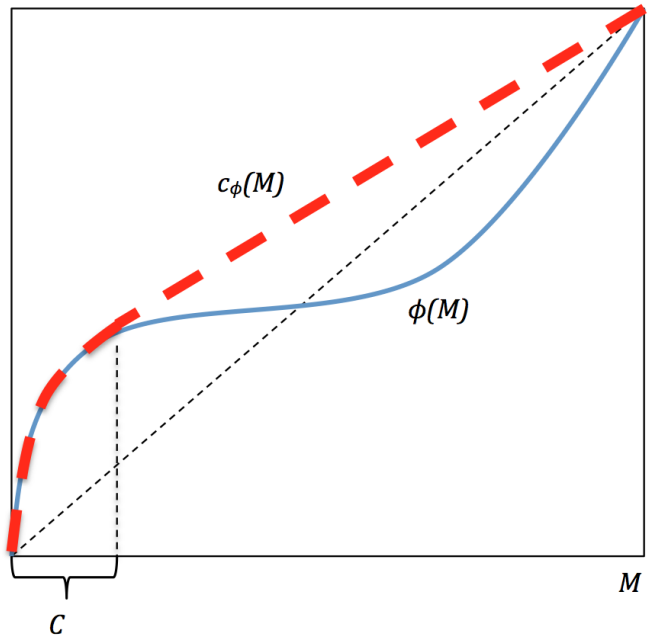

FIGURE 1. OPTIMAL EQUILIBRIA.

each round player 2 reveals exactly the same amount of evidence as player 1. Suppose that in some odd round, player 1 is to reveal $M' - M$. The ratio of the extrinsic value of this amount of evidence to its intrinsic value is $\frac{\phi(M')-\phi(M)}{M'-M}$. Suppose now that the players consider splitting this evidence into two smaller amounts, $M'' - M$ and $M' - M''$, with the corresponding ratios $\frac{\phi(M'')-\phi(M)}{M''-M}$ and $\frac{\phi(M')-\phi(M'')}{M'-M''}$. If the first ratio is lower than the second, then the first amount of evidence is less useful in terms of its extrinsic value than the second. Hence, it is better to ask one player to reveal both amounts first, before asking the opponent to reveal her less useful first amount. Formally, the first ratio is lower than the second if the concavification of $\phi(M)$ lies strictly above M''; hence, the proposition follows. The result that making exchange more incremental is beneficial when $\phi(M)$ is stricly concave follows from the same argument as in part (b) of Proposition II.5.

COMMENTS ON THE SHAPE OF THE UTILITY FUNCTION. — We have established that the nature of the optimal evidence exchange depends crucially on the shape of $\phi(\cdot)$. The shape of $\phi(\cdot)$ is an empirical question, but we would like to develop intuition for when it is likely that players use incremental exchange.

Let us start by assuming that $F(K) = K$. In such a case, $\phi(\cdot)$ is concave when h is convex. We should expect h to be convex if the proprietary evidence is of

little value unless a large quantity of it is obtained. Suppose now that h is linear. Then, $\phi(\cdot)$ is concave if F is concave. Recall that K^i is interpreted as the smallest amount of evidence of i that allows j to verify that i is unviable. We expect F to be concave in environments in which most of the invalid evidence can be spotted quickly (small K^i). Hence, we expect the evidence exchange to be incremental if the initial pieces of information are more likely to reveal the viability of the opponent and less able to inflict harm due to privacy loss than the subsequent pieces. Otherwise, we should expect simple evidence exchange.

Despite our results, our intuition tells us that in most applications the evidence exchange takes incremental form. One reason for this discrepancy between our findings and casual observations may be that in actual situations players may have flexibility as to the order in which evidence is presented. In such cases, they may order it in a socially optimal way. Since welfare is higher in the environments of Proposition II.5 than of Proposition II.4, players will find it beneficial to order evidence in such a way that incremental exchange is optimal.[10]

In other applications, however, the order in which evidence is revealed may be fixed by the nature of evidence. For example, revealing a proof may require revealing its steps in a predetermined order, as the steps may build upon themselves. However, even in such applications observing simple communication may be unlikely. The reason for this is that the simple equilibrium is the least equitable of all: The first player reveals all her evidence at once, while the second player incurs no privacy loss. If players are concerned with some ex-post notion of equitability, or if they negotiate which equilibrium to play using some outside options as threat points, they may not select the optimal equilibrium. Hence, a lesson that one may draw from Proposition II.4 is that in environments with $M > \phi(M)$, equitability comes at the cost of efficiency.

C. Properties of Optimal Incremental Evidence Exchange

Proposition II.5 implies that when $\phi(\cdot)$ is strictly concave, the optimal mode of equilibrium exchange does not exist. For any number of rounds T, players would always benefit from splitting evidence even more finely. Since the model is only an abstraction of actual situations, one might conjecture that in reality there is a limit to the number of rounds in which players can engage. Hence, one may ask how players would split evidence in such situations. The next proposition shows that, generically, the evidence will not be split into equal pieces, but, depending

[10]Since evidence is continuous in our model for tractability reasons, we cannot formally talk about "ordering pieces of evidence", but our results would extend to a discrete model. To see this, suppose that there are 2 pieces of evidence, a and b, and for simplicity, they have the same intrinsic value, but an unviable type is more likely to possess (or fake) a (probability α) than b (probability β). Unviable types possess at most 1 of the pieces. The corresponding function $F(\cdot)$ is as follows: For both orders, $F(0) = 1 - \alpha - \beta$ and $F(2) = 1$. If the order is $\{a, b\}$, then $F(1) = \alpha$, while if the order is $\{b, a\}$, then $F(1) = \beta$. Propositions II.4 and II.5 imply that welfare will be higher if $F(\cdot)$ is concave, i.e., if the order is $\{a, b\}$. A quick glance at the proofs of Proposition II.4 and the first part of Proposition II.5 should convince the reader that the above argument can be made formal.

on the fine details of the $\phi(\cdot)$ function, it will be revealed either in increasing or decreasing increments (in terms of intrinsic information).

PROPOSITION II.7: *Consider equilibria of length T, and suppose that $\phi(\cdot)$ is differentiable and strictly concave. If $(\phi(\cdot))'$ is (weakly) concave/convex, then in the equilibrium that maximizes the joint welfare of the viable types, the sequence $\left\{h(\bar{N}_t) - h(\bar{N}_{t-1})\right\}_{t=1}^{T}$ is (weakly) increasing/decreasing.*

Hence, if $(\phi(\cdot))'$ is concave, then the amount of intrinsic information revealed in each round should increase as the communication progresses. In this case it is as if players "build trust" in the initial rounds, and once this trust is built, they exchange evidence more freely. In the other case, players become more "cautious" towards the end.

The intuition for Proposition II.7 is as follows. Recall that if $\phi(\cdot)$ is strictly concave, then the extrinsic value of an additional unit of intrinsic value revealed by the player who lags is higher than the one revealed by the player who leads. Hence, an equilibrium is optimal if in each round, the difference in the amount of evidence revealed by the players is small. For a finite T, however, there is a limit to how small this difference can be, but one can make it larger in some rounds and smaller in others. It is crucial then that the rounds in which this difference is high coincide with the rounds in which the difference between the extrinsic value of an additional unit of intrinsic value revealed by the lagging player and the leading player is the smallest. And when $(\phi(\cdot))'$ is concave, this difference is the smallest at higher levels of $h(\bar{N}_t)$, that is, in later rounds. Hence, in later rounds players can split information less finely.

We conclude this section by illustrating our results using an example of a relatively simple family of functions: $h(N) = F(N)^{\alpha}$, where $\alpha > 0$ is a real number. In this family, $\phi(M) = M^{\frac{1}{\alpha}}$. When $\alpha < 1$, then $\phi(\cdot)$ is convex, and by Proposition II.4 we know that simple equilibrium is optimal, delivering privacy leakage of 1. When $\alpha > 1$, then $\phi(\cdot)$ is concave, and hence incremental exchange is better. Moreover, $(\phi(\cdot))'$ is strictly convex, hence for any finite T, it is optimal to split information in decreasing increments: The first piece of evidence exchange will have higher intrinsic value than each subsequent one. And finally, for this simple family of functions, we can calculate the welfare gain from incremental exchange:

PROPOSITION II.8: *Suppose $h(N) = F(N)^{\alpha}$ (or equivalently, $\phi(M) = M^{\frac{1}{\alpha}}$) for some real number $\alpha > 1$. Then in the optimal equilibrium among those with T rounds, it holds that $\Psi^i(\boldsymbol{N}_T) + \Psi^j(\boldsymbol{N}_T) \to 2/(\alpha+1)$ as $T \to \infty$.*

Hence, for higher α—which corresponds to a more concave $\phi(\cdot)$—the welfare gain from incremental communication is higher. In the limit as $\alpha \to \infty$, evidence that has very little intrinsic value carries a lot of extrinsic value; hence, in this case one can avoid privacy leakage almost entirely.

III. Extensions

A. Other Optimality Criteria

In this section we discuss how our results extend under other optimality criteria, namely the joint payoff of all types and the joint payoff of the unviable types.

Note first that when both players are viable or both players are unviable, the details of evidence exchange do not matter. This is because the viable types end up cooperating no matter what exchange protocol they follow, and the unviable types do not gain any valuable evidence. Hence, the mode of communication matters only if one player is viable and the other is not.

With probability p, player 2 is unviable. In this case, player 1's privacy leakage is described by the equation for $\Psi_1(\cdot)$ (equation 4). Player 2's privacy leakage (which is a gain in this case) is derived solely from the evidence revealed by player 1; hence, her privacy leakage will be like in equation (4) but with $(-g)$ in place of h. So the total privacy leakage will be like in (4) but with $h(N) - g(N)$ in place of $h(N)$. One can derive the privacy leakage in the case when player 1 is unviable similarly. Summing them, we obtain that the total expected privacy leakage is

$$\sum_{t=1}^{T-1} \left(F(\bar{N}_{t+1}) - F(\bar{N}_{t-1})\right) \left(h(\bar{N}_t) - g(\bar{N}_t)\right). \tag{7}$$

Maximizing total welfare of all types is thus equivalent to minimizing (7). By the same argument, maximizing the joint payoff of only the unviable types requires minimizing

$$\sum_{t=1}^{T-1} \left(F(\bar{N}_{t+1}) - F(\bar{N}_{t-1})\right) \left(-g(\bar{N}_t)\right). \tag{8}$$

Hence, all our results still hold with the caveat that the conditions in the propositions need to be placed on the function $F(w^{-1}(\cdot))$ instead of $\phi(\cdot)$, where $w(N) \stackrel{\text{def}}{=} h(N) - g(N)$ when we maximize the total payoff and $w(N) \stackrel{\text{def}}{=} -g(N)$ when we maximize the total payoff of the unviable types.

Note that evidence exchange is socially beneficial if $h(N) - g(N) < 0$. Using $w(N) \stackrel{\text{def}}{=} h(N) - g(N)$, our model can be used to characterize when incremental exchange is socially beneficial in such settings.

B. All Types Have Privacy Concerns

So far we have assumed that the unviable types do not have privacy concerns. Even though this is likely to hold in most of our examples, one can easily imagine situations in which this is not true. The following provides one such example.

EXAMPLE III.1: *Two firms are contemplating a merger based on the perceived potential synergies. To complete the merger, they have to share all private information; e.g., their financial statements, accounting procedures, initiated investments, corporate culture. Each firm i privately knows whether it satisfies conditions for the synergies to be realized, and if it does not, then this becomes apparent to firm j after i reveals K^i amount of information. In this example, both firms may be concerned about their privacy independently of their type: In the event of the merger not occurring, both firms can use the acquired information in the marketplace.*

In this section, we discuss how our results would change if we allowed all types to have privacy concerns.

Let $h_{\tau_i}\left(N^i\right)$ be the disutility that player i of type τ_i suffers if she reveals N^i, and let $g_{\tau_j}\left(N^j\right)$ be the utility that player i receives if she obtains N^j from her opponent of type τ_j. The payoffs may depend on whether the source of evidence is U or V:

$$u_i^V = v\mathbf{1}_{(\tau_j=V,N^1=N^2=1)} - h_V\left(N^i\right) + g_V\left(N^j\right)\mathbf{1}_{(\tau_j=V)} + g_U\left(N^j\right)\mathbf{1}_{(\tau_j=U)}, \tag{9}$$

and

$$u_i^U = -h_U\left(N^i\right) + g_V\left(N^j\right)\mathbf{1}_{(\tau_j=V)} + g_U\left(N^j\right)\mathbf{1}_{(\tau_j=U)}. \tag{10}$$

As before, all functions are continuous, take value 0 at 0, g_V and h_V are strictly increasing, and we normalize $h_V(1) = 1$. Until now, we were assuming that $h_U \equiv g_U \equiv 0$.

It should be clear that Proposition II.1 and Lemma II.2 still hold. We show in the appendix that the behavior outlined in parts a, b, and c of Proposition II.3 holds with possibly one exception: The unviable types do not have to adhere to the sequence $\mathbf{N}_T$ (the second part of a). They follow the sequence only up to a certain round, after which they reveal no new evidence. This is because the unviable types now have privacy concerns as well, and therefore adhering to the equilibrium sequence may result in too much privacy loss for them to be optimal.

In the presence of privacy concerns for the unviable types, making evidence exchange more incremental has an additional effect: If the difference between the amount of information revealed in each round is small enough, some of the unviable types may find it optimal to stop revealing evidence early in the game (possibly in their first round). This decreases the information leakage of the viable types; hence, it should be beneficial. Indeed, Proposition III.2 below states that when the unviable types have any privacy concerns, the simple equilibrium is *never* optimal.

PROPOSITION III.2: *Suppose that h_U is strictly increasing, and the simple equilibrium exists. Then there exists $\bar{N}_1 \in (0,1)$ and $\bar{N}_2 \in (0,1)$ such that $\boldsymbol{N}_4 = \left\{\bar{N}_1, \bar{N}_2, 1, 1\right\}$ strictly Pareto dominates (in terms of the payoffs of the viable types) the simple equilibrium.*

Note that if $\bar{N}_1$ is sufficiently large and $\bar{N}_2$ sufficiently small, the unviable types of player 1 and 2 do not reveal any evidence in the game, as the privacy loss from revealing $\bar{N}_1$ and $\bar{N}_2$, respectively, is larger than the possible evidence gain from $\bar{N}_2$ and $1-\bar{N}_1$, respectively. Hence, the viable type of player 2 knows the type of her opponent already in the second round. As a result, she reveals no evidence to the unviable opponent, obtaining the same privacy leakage of 0 as in the simple equilibrium. Similarly, the viable type of player 1 knows the type of her opponent in the third round; hence, her privacy leakage is only $h_V(\bar{N}_1)$. And this is strictly less than the privacy leakage in the simple equilibrium, namely $\Psi_1(\boldsymbol{N}_2) = h_V(1)$.

Let us call all equilibria that completely discourage the unviable types from evidence exchange *screening*. In such equilibria, as in the equilibrium $\boldsymbol{N}_4$ of Proposition III.2, it is always the case that $\Psi_2(\boldsymbol{N}_T) = 0$. The crucial feature of the screening equilibrium $\boldsymbol{N}_4$ is that the difference between what a player has to reveal in a given round and what she expects to receive in the next round is small enough that the expected benefit from the next round does not outweigh the privacy loss incurred by participating in the current round. This suggests that by splitting information more finely, one can discourage the unviable types from participating using $\bar{N}_1$ lower than in $\boldsymbol{N}_4$, and hence achieving a lower privacy leakage $h_V(\bar{N}_1)$. The following Proposition III.3 confirms this intuition.

PROPOSITION III.3: *Let h_V be a strictly increasing and continuous function with $h_V(0) = 0$, and let $h_U(N) = bh_V(N)$ for some $b > 0$.*[11] *Then incremental communication improves the welfare guaranteed by screening equilibria:*

a. *If $\boldsymbol{N}_T$ is a screening equilibrium, then there exists $\boldsymbol{N}_{T+1}$ that is also a screening equilibrium with $\bar{N}_1(\boldsymbol{N}_{T+1}) < \bar{N}_1(\boldsymbol{N}_T)$. Moreover, for each fixed T, the lowest $\bar{N}_1$ in any screening equilibrium is a decreasing function of b.*

b. *Suppose $g_U \equiv g_V \equiv h_U \equiv h_V$. Then in the best screening equilibrium $\Psi_1(\boldsymbol{N}_T) \to 0$ as $T \to \infty$.*

Hence, by making communication more incremental, one can incentivize the unviable types to drop out of the evidence exchange at an increasingly smaller cost. This cost is smaller when the privacy concerns of the unviable types are larger. In the case in which the gain from evidence received is type-independent and equal to the privacy loss (case (b) in Proposition III.3), one can construct an equilibrium in which the payoffs of the viable types are arbitrarily close to the payoffs they would obtain if they had no privacy concerns.

[11] The assumption $h_U(N) = bh_V(N)$ can be relaxed, but this would make the proof less transparent.

Unfortunately, it is not trivial to show that screening equilibria are optimal. First, if g_U takes large values, the viable types may want the unviable types to stay in the game and reveal information. Second, to incentivize the unviable types to drop out, N_1 may need to be large. Hence, in situations in which the privacy loss from N_1 is significantly larger than from revealing a smaller amount of evidence, the viable types may be better off playing an equilibrium that is not screening but has smaller N_1. Deriving general conditions under which screening equilibria are optimal is left for future research.

C. Discounting

Our interpretation of real-world communication is that exchanging evidence always takes the same amount of time, independent of whether one player reveals all evidence first or players split evidence into smaller pieces that they then exchange in an alternating fashion. For that reason, we do not incorporate discounting into our model. However, if alternating evidence exchange requires more time or carries some other cost (e.g., cost of attention), then players have an incentive to minimize the number of rounds. In this case, the characterization of the optimal equilibria is less straightforward, but the general insights hold. Since discounting incentivizes players to have as few rounds as possible, the simple equilibria are still optimal when $\phi(\cdot)$ is convex. When $\phi(\cdot)$ is sufficiently concave, the optimal equilibria are still incremental, but unlike before, for each $\phi(\cdot)$, there exists a finite optimal number of rounds T.

Discounting also affects the analysis of Section III.B. The equilibrium $\boldsymbol{N}_4$ from Proposition III.2 is still an equilibrium, but any other incremental equilibrium that discourages all unviable types from exchange is not: Once both players are revealed viable, they have a strong incentive to reveal all remaining evidence at once. In such a case, one can still construct equilibria in which some unviable types drop out in the first two rounds, but one has to leave some unviable types with large K^i in the game to incentivize the viable types to reveal evidence incrementally.

IV. Conclusions

Our results indicate that optimal communication can be simple or incremental. When players can decide on the order in which to reveal information or when all players have privacy concerns, incremental evidence exchange dominates. In merger negotiations, for example, the order of topics discussed and documents disclosed is to some extent in the hands of the negotiating parties, and all players are likely to have privacy concerns. In criminal interactions, undercover agents may not have privacy concerns as they are only revealing information that is already likely to be known to the criminal underworld. The parties, however, are relatively unconstrained about what issues they should raise first. Our paper suggests that in such situations we should observe incremental communication.

On the other hand, when two agents with complementary expertise consider cooperation, the order in which they prove their expertise may be constrained by its nature. In this case, the optimal exchange depends on the fine details of the privacy concerns. For example, if simply revealing the purpose of the product that an innovator proposes to create or the premise of the experiment that a researcher proposes to run has a lot of value, it may be optimal that the innovator and the researcher prove the viability of their idea before the opponent contributes hers. Hence, simple exchange may be optimal.

Appendix

Proofs from Section II.A

A1. Proof of Lemma II.2

If there is no cooperating equilibrium, then the lemma is vacuously true. Suppose then that there exists a cooperating equilibrium that is weakly Pareto dominated by some non-cooperating equilibrium. Since by Proposition II.1, in the non-cooperating equilibria no valuable evidence is exchanged, and so the payoff of each viable type is 0. Hence, if a cooperating equilibrium is dominated by a non-cooperating one, then the payoffs in the former must be $E\left[u_1^V\right] \leq 0$ and $E\left[u_2^V\right] \leq 0$. First, it cannot be that $E\left[u_i^V\right] < 0$, as then player i would prefer not to reveal her evidence at all and obtain at least 0. Second, in any cooperating equilibrium $E\left[u_2^V\right] > 0$, as player 2 has an option to walk away from communication after receiving the first round of evidence, which with probability p is valuable. Hence, her expected payoff in the equilibrium must be strictly positive. Actually, the lemma is true even if we consider joint payoff of all types (as in Section III.A), as the unviable types can only benefit from evidence exchange.

A2. Proof of Proposition II.3

In what follows, p_t^i denotes the posterior belief held at the end of period t by j about i being the viable type. Let σ_1 and σ_2 denote the pure strategies of the viable type of players 1 and 2 in a cooperating equilibrium. Let $\left\{\bar{N}_t^1\right\}_{t\in\{1,3,\ldots\}}$ and $\left\{\bar{N}_t^2\right\}_{t\in\{2,4,\ldots\}}$ be the amount of information that the viable types of players 1 and 2 reveal on the equilibrium path if both players are viable and use σ_1 and σ_2. By Assumption I.1, viable types achieve cooperation in a finite number of rounds; hence, there exists t such that $\bar{N}_{t-1}^i = \bar{N}_t^i = 1$. Let T denote the smallest such t. By definition, in any cooperating equilibrium at any t, the viable type of the speaking player i must adhere to $\left\{\bar{N}_t^1\right\}_{t\in\{1,3,\ldots\}}$ and $\left\{\bar{N}_t^2\right\}_{t\in\{2,4,\ldots\}}$ as long as the opponent has adhered to it so far and $s_{t-2}^i = s_{t-1}^j = \emptyset$. Setting $\left\{\bar{N}_t\right\}_{t=1}^T = \{\bar{N}_1^1, \bar{N}_2^2, \bar{N}_3^1, \ldots\}$ proves the behavior of the viable types outlined in part (a).

Step A: If in some t, $p_t^i = 0$, then the viable type of j reveals no evidence in $t+1$. This is straightforward as given such beliefs, j expects only privacy leakage from continuing the evidence revelation.

Step B: If $s_t^i = U$, then the viable type j reveals no more evidence in the game. This comes directly from Bayes' rule and Step A. This proves part (c).

Step C: If in equilibrium the strategy of the unviable type of player i prescribes revealing $N_t^i \neq \bar{N}_t$, then if in round t the opponent j sees N_t^i, then she reveals no new evidence after t unless i reveals all her evidence and turns out to be viable. This follows directly from Bayes' rule and Step A.

Step D: To prove the behavior of the unviable types outlined in part (a), note that such types are indifferent between any amount of evidence they reveal, and strictly prefer to receive more evidence from the opponent. By Step C, if a strategy of j that does not adhere to $\left\{\bar{N}_t\right\}_{t=1}^T$ is an equilibrium strategy, it results in the opponent i withholding her evidence in the first round in which j deviates from $\left\{\bar{N}_t\right\}_{t=1}^T$. Since adhering to $\left\{\bar{N}_t\right\}_{t=1}^T$ as long as $K^j > \bar{N}_t^j$ makes the viable opponent reveal more evidence in $t+1$, a strategy that does not adhere to $\left\{\bar{N}_t\right\}_{t=1}^T$ as long as $K^j > \bar{N}_t^j$ cannot be an equilibrium.

Step E: To prove part (b), note that if $N_{t-1}^i < \bar{N}_t$ is on the equilibrium path, then part (b) follows from Step C. Suppose then that N_{t-1}^i is off the equilibrium path. Take the first round $t-1$ in which $N_{t-1}^i < \bar{N}_t$ is observed. Clearly, the beliefs of j at $t-1$ can be arbitrary. If they are $p_{t-1}^i = 0$, then part (b) follows from Step A. Suppose then that $p_{t-1}^i > 0$, and that the viable j's strategy is to continue revealing new evidence. Consider an unviable type with $K_t^i \in (\bar{N}_{t-2}, \bar{N}_t)$. If this type adheres to her equilibrium strategy, then in any case she obtains no more evidence from the opponent (either because of Step C or because her strategy requires revealing $\bar{N}_t$ which results in $s^i = U$). If she reveals N_t^i instead, she obtains some new evidence from the viable type of j. Hence, revealing N_t^i is profitable, which contradicts the assumption that N_t^i was off the equilibrium path.

Last claim

We will show now that if $pv \geq h(1)$, then any sequence $\left\{\bar{N}_t\right\}_{t=1}^T$ with $\bar{N}_{T-1} = \bar{N}_T = 1$ can be supported as the following cooperating equilibrium. Players adhere to the behavior outlined in parts (a), (b), and (c). Moreover, (d) as soon as $N_t^i \neq \bar{N}_t$, then the opponent (of either type) reveals no new evidence unless at some $\tau > t$, i reveals $N_\tau^i = 1$ and $s_\tau^i = \emptyset$; and (e) if at any t, $N_\tau^i = 1$ and $s_i = \emptyset$, then j (of either type) reveals $N_{\tau+1}^j = 1$. There may be equilibria in which the behavior described in (d) and (e) does not hold, but it is straightforward to see that they will be outcome equivalent to the equilibrium outlined here. And since we are proving existence, it is enough to prove the existence of one equilibrium.

The proof of the behavior described in parts (b) and (c), and the behavior of the unviable types described in (a) did not rely on the details of the sequence; hence, it will hold for any sequence. The behavior described in (d) and (e) is clearly

optimal for any sequence. We will now show that each player has an incentive to adhere to the behavior outlined in part (a) for any sequence.

Consider one such sequence. Suppose that the players followed this sequence up to (but excluding) round t, and that a viable type of j moves in t. By Bayes' rule, $p_{t-1}^i = \frac{p}{p+(1-p)\left(1-F\left(\bar{N}_{t-1}\right)\right)}$. If j adheres to the sequence, with probability p_{t-1}^i she will cooperate and with probability $\left(1-p_{t-1}^i\right)$ the evidence exchange will stop at some time $\tau > t$, the time at which the opponent reveals herself unviable. Hence, her expected payoff from adhering to the sequence is

$$\text{(A1)}\quad p_{t-1}^i\left(v+g\left(1\right)-h\left(1\right)\right)-\left(1-p_{t-1}^i\right)\sum_{k=\frac{t-1}{2}}^{K}\frac{F\left(\bar{N}_{2k+2}\right)-F\left(\bar{N}_{2k}\right)}{1-F\left(\bar{N}_{t-1}\right)}h\left(\bar{N}_{2k+1}\right),$$

where again $K=\frac{T-2}{2}$ if T is even and $K=\frac{T-3}{2}$ if T is odd.

If j deviates in t to revealing all her evidence, then her expected payoff is

$$p_{t-1}^i\left(v+g\left(1\right)-h\left(1\right)\right)-\left(1-p_{t-1}^i\right)1,$$

which is clearly lower than her payoff from not deviating. If she deviates to anything else, then by (d), she expects no more evidence exchange. Hence her best deviation is to reveal no new evidence at t. In such a deviation, she suffers disutility $h\left(\bar{N}_{t-2}\right)$, but with probability p_{t-1}^i she faces a viable type in which case she benefits from the evidence gained so far $g\left(\bar{N}_{t-1}\right)$. Comparing this to (A1) and using the formula for p_{t-1}^i, we obtain that she does not deviate if and only if

$$v+g\left(1\right)-h\left(1\right)\geq\frac{1-p}{p}\sum_{k=\frac{t-1}{2}}^{K}\left(F\left(\bar{N}_{2k+2}\right)-F\left(\bar{N}_{2k}\right)\right)h\left(\bar{N}_{2k+1}\right)+g\left(\bar{N}_{t-1}\right)-\frac{h\left(\bar{N}_{t-2}\right)}{p_t}.$$

Note that the summation on the right-hand side of the IC constraint is smaller than the privacy leakage in the entire game, which in turn we have shown is smaller than 1. Using this, we obtain

$$\begin{aligned}
&\frac{1-p}{p}\sum_{k=\frac{t-1}{2}}^{K}\left(F\left(\bar{N}_{2k+2}\right)-F\left(\bar{N}_{2k}\right)\right)h\left(\bar{N}_{2k+1}\right)+g\left(\bar{N}_{t-1}\right)-\frac{h\left(\bar{N}_{t-2}\right)}{p_t}\\
&\quad<\frac{1-p}{p}+g\left(\bar{N}_{t-1}\right)-\frac{h\left(\bar{N}_{t-2}\right)}{p_t}\\
&\quad<\frac{1-p}{p}h(1)+g\left(1\right)\\
&\quad<v+g\left(1\right)-h\left(1\right),
\end{aligned}$$

where the last inequality is true if $v > \frac{1}{p}h(1)$. Hence, if $v > \frac{1}{p}h(1)$, the incentive compatibility constraint is satisfied.

A3. Proof of Proposition II.4

As established before, the equilibrium that maximizes payoff of the viable types minimizes their total privacy leakage. Using the change of the variables introduced in (6), the formula for privacy leakage (5) becomes

$$\Psi_1(\boldsymbol{M}_T) + \Psi_2(\boldsymbol{M}_T) = \sum_{t=1}^{T-1}(\phi(\bar{M}_{t+1}) - \phi(\bar{M}_{t-1}))\bar{M}_t. \tag{A2}$$

If for all M, it holds that $M \geq \phi(M)$, then

$$\begin{aligned}\Psi_1(\boldsymbol{M}_T) + \Psi_2(\boldsymbol{M}_T) &\geq \sum_{t=1}^{T-1}(\phi(\bar{M}_{t+1}) - \phi(\bar{M}_{t-1}))F(h^{-1}(\bar{M}_t)) \\ &= \phi(\bar{M}_{T-1})\phi(\bar{M}_T) = \phi(h(1))\phi(h(1)) = 1,\end{aligned}$$

where the inequality is strict if $M > \phi(M)$ for all M and it is an equality if $M = \phi(M)$ for all M. Since $\Psi_1(\boldsymbol{N}_2) + \Psi_1(\boldsymbol{N}_2) = 1$, the proposition follows.

A4. Proof of Proposition II.5

To prove part (a), take M for which $M < \phi(M)$, and let $\bar{N}_1 = h^{-1}(M)$. Then clearly $h(\bar{N}_1) < F(\bar{N}_1)$. Consider an incremental equilibrium $\boldsymbol{N}_3 = \{\bar{N}_1, 1, 1\}$. We have

$$\Psi_1(\boldsymbol{N}_3) + \Psi_2(\boldsymbol{N}_3) = h(\bar{N}_1) + 1 - F(\bar{N}_1) < 1, \tag{A3}$$

which is less than in the simple equilibrium $\boldsymbol{N}_2$.

To prove part (b), take an equilibrium $\boldsymbol{N}_T$, and consider a sequence $\hat{\boldsymbol{N}}_{T+1}$ for which $\hat{N}_t = \bar{N}_t$ for all $t \in \{1, \ldots, T-2\}$ and $\hat{N}_{T-1} = x$, $\hat{N}_T = 1$, and $\hat{N}_{T+1} = 1$, where $x \in \left(\bar{N}_{T-3}, 1\right)$. Then

$$\begin{aligned}&\Psi_1(\boldsymbol{N}_T) + \Psi_2(\boldsymbol{N}_T) - (\Psi_1(\hat{\boldsymbol{N}}_{T+1}) + \Psi_2(\hat{\boldsymbol{N}}_{T+1})) = \\ &\quad \left(1 - F\left(\bar{N}_{T-2}\right)\right) + \left(1 - F\left(\bar{N}_{T-3}\right)\right) h\left(\bar{N}_{T-2}\right) \\ &\quad\quad - \left(1 - F\left(\bar{N}_{T-2}\right)\right) h(x) - \left(F(x) - F\left(\bar{N}_{T-3}\right)\right) h\left(\bar{N}_{T-2}\right) - (1 - F(x)).\end{aligned}$$

Thus, the privacy leakage in $\boldsymbol{N}_T$ is strictly higher than in $\hat{\boldsymbol{N}}_{T+1}$ if and only if

$$(1 - F(x))\left(1 - h\left(\bar{N}_{T-2}\right)\right) < \left(1 - F\left(\bar{N}_{T-2}\right)\right)(1 - h(x)),$$

which holds if and only if

$$\frac{1-F\left(x\right)}{1-h\left(x\right)}<\frac{1-F\left(\bar{N}_{T-2}\right)}{1-h\left(\bar{N}_{T-2}\right)}.$$

Using the change of variables $\bar{M}_{T-2}\overset{\text{def}}{=}h\left(\bar{N}_{T-2}\right)$ and $z\overset{\text{def}}{=}h\left(x\right)$, the above inequality can be rewritten as

$$\frac{1-\phi(z)}{1-z}<\frac{1-\phi(\bar{M}_{T-2})}{1-\bar{M}_{T-2}}.$$

But when $\phi(\cdot)$ is strictly concave, one can find $x\in\left(\max\{\bar{N}_{T-3},\bar{N}_{T-2}\},1\right)$ and the corresponding $z\in\left(\max\{\bar{M}_{T-3},\bar{M}_{T-2}\},1\right)$ such that the above is satisfied. By Proposition II.3, if $pv>h(1)$, then $\hat{\boldsymbol{N}}_{T+1}$ is also an equilibrium, which completes the proof.

A5. Proof of Proposition II.6

First, we will use the following definition.

DEFINITION A.1: $\boldsymbol{M}_T$ *is a* matching equilibrium *if for all even* t, $M_t=M_{t-1}$. *Let* $\boldsymbol{Z}\left(\boldsymbol{M}_T\right)=\{M_1,M_3,..M_{T-2},1\}$.

LEMMA A.2: *Fix any equilibrium* $\boldsymbol{M}_T$. *Then there exists a matching equilibrium* $\mathbf{M}_{T'}$ *that Pareto dominates* $\boldsymbol{M}_T$. *All elements of* $\mathbf{M}_{T'}$ *are drawn from* $\boldsymbol{M}_T$. *Moreover, for all elements of* $\mathbf{M}_{T'}$, $\frac{\phi(M_t)-\phi(M_{t-1})}{M_t-M_{t-1}}$ *is decreasing in* t.

PROOF: Pick $M_t=\arg\max_s\frac{\phi(M_s)}{M_s}$. Construct the following sequence $\boldsymbol{M'}=\{M_t,M_t,M_{t+1},M_{t+2},...M_{T-2},1,1\}$ (note that this does not have to be an equilibrium). Then

$$\begin{aligned}\Psi\left(\boldsymbol{M}_T\right) &= \sum_{k=1}^{t}\left(\phi\left(M_{k+1}\right)-\phi\left(M_{k-1}\right)\right)M_k+\sum_{k=t+1}^{T-1}\left(\phi\left(M_{k+1}\right)-\phi\left(M_{k-1}\right)\right)M_k,\\ \Psi\left(\boldsymbol{M'}\right) &= \phi\left(M_t\right)M_t+\left(\phi\left(M_{t+1}\right)-\phi\left(M_t\right)\right)M_t+\sum_{k=t+1}^{T-1}\left(\phi\left(M_{k+1}\right)-\phi\left(M_{k-1}\right)\right)M_k.\end{aligned}$$

Hence,

$$\begin{aligned}
\Psi(\boldsymbol{M}_T) - \Psi(\boldsymbol{M}') &= \sum_{k=1}^{t} (\phi(M_{k+1}) - \phi(M_{k-1}))\, M_k - \phi(M_{t+1})\, M_t \\
&= \frac{M_t}{\phi(M_t)} \sum_{k=1}^{t} (\phi(M_{k+1}) - \phi(M_{k-1}))\, M_k \frac{\phi(M_t)}{M_t} - \phi(M_{t+1})\, M_t \\
&\geq \frac{M_t}{\phi(M_t)} \sum_{k=1}^{t} (\phi(M_{k+1}) - \phi(M_{k-1}))\, \phi(M_k) - \phi(M_{t+1})\, M_t \\
&= \frac{M_t}{\phi(M_t)} \phi(M_{t+1})\, \phi(M_t) - \phi(M_{t+1})\, M_t \\
&= 0,
\end{aligned}$$

where the inequality follows from the fact that by the definition of M_t, we have $M_k \frac{\phi(M_t)}{M_t} \geq \phi(M_k)$. Now pick $M_s = \arg\max_{M_k \in \boldsymbol{M}': M_k > M_t} \frac{\phi(M_k) - \phi(M_t)}{M_k - M_t}$. Construct $\boldsymbol{M}'' = \{M_t, M_t, M_s, M_s, M_{s+1}, M_{s+2}, ...1\}$. Then we have

$$\begin{aligned}
\Psi(\boldsymbol{M}'') &= \phi(M_t)\, M_t + (\phi(M_s) - \phi(M_t))\, M_t + (\phi(M_s) - \phi(M_t))\, M_s \\
&\quad + (\phi(M_{s+1}) - \phi(M_s))\, M_s + \sum_{k=s+1}^{T-1} (\phi(M_{k+1}) - \phi(M_{k-1}))\, M_k,
\end{aligned}$$

and hence

$$\begin{aligned}
\Psi(\boldsymbol{M}') - \Psi(\boldsymbol{M}'') &= \phi(M_{t+1})\, M_t + \sum_{k=t+1}^{s} (\phi(M_{k+1}) - \phi(M_{k-1}))\, M_k \\
&\quad - (\phi(M_s)\, M_t - \phi(M_t)\, M_s + \phi(M_{s+1})\, M_s).
\end{aligned}$$

By the definition of M_s, we have that $M_k \geq \phi(M_k) \frac{M_s - M_t}{\phi(M_s) - \phi(M_t)} - \frac{\phi(M_t) M_s - \phi(M_s) M_t}{\phi(M_s) - \phi(M_t)}$. Hence,

$$\begin{aligned}
&\Psi(\boldsymbol{M}') - \Psi(\boldsymbol{M}'') \\
&\geq \phi(M_{t+1})\, M_t - (\phi(M_s) M_t - \phi(M_t) M_s + \phi(M_{s+1}) M_s) \\
&\quad + \sum_{k=t+1}^{s} (\phi(M_{k+1}) - \phi(M_{k-1})) \cdot \left(\phi(M_k) \frac{M_s - M_t}{\phi(M_s) - \phi(M_t)} - \frac{\phi(M_t)\, M_s - \phi(M_s)\, M_t}{\phi(M_s) - \phi(M_t)} \right) \\
&= 0.
\end{aligned}$$

Repeating these steps, we construct a sequence that is strictly increasing, and hence constitutes a matching equilibrium, and that Pareto dominates $\boldsymbol{M}_T$, and

for which $\frac{\phi(M_t)-\phi(M_{t-1})}{M_t-M_{t-1}}$ is decreasing in t.

LEMMA A.3: *Take a matching equilibrium $\boldsymbol{M}_T$, and the corresponding $\boldsymbol{Z}(\boldsymbol{M}_T)$. Suppose that there exists t and $m \in (Z_t, Z_{t+1})$ such that $\frac{\phi(m)-\phi(Z_t)}{m-Z_t} > \frac{\phi(Z_{t+1})-\phi(m)}{Z_{t+1}-m}$. Then a matching equilibrium $\boldsymbol{M}'_{T+2}$ with a corresponding sequence $\boldsymbol{Z}\left(\boldsymbol{M}'_{T+2}\right) = \{Z_1, .., Z_t, m, Z_{t+1}, ..., 1\}$ welfare dominates $\boldsymbol{M}_T$.*

PROOF: $\Psi(\boldsymbol{M}_T) - \Psi\left(\boldsymbol{M}'_{T+2}\right) = \phi(Z_{t+1}) Z_t - \phi(Z_t) Z_{t+1} - \phi(m) Z_t + \phi(Z_t) m - \phi(Z_{t+1}) m + \phi(m) Z_{t+1} > 0$, where the inequality comes from the assumption of the lemma.

LEMMA A.4: *Take a matching equilibrium $\boldsymbol{M}_T$, and the corresponding $\boldsymbol{Z}(\boldsymbol{M}_T)$. Suppose that there exists t such that $\frac{\phi(Z_{t+1})-\phi(Z_t)}{Z_{t+1}-Z_t} > \frac{\phi(Z_t)-\phi(Z_{t-1})}{Z_t-Z_{t-1}}$. Then a matching equilibrium $\boldsymbol{M}'_{T-2}$ with a corresponding sequence $\boldsymbol{Z}\left(\boldsymbol{M}'_{T-2}\right) = \{Z_1, .., Z_{t-1}, Z_{t+1}, ..., 1\}$ welfare dominates $\boldsymbol{M}_T$.*

PROOF: $\Psi(\boldsymbol{M}_T) - \Psi\left(\boldsymbol{M}'_{T-2}\right) = \phi(Z_t) Z_{t-1} - \phi(Z_{t-1}) Z_t + \phi(Z_{t+1}) Z_t - \phi(Z_t) Z_{t+1} - (\phi(Z_{t+1}) Z_t - \phi(Z_t) Z_{t+1}) > 0$,where the inequality comes from the assumption of the lemma.

We will now show that in the optimal equilibrium, $M_t \in C$ for all t. Suppose that in the optimal equilibrium there exist some $M_k \notin C$. If $\frac{\phi(M_t)-\phi(M_{t-1})}{M_t-M_{t-1}}$ is not decreasing in t, then by Lemma A.2, we can construct a matching equilibrium that delivers a higher welfare and in which $\frac{\phi(M_t)-\phi(M_{t-1})}{M_t-M_{t-1}}$ is decreasing. Let $\boldsymbol{Z}$ be the corresponding sequence of this matching equilibrium. If all elements of $\boldsymbol{Z}$ are in C, then we are done. If not, then since $\frac{\phi(Z_{t+1})-\phi(Z_t)}{Z_{t+1}-Z_t}$ is decreasing there must exist $Z_k \notin C$ and $m \in C$ such that $m \in (Z_k, Z_{k+1})$ and $\frac{\phi(m)-\phi(Z_k)}{m-Z_k} > \frac{\phi(Z_{k+1})-\phi(m)}{Z_{k+1}-m}$. Hence, by Lemma A.3, we construct a matching equilibrium $\boldsymbol{Z}' = \{..., Z_k, m, Z_{k+1}, ...\}$ that increases welfare. It is easy to see that in this new equilibrium $\frac{\phi(m)-\phi(Z_k)}{Zm} < \frac{\phi(Z_k)-\phi(Z_{k-1})}{Z_k-Z_{k-1}}$. Hence, by Lemma A.4 we can eliminate Z_k. Repeating this process, we eventually eliminate all $M_k \notin C$. The result for strict concavity follows directly from Lemma A.3.

A6. Proof of Proposition II.7

PROOF: Let $\left\{\bar{N}_t\right\}_{t=1}^{T}$ be the equilibrium of length T that minimizes the total information leakage. Then by definition, we have

$$\left\{\bar{N}_t\right\}_{t=1}^{T-2} = \arg \min_{\left\{\bar{N}_t\right\}_{t=1}^{T-2}} \sum_{t=1}^{T-1} \left(F\left(\bar{N}_{t+1}\right) - F\left(\bar{N}_{t-1}\right)\right) h\left(\bar{N}_t\right),$$

where $\bar{N}_0 = 0$ and $\bar{N}_T = \bar{N}_{T-1} = 1$. From Proposition II.5 we know that the solution is interior. Using the change of variables (6) in the above expression and

differentiating with respect to M_t, we obtain that for each integer t such that $1 \leq t \leq T-2$ it must be the case that

$$\begin{aligned}\frac{d}{dM_t}\left(\sum_{t=1}^{T}\left(\phi(M_{t+1})-\phi(M_{t-1})\right)M_t\right)&\\ =\phi(M_{t+1})-\phi(M_{t-1})+\frac{d\phi(M_t)}{dM_t}(M_{t-1}-M_{t+1})&\\ =0.&\end{aligned}$$

Hence,

$$\begin{aligned}\frac{d\phi(M_t)}{dM_t}&=\frac{\phi(M_{t+1})-\phi(M_{t-1})}{M_{t+1}-M_{t-1}}\\ =\frac{\int_{M_{t-1}}^{M_{t+1}}\frac{d\phi(x)}{dx}dx}{M_{t+1}-M_{t-1}}&\leq\frac{d\phi(\frac{M_{t+1}+M_{t-1}}{2}))}{dM_t},\end{aligned}\tag{A4}$$

where the last inequality follows if the derivative of $\phi(\cdot)$ is weakly concave (the inequality is strict if the derivative of $\phi(\cdot)$ is strictly concave). When we compare the far left-hand side with the far right-hand side of the above inequality and use the fact that $\phi(\cdot)$ is strictly increasing, we obtain that

$$M_t \leq \frac{M_{t+1}+M_{t-1}}{2}\tag{A5}$$

(with strict inequality if the derivative of $\phi(\cdot)$ is strictly concave). Thus, since this holds for every $t \in \{1,\ldots,T\}$ we get that $\{M_t - M_{t-1}\}_{t=1}^{T}$ is (weakly) increasing. Using the change of variables (6), equation (A5) implies the first part of the proposition. The proof for the convex case is analogous.

A7. Proof of Proposition II.8

Let $\boldsymbol{N}_T = \left\{\bar{N}_t\right\}_{t=1}^{T}$ be the equilibrium of length T that minimizes the total information leakage. Then by definition, we have

$$\boldsymbol{N}_T = \arg\min_{\left\{\bar{N}_t\right\}_{t=1}^{T-2}} \sum_{t=1}^{T-1} F\left(\bar{N}_{t+1}\right) - F\left(\bar{N}_{t-1}\right) h\left(\bar{N}_t\right),$$

where $\bar{N}_0 = 0$ and $\bar{N}_T = \bar{N}_{T-1} = 1$. We have

$$\begin{aligned}\Psi^i(\boldsymbol{N}_T) + \Psi^j(\boldsymbol{N}_T) &= \sum_{t=1}^{T-1} \left(F\left(\bar{N}_{t+1}\right) - F\left(\bar{N}_{t-1}\right)\right) h\left(\bar{N}_t\right) \\ &= \sum_{t=1}^{T-1} \left(L_{t+1} - L_{t-1}\right) h(F^{-1}(L_t)) \\ &= \sum_{t=1}^{T-1} \left(L_{t+1} - L_{t-1}\right) L_t^{\alpha}\end{aligned}$$

using the change of variables $L_t \overset{\text{def}}{=} F(N_t)$ and the assumption that $h(N) = F(N)^{\alpha}$. Now, similarly to the proof of Proposition II.7, after differentiating with respect to L_t the first-order conditions yield

$$L_{t+1} - L_{t-1} = \frac{h(F^{-1}(L_{t+1})) - h(F^{-1}(L_{t-1}))}{h(F^{-1}(L_t))'} = \frac{L_{t+1}^{\alpha} - L_{t-1}^{\alpha}}{\alpha L_t^{\alpha-1}}.$$

Thus,

$$\Psi^i(\boldsymbol{N}_T) + \Psi^j(\boldsymbol{N}_T) = \sum_{t=1}^{T-1} (L_{t+1} - L_{t-1}) L_t^{\alpha} \tag{A6}$$

$$= \frac{1}{\alpha} \sum_{t=1}^{T-2} \left(L_{t+1}^{\alpha} L_t - L_{t-1}^{\alpha} L_t\right) + (1 - L_{T-2}). \tag{A7}$$

Summing (A6) with α times (A7) and observing that $L_T = L_{T-1} = 1$ yields

$$(\alpha+1)(\Psi^i(\boldsymbol{N}_T) + \Psi^j(\boldsymbol{N}_T)) = L_{T-2}^{\alpha} L_{T-1} + L_{T-1}^{\alpha} L_{T-2} + (\alpha+1)(1 - L_{T-2}),$$

and so

$$\Psi^i(\boldsymbol{N}_T) + \Psi^j(\boldsymbol{N}_T) = \frac{2 + \left(L_{T-2}^{\alpha} - 1\right) + \alpha\left(1 - L_{T-2}\right)}{\alpha + 1}.$$

The claim then follows from the fact that $L_{T-2} \to 1$ as $T \to \infty$ (by the proof of Proposition II.5).

Proof from Section III

B1. Proof of the version of Proposition II.3 outlined in Section III.B

The proof of Proposition II.3 depends on the unviable types not having privacy concerns only in two steps: A and D. When the unviable types have privacy concerns, Step A still holds but needs a different proof. Step D needs to be

altered as the behavior of the unviable types is different than in Proposition II.3. Below we present the new versions of these steps.

Step A: When $p_t^i = 0$, both players believe that evidence exchange will not lead to cooperation. Hence, at this point a player has an incentive to reveal new evidence only if the opponent is expected to reveal new evidence in return. But by the same argument as in the proof of Proposition II.1, the player who is supposed to reveal new evidence last, has an incentive to deviate to revealing no new evidence. Hence, no evidence exchange can occur.

Step D: By Step C, if the equilibrium strategy of j does not adhere to $\left\{\bar{N}_t\right\}_{t=1}^{T}$ in round t, it results in the opponent i withholding her evidence in $t+1$. Hence, it is better for j to either reveal no new information in this round, as this decreases her privacy leakage, or to adhere to the sequence $\left\{\bar{N}_t\right\}_{t=1}^{T}$.

B2. Proof of III.2

Let $\boldsymbol{N}_4 = \left\{\bar{N}_1, \bar{N}_2, 1, 1\right\}$ be such that

$$h_U\left(\bar{N}_1\right) \geq p g_V\left(\bar{N}_2\right) \tag{B1}$$

$$h_U\left(\bar{N}_2\right) \geq g_V(1) - g_V\left(\bar{N}_1\right). \tag{B2}$$

By continuity of g_V and h_U, it is possible to find $\bar{N}_1$ close to 1 and $\bar{N}_2$ close to 0, for which (B1) and (B2) are satisfied. Below we will show that there exists an equilibrium with the above $\mathbf{N}_4$ in which the viable types of both players adhere to $\boldsymbol{N}_4$, and the unviable types of both players reveal no evidence. Hence, in this equilibrium the information leakage of player 1 is $h_V(\bar{N}_1)$ and of player 2 is 0. Since in the simple equilibrium the information leakage of player 1 is $h_V(1)$, and player 2 is 0, $\boldsymbol{N}_4$ strictly Pareto dominates the simple equilibrium.

Suppose first that both types of player 2 follow the strategies outlined in the previous paragraph. The unviable type of player 1 knows that by revealing $\bar{N}_1$, she suffers $h_U(\bar{N}_1)$, and gains $g_V\left(\bar{N}_2\right)$ only if she faces the viable opponent. Hence, her IC is

$$0 \geq p g_V\left(\bar{N}_2\right) - h_U\left(\bar{N}_1\right),$$

which is satisfied by (B1). Suppose now that both types of player 1 follow the strategies outlined in the previous paragraph. Then at $t = 2$, the unviable type of 2 knows the type of her opponent. If the opponent revealed $\bar{N}_1$, then player 2 benefits $g_V\left(\bar{N}_1\right)$ if she does not reveal anything. If she pretends to be viable and reveals $\bar{N}_2$, then she suffers a privacy loss, but she will receive all information from the opponent. Hence, her IC is

$$g_V\left(\bar{N}_1\right) \geq g_V(1) - h_U\left(\bar{N}_2\right),$$

which again is satisfied by (B2).

It remains to show that the viable types of both players have an incentive to adhere to $\mathbf{N}_4$, but it should be clear (and is straightforward to show) that whenever players have an incentive to reveal evidence in a simple equilibrium, then they do in $\mathbf{N}_4$.

B3. Proof of Proposition III.3

Suppose $\boldsymbol{N}_T$ is a screening equilibrium. This requires that every unviable type prefers to reveal nothing in the first round at which she speaks instead of planning to reveal evidence until some round t. At $t = 1$, the unviable type of player 1 knows that with probability $(1 - p)$ she faces an unviable type, in which case no evidence will be revealed in $t = 2$. With probability p she faces the viable type, in which case she can stay in the conversation until any t such that $K^i > \bar{N}_t$ and receive $\bar{N}_{t+1}$ from the opponent. Hence, for all odd t, the following IC constraint must be satisfied:

$$p\left(g_V\left(\bar{N}_{t+1}\right) - h_U\left(\bar{N}_t\right)\right) - (1-p)\, h_U\left(\bar{N}_1\right) \leq 0. \tag{B3}$$

In round 2, no player will reveal evidence if no evidence is disclosed in round 1. For the unviable player 2 not to disclose any evidence in round 2 after she observes $\bar{N}_1$ in round 1, it must be that she prefers to walk away with $\bar{N}_1$ instead of planning to follow $\mathbf{N}_T$ until some t. That is, for all even t, the following IC constraint must be satisfied:

$$g_V\left(\bar{N}_{t+1}\right) - h_U\left(\bar{N}_t\right) \leq g_V\left(\bar{N}_1\right). \tag{B4}$$

Step 1: If $\boldsymbol{N}_T$ satisfies the IC constraints and at least one inequality is strict, then there exists another $\boldsymbol{N}'_T$ that satisfies the IC constraints with inequality and $\bar{N}'_1 < \bar{N}_1$.

Take the first t at which the IC constraint is satisfied with strict inequality. If $t = 1$, then by continuity and strict monotonicity of h_U and g_V one can decrease $\bar{N}_1$ and keep (B3) satisfied. Suppose then $t = 2$. Then one can decrease $\bar{N}_2$ and keep (B4) satisfied. Since $\bar{N}_2$ only enters the first and second period IC's, we need to make sure that (B3) for $t = 1$ is satisfied. But decreasing $\bar{N}_2$, relaxes (B3) for $t = 1$, so one can decrease $\bar{N}_1$. Repeating the same argument for any t proves the step.

Step 2: Suppose $\boldsymbol{N}_T$ is a screening equilibrium. By Step 1, we can assume that $\boldsymbol{N}_T$ satisfies the IC constraints with strict equality. Consider decreasing $\bar{N}_1$ by ε_1. In order to keep the IC for $t = 1$ satisfied, one needs to decrease $\bar{N}_2$ by some ε_2, which by continuity of all functions is a continuous function of ε_1 Therefore, to keep the IC for $t = 2$ satisfied, one needs to decrease $\bar{N}_2$ by ε_2, which again by continuity is a continuous function of ε_1. Continuing this logic until $T - 1$, we obtain that we need to decrease $\bar{N}_{T-1}$ by some ε_{T-1}. So far, by constructions,

all IC constraints are satisfied until $T-1$. We need to make sure that the IC constraint at $T-1$ is satisfied as well. Suppose that $T-1$ is odd. Then the IC is

$$p\left(g_V(1) - h_U\left(\bar{N}_{T-1} - \varepsilon_{T-1}(\varepsilon_1)\right)\right) - (1-p)\, h_U\left(\bar{N}_1 - \varepsilon_1\right) \leq 0.$$

Since in $\boldsymbol{N}_T$, $\bar{N}_{T-1} = 1$, and by continuity of ε_{T-1} with respect to ε_1, one can find $\varepsilon_1 > 0$ that will make this IC satisfied. And by construction,, $\bar{N}_1 - \varepsilon_1 < \bar{N}_1$.

Step 3: It remains to show that the lowest N_1 is a decreasing function of b. Since by Step 1, the lowest N_1 is achieved in the equilibrium in which all constraints are binding, it remains to show that increasing b, relaxes all the constrains. But this is immediate from (B3) and (B4).

We will prove part (b) now. Define $h \overset{\text{def}}{=} g_U \equiv g_V \equiv h_U \equiv h_V$. Take T even (an analogous proof can be constructed for T odd) and a sequence $\boldsymbol{N}_T$ with $\bar{N}_T = \bar{N}_{T-1} = 1$ such that (B3) and (B4) are satisfied with equality for all t. Then by adding the left-hand sides and the right-hand sides of (B3) and (B4) for the corresponding ts one obtains that for any odd K,

$$h(1) - h\left(\bar{N}_{T-K}\right) = \frac{K+2-2p}{p} \cdot h\left(\bar{N}_1\right).$$

Hence, for $K = T-1$, we get

$$\begin{aligned} h(1) - h\left(\bar{N}_1\right) &= \frac{T+1-2p}{p} \cdot h\left(\bar{N}_1\right) \\ \Leftrightarrow \frac{p}{T+1-p} &= h\left(\bar{N}_1\right). \end{aligned}$$

By setting $\bar{N}_1 = h^{-1}\left(\frac{p}{T+1-p}\right)$, we obtain a sequence for which all IC constraints for the unviable types are satisfied, and the unviable types reveal no evidence. If the viable types adhere to $\boldsymbol{N}_T$, then the privacy leakage is $\Psi_1(\boldsymbol{N}_T) = h\left(\bar{N}_1\right)$ and $\Psi_2(\boldsymbol{N}_T) = 0$, and $\lim_{T\to\infty} h\left(\bar{N}_1\right) = \lim_{T\to\infty} \frac{p}{T+1-p} = 0$. Hence, it remains to show that the IC constraints of the viable types are satisfied.

At $t=1$, if the viable type of player 1 reveals $\bar{N}_1$, with probability $1-p$ she faces the unviable type, and no more evidence is exchanged, and with probability p she faces the viable type and cooperates on the project. Hence, her IC constraint at $t=1$ is

$$0 \leq pv - (1-p)h(\bar{N}_1). \tag{B5}$$

In all other odd rounds she knows whether her opponent is viable. If she faces the unviable type, she clearly has an incentive to reveal no evidence. If she faces the viable type, she must prefer to continue exchanging evidence and end up with $v + g_V(1) - h_V(1) = v$, instead of walking away at some t with the evidence that

the opponent revealed in $t-1$. Hence, her IC constraint in round t is

$$v \geq h(\bar{N}_{t-1}) - h(\bar{N}_{t-2}). \tag{B6}$$

In each round in which player 2 speaks, she knows the type of her opponent, hence, her IC constraint in all even rounds is identical to (B6). By (B3) and (B4), $h(N_{t-1}) - h(N_{t-2}) \leq \max\left\{\frac{1-p}{p} h\left(\bar{N}_1\right), h\left(\bar{N}_1\right)\right\}$, and since $\lim_{T\to\infty} h\left(\bar{N}_1\right) = 0$, there exists T_0, such that for all $T \geq T_0$, (B6) is satisfied. Similarly, (B5) is also satisfied.

*

REFERENCES

Admati, Anat R, and Motty Perry. 1991. "Joint projects without commitment." *The Review of Economic Studies*, 58(2): 259–276.

Augenblick, Ned, and Aaron Bodoh-Creed. 2014. "To Reveal or not to reveal: Privacy preferences and economic frictions."

Bardsley, Peter, Andrew P Clausen, and Vanessa Teague. 2008. "Cryptographic commitment and simultaneous exchange." *Available at SSRN 1153162*.

Blum, Manuel. 1983. "How to exchange (secret) keys." *ACM Transactions on Computer Systems (TOCS)*, 1(2): 175–193.

Chen, Ying, and Wojciech Olszewski. forthcoming. "Effective persuasion." *International Economic Review*.

Compte, Olivier, and Philippe Jehiel. 2003. "Voluntary contributions to a joint project with asymmetric agents." *Journal of Economic Theory*, 112(2): 334–342.

Compte, Olivier, and Philippe Jehiel. 2004. "Gradualism in bargaining and contribution games." *The Review of Economic Studies*, 71(4): 975–1000.

Crawford, Vincent P, and Joel Sobel. 1982. "Strategic information transmission." *Econometrica: Journal of the Econometric Society*, 1431–1451.

Damgård, Ivan Bjerre. 1995. "Practical and provably secure release of a secret and exchange of signatures." *Journal of Cryptology*, 8(4): 201–222.

Dziuda, Wioletta. 2011. "Strategic argumentation." *Journal of Economic Theory*, 146(4): 1362–1397.

Ganglmair, Bernhard, and Emanuele Tarantino. 2014. "Conversation with secrets." *The RAND Journal of Economics*, 45(2): 273–302.

Georgiadis, George. 2013. "Projects and team dynamics." Working paper.

Goldwasser, Shafi, Silvio Micali, and Charles Rackoff. 1989. "The knowledge complexity of interactive proof systems." *SIAM Journal on computing*, 18(1): 186–208.

Hörner, Johannes, and Andrzej Skrzypacz. 2011. "Selling information." *Cowles Foundation Discussion Paper No. 1743R.*

Jovanovic, Boyan. 1982. "Truthful disclosure of information." *The Bell Journal of Economics*, 36–44.

Kartik, Navin. 2009. "Strategic communication with lying costs." *The Review of Economic Studies*, 76(4): 1359–1395.

Kessing, Sebastian G. 2007. "Strategic complementarity in the dynamic private provision of a discrete public good." *Journal of Public Economic Theory*, 9(4): 699–710.

Li, Hao, Sherwin Rosen, and Wing Suen. 2001. "Conflicts and common interests in committees." *American Economic Review*, 1478–1497.

Lockwood, Ben, and Jonathan P Thomas. 2002. "Gradualism and irreversibility." *The Review of Economic Studies*, 69(2): 339–356.

Marx, Leslie M, and Steven A Matthews. 2000. "Dynamic voluntary contribution to a public project." *The Review of Economic Studies*, 67(2): 327–358.

Milgrom, Paul. 1981. "Good news and bad news: Representation theorems and applications." *The Bell Journal of Economics*, 380–391.

Milgrom, Paul, and John Roberts. 1986. "Relying on the information of interested parties." *The RAND Journal of Economics*, 18–32.

Pitchford, Rohan, and Christopher M Snyder. 2004. "A solution to the hold-up problem involving gradual investment." *Journal of Economic Theory*, 114(1): 88–103.

Shin, Hyun Song. 1994. "The burden of proof in a game of persuasion." *Journal of Economic Theory*, 64(1): 253–264.

Stein, Jeremy C. 2008. "Conversations among competitors." *The American Economic Review*, 98(5): 2150–62.

Watson, Joel. 1999. "Starting small and renegotiation." *Journal of economic Theory*, 85(1): 52–90.

Watson, Joel. 2002. "Starting small and commitment." *Games and Economic Behavior*, 38(1): 176–199.

Yao, Andrew C. 1982. "Protocols for secure computations." 160–164, IEEE.

Yildirim, Huseyin. 2006. "Getting the ball rolling: Voluntary contributions to a large-scale public project." *Journal of Public Economic Theory*, 8(4): 503–528.