Essays in Information Economics

Aidan Smith

Thesis presented for the degree of

Doctor of Philosophy in Economics

St John's College

University of Oxford

8th August 2022

Acknowledgements

I am grateful to my supervisor Meg Meyer for invaluable support and guidance across my MPhil and DPhil, along with Peter Eso, Stephen Nei, Paula Onuchic, Ludvig Sinander, Alex Teytelboym, and many other faculty and students at the University of Oxford and beyond. I am also grateful to seminar participants at the University of Oxford, Cambridge INET, the Transatlantic Theory Workshop 2021, CTN Annual Workshop 2019 and SING15 and SING16 for helpful comments and suggestions. I thank the ESRC and the Department of Economics, University of Oxford for funding, and Paul Klemperer, without whom I would not have taken up a place on the MPhil or DPhil. Finally, I thank Anna Murgatroyd for everything else. All errors are my own.

Contents

Abstract

This thesis collects three papers which study settings of imperfect information transmission. In "Strategic Information Release on a Communication Network", we study the problem of a sender who seeds a network with information, knowing that as the information spreads, it will acquire noise due to the 'broken telephone' effect. In "Noisy Disclosure", we also study communication via a noisy channel, and analyse the disclosure problem of a sender who can send a verifiable message about the state of the world, which may be corrupted by noise before reaching a receiver. In "Reputational Incentives with Networked Customers", we study networked learning for customers in a setting in which customers can observe only the reviews written by their neighbours on a network, and a firm can engage in non-contractible effort when serving customers to increase the probability of good reviews. In the first two papers, we show that a sender looking to influence the behaviour of decision makers can attempt to overcome noise in transmission by seeding the most central receiver on a network, or by failing to disclose their private information with positive probability. In the final paper, we show that restricting the ability of customers to observe only their neighbours' reviews may be welfare improving, if doing so means a firm must exert more effort to build a good reputation. Our results have important implications for consumer welfare: for example; mandatory disclosure policies for firms may be harmful if communication is noisy, while making it easier for customers to access a firm's past reviews may mean a firm with a good reputation does not exert as much effort for customers.

Chapter 1

Introduction

How should privately informed parties act when trying to influence the behaviour of decision makers in the presence of informational frictions? For example, which politician should a lobbyist focus their message on when word-of-mouth communication is unreliable? How should a scientist communicate their research to a non-expert decision maker? When should a firm exert effort if customers do not know the firm's inherent quality? In this thesis, we study these questions, modelling interactions as games of incomplete information, and characterise equilibria and analyse social welfare.

The thesis is divided into three papers: "Strategic Information Release on a Communication Network", "Noisy Disclosure", and "Reputational Incentives with Networked Customers". Each of the three papers is presented in an entirely self-contained format (including references and bibliography), and can be read independently (and in any order). In the remainder of this introduction we provide a summary of each paper, along with an extended motivation for the frictions we study. Following the papers, we conclude by discussing links between them and providing a longer discussion of possible extensions than in the papers themselves.

Strategic Information Release on a Communication Network: As information spreads through social networks, its quality often deteriorates, because mis-

understandings occur in transmission. Networked agents adjust the importance they attach to gossip in response; it is common to treat fourth-hand information as less reliable than first-hand information, because fourth-hand information may feature up to three additional accumulated misunderstandings. We propose a new model of mechanical noisy diffusion to capture this 'broken telephone' effect, in which a sender looking to influence behaviour by networked agents chooses an agent to seed with information, and information then spreads through the network, acquiring noise as it spreads.

The model is as follows: the sender chooses a receiver on the network to be seeded with a normally distributed signal, and the receiver passes a message on to each of their neighbours, comprising the signal plus an additional normal noise term. Their neighbours in turn pass the information on to *their* remaining uninformed neighbours, and this process repeats, until each receiver connected to the seeding point has received a message. Receivers thus each observe a message whose noise is proportional to the length of the shortest path between that receiver and the sender's seeding point. Each receiver wants to match their action to a state-contingent bliss point, while the sender wants *all* receivers to take their state-contingent ideal action.

When receiver preferences are homogeneous, we show that when seeding is publicly observed, the sender's optimal seeding decision takes a simple form: if their preferences are sufficiently aligned with the receivers', seed the most central receiver; else, seed the least central receiver. We characterise the relevant centrality measure, which is closely related to both harmonic and closeness centrality in a sense we make precise. Seeding the most central receiver minimises the sum of posterior variances, giving the highest quality aggregate information possible in the presence of noise. In contrast, we show with heterogeneous receivers, the sender's preference for seeding the most aligned receiver may dominate their centrality concerns, and we give an example in which increasing aggregate preference *misalignment* can *improve* the quality of aggregate information provision via the sender's equilibrium seeding

decision.

We contrast our findings with a version of our model in which receivers do not observe the path their messages travelled, and show that with unobserved seeding, regardless of preference alignment the sender prefers to seed the most central receiver on a homogeneous receiver network. With observed seeding, the receiver benefits from providing misaligned receivers with low quality information because their updating rule responds optimally to place less weight on noisier messages, reducing the variance of their actions. With unobserved seeding, however, receivers have fixed conjectures about the amount of noise in messages, and for fixed conjectures the sender has an incentive to surprise them with *less* noisy messages, in order to further reduce the volatility of their actions. As a result, in equilibrium, receivers who do not observe seeding must expect the sender to seed the most central receiver (on a centrality measure weighted by conjectures), regardless of preference misalignment.

We consider two extensions to our base model: seeding by two senders simultaneously, and seeding agents who want to co-ordinate their actions with one another. When two senders simultaneously seed the network and receivers observe a message originating from each sender, we show that if one has preferences aligned with receivers and the other does not, the misaligned sender has an incentive to seed the same receiver as the aligned sender, because observing a second signal is less useful for a receiver who has already observed one informative signal. An implication is that if we observe a central receiver being seeded by multiple senders (for example, a well-connected politician sitting on many committees which reveal privileged information), we should *not* infer that all of the senders seeding that receiver have interests that are aligned with receivers; some may be trying to minimise the aggregate informativeness of the information they seed by giving it to a receiver who is already very well informed and so pays less attention to it.

When the sender seeds receivers who want to co-ordinate with one another (but the sender themselves does not care about co-ordinating receiver actions) we show

that co-ordination concerns can lead the sender to optimally seed receivers with *intermediate* centralities. The reason for this is that when receivers have co-ordination concerns, the sender's seeding decision both affects the *quality* of receiver information *and* how easy it is for receivers to co-ordinate their actions, because the sender's seeding decision affects correlation between receiver messages. As a result, if from a sender's perspective they would like receivers to react less to the messages they receiver, by seeding a less central receiver the sender both reduces the weight placed on messages due to additional noise on aggregate, *and* reduces the weight on messages by reducing correlation in receiver messages. If this second channel is sufficiently strong, it may be that moving the seeding point only *slightly* away from the most central receiver is sufficient to lead receivers to place the weight on their messages that is ex-ante optimal from the sender's perspective, meaning the sender then has no reason to want to reduce aggregate information quality further. As a result, we find that even if senders have no co-ordination concerns themselves, they may want to encourage co-ordination concerns among receivers, if from their perspective they view receivers as overreacting to information.

Noisy Disclosure: When a privately informed expert (or sender) offers advice to a decision maker (or receiver), their expertise can have two components; they may have access to private information the decision maker does not, and they may have a greater capacity to *understand* that information than the decision maker. If this second component of expertise is significant, it may impact the expert's ability to communicate clearly, because the decision maker may *miscomprehend* the messages they receive (for example, a computer scientist providing technical advice about quantum computing and online banking to a politician may risk large misunderstandings). Misunderstandings may occur even if the decision maker is restricted to only sending *true* messages to the receiver; for example, a regulator may fact-check advertising material or financial disclosures by firms, but these may

still be miscomprehended by customers or casual investors.

We model the expert's problem as a game of noisy verifiable disclosure, in which a sender privately observes a state of the world θ, and chooses whether or not to disclose this state to a receiver. If they do not disclose, the receiver observes nothing; if they do disclose, however, instead of observing θ directly, the receiver observes a noisy signal s. The receiver wants to match their action to the state; the sender wants the receiver to take an action which is b larger than the state. Taking the joint distribution of s and θ as fixed, we explore what disclosure rules can be supported for the sender in equilibrium.

We contrast our model to two natural benchmarks: noiseless verifiable disclosure and cheap talk. With noiseless disclosure, the famous unravelling result tells us that if a sender wants a larger action than the receiver in any state, the unique equilibrium outcome is full disclosure, with the receiver believing no disclosure implies the lowest possible state. We show that when disclosure is noisy, full disclosure is *only* an equilibrium outcome if the sender is *sufficiently* biased. If the sender's bias is small, they view noisy disclosure as risky, because a large noise realisation may lead the receiver to overestimate the state by a large amount, and the sender dislikes large overestimates. As a result, if the receiver believes no message implies the lowest possible state, a sender observing a very small state with small bias may strictly prefer to receive the non-disclosure action with certainty to receiving a draw from the action distribution induced by disclosure. If the receiver is sceptical on seeing no message, for sufficiently large bias we have full disclosure in equilibrium, but below some threshold for bias we have lower censorship, with the sender only disclosing sufficiently large states.

While lower censorship equilibria for a positively biased sender feature the same sceptical reasoning by the receiver as in the noiseless benchmark, with no news treated as bad news, we *also* show that for a sender with positive bias, we can support equilibria featuring *upper* censorship, with the sender only disclosing states

below some threshold. To see why these equilibria exist, it is useful to consider our other benchmark; cheap talk. If our noisy signal is arbitrarily noisy, such that θ and s are independent, we can note that even though s is uninformative about the state the receiver can *still* draw inference from the sender's *decision* to disclose. In other words, the sender can signal information about the state by only disclosing in some states. We can view this as a form of cheap talk; the sender has access to two cheap and coarse messages: disclosing an uninformative signal, and not disclosing. When disclosure decisions are purely cheap talk, we can reason that if a lower censorship equilibrium exists, a corresponding upper censorship equilibrium must *also* exist, since the disclosure decision has no intrinsic meaning with an uninformative signal.

We show that upper censorship equilibria for a positively biased sender may persist even if s is informative about θ. We show that if observing s does not narrow down the support of a receiver's posterior, then even as s becomes *arbitrarily* informative about θ and we approach noiseless disclosure, upper censorship equilibria can continue to exist. Such equilibria feature a kind of self-fulfilling prophecy by the receiver; they conjecture that the sender never discloses for states above a threshold, and so for all signal realisations the receiver takes an action below that threshold. As a result, although the sender has positive bias, high types cannot increase the expected action they receive by disclosing, because the receiver interprets their disclosures as noisy messages from a sender in a low state.

We study the receiver's ex-ante preferred disclosure rule for the sender, and show that in the presence of noisy transmission, the receiver is *not* best off when the sender always discloses. The reason is that the receiver can learn from the sender's disclosures in two ways: they can learn from the *contents* of the message, and from the sender's disclosure *decision*. Since message contents are noisy while the disclosure decision is perfectly observed, the receiver is best off when the sender utilises the (perfectly observed) message of no disclosure with positive probability. Again, the intuition is clearest considering the cheap talk benchmark with uninformative s; in

cheap talk the receiver is always better off when the sender uses two messages rather than garbling in equilibrium. Our results imply that mandatory disclosure policies may be harmful if receivers have imperfect comprehension.

We find that (taking the distribution of s as fixed) equilibrium disclosure rules with a biased sender are inefficient, and that in lower censorship equilibria a sender with positive bias discloses *too often* from the receiver's perspective. An implication is that a biased sender would benefit from being able to commit to a disclosure rule, and may be better off ex-ante when subject to a policy which restricts disclosures (for example, a policy of only publishing statistically significant results). In contrast, we show that if the sender is unbiased, their equilibrium disclosure rule is efficient and they would not benefit from commitment power, because their indifference condition exactly matches the first order condition for an optimal cutoff. Our results imply that, while a receiver is indifferent over sender bias with noiseless disclosure, they *strictly* prefer a less biased sender when disclosure is noisy.

Reputational Incentives with Networked Customers: When customers learn about a firm's quality from one-another's reviews, this can create *reputational* incentives for the firm to exert effort while serving them. Good reviews indicate a firm is more likely to be of high quality, increasing the probability that future customers visit. This can motivate firms to engage in costly, non-contractible effort when serving customers; a firm providing poor service to a customer risks not just *that* customer not returning, but all of their friends also staying away. A firm's incentives for effort will depend upon the information structure through which reviews are shared; if customers share reviews with friends on a social network, then a firm's incentives to exert effort for a customer depend *both* on how many others will see a review, *and* on how much those other customers already know about the firm.

We propose a model of reputational incentives in which a firm serves networked customers. In our model, customers are uncertain about the underlying skill level of

a firm, and are willing to pay more for high skill firms, and for firms exerting high effort. They visit the firm according to an exogenous order, and when they visit, choose whether or not to purchase at a fixed price. *After* they take their purchase decision, the firm chooses whether to exert effort for them, and the customer then reports whether or not they received good service to their neighbours on a network. We study how firm incentives for effort depend upon properties of the customer network, and explore which networks maximise social welfare.

Customers in our setting can learn about the firm both from the reported quality of service (reviews) and from the decisions of their neighbours. If a customer knows their friend only purchases in equilibrium having seen good reviews, the customer may purchase despite seeing only bad reviews themselves, because they infer the *existence* of good reviews from their friend's purchase decisions. For certain network structures we show this can lead to herding in equilibrium, with customers beyond a certain date copying the purchase decisions of well-informed earlier neighbours, regardless of the contents of the additional reviews they may have seen. Herding kills firm incentives for effort, since if customers start to base their purchase decisions on the *decisions* of their neighbours rather than on *reviews*, subsequent firm effort choices will not affect the probability that future customers purchase.

We show that (if customers are not herding) while firm incentives for effort are in general stronger for a customer when more other customers read their review, they are *also* stronger when the customers who read the review do not observe too many others themselves. The more well-informed a customer is (either by learning from reviews or from purchase decisions), the less likely it is that any one review will influence their decision, and so the less they contribute to firm incentives for effort. We show that a firm may exert high effort for customers significantly *more* often when customer networks are *less* well connected.

From the perspective of social welfare, this creates a *trade-off* for a social planner designing a network between incentives and learning. Other things equal, a plan-

ner would like *both* to incentivise effort from firms *and* for customers to take well informed decisions, but if customers observe too many reviews each, firms will have no incentives to exert effort.

We show that one way in which a planner may try to strengthen incentives without reducing information quality is if they can have customers visit simultaneously. When customers visit sequentially, the firm can condition tomorrow's effort choice on today's review; if they receive a sufficiently glowing review today, they know they are safe to shirk tomorrow. By having all customers for whom a firm may exert high effort visit in the same period, a planner can shut down this possibility of shirking following good reviews, and may be able to incentivise more effort from a firm in expectation.

Optimal networks (and customer orders) in our model have a simple structure; the planner divides customers into 'always buys' and 'sometimes buys', and constructs a bipartite network linking these two sets. They then have the firm receive visits by up to two waves of 'always buys', followed by a final wave of 'sometimes buys' who base their decisions on the reviews of the 'always buys'. We show that a complete bipartite network is socially optimal when there are not too many customers, but that as customer numbers grow, a planner may do better allocating customers to several disjoint subgraphs, because with too many customers a complete bipartite network will not be able to incentivise effort. An important policy implication is that if firm effort is sufficiently productive, tools which increase review visibility (such as auto-translate features) may *harm* social welfare, because by making it easier for customers to learn about firms they can remove firm incentives for effort, *reducing* average customer payoffs.

1.1 Motivation

We now offer a brief discussion containing an overview of our general modelling approach, some additional general motivation for studying the informational frictions we choose to model, and responses to some common challenges posed to models in these settings. We emphasise that this section should be viewed as entirely supplementary to our papers, both in the sense that the papers are self-contained, and in the sense that the reader can derive insights from our papers while remaining agnostic with regards to our general modelling philosophy and motivations. As such, the following discussion is more context for the questions we choose to ask than it is necessary to understand the questions themselves.

Modelling Approach In this thesis, our underlying concept of economic decision making is that consumers are, at least approximately, rationally inattentive. In other words, we think of consumers as making an assessment of which decisions in their life will have a large impact on their wellbeing, and then taking proportionate steps to improve the quality of their decision making accordingly. For example, a consumer may assess that their individual votes are very unlikely to significantly impact political decision making, and so take minimal steps to reduce misunderstandings when they read about politics; in contrast, even consumers with little mathematical interest may invest effort in reducing misunderstandings when filling out tax returns, if they expect doing so to significantly affect their wellbeing.

This underlying concept gives us an answer to a common criticism levied at models featuring informational frictions, which is: if the decisions being modelled were important, why would customers not take steps to avoid the friction? For example, if the friction is that customers can only learn from observing the decisions of their friends, why would a customer taking an important decision not ask a friend directly what their posterior belief is, rather than just updating based on observed actions?

Our answer is that customers *are* taking all *proportionate* steps to avoid frictions, but encounter uncertainty in almost every economic decision they take, and have a limited capacity to reduce it. For example, while a customer could engage in an intense interrogation of their friends to better understand the beliefs that informed an observed decision, they may only be able to spend so long doing so before they have few friends left. As a result, frictions persist in day-to-day life because it is not proportionate for customers to remove them, and are more significant either when they are very costly to remove (for example, it may take a lot of effort to remove a language barrier) or because the customer does not view the stakes of a decisions as particularly high. We note that there are many settings in which customer decisions are far more important to privately informed parties seeking to influence them than they may be to the customers themselves. For example, customers may care much less about their votes than political parties do, or may care much less about their assessment of a particular product's quality than the manufacturer does, if they have a close substitute.

We treat frictions as not just unavoidable (in the sense that it is not proportionate for customers to invest further in avoidance) but commonly understood. In many of the settings we are interested in modelling, it is reasonable to assume that customers take informational frictions into account when forming opinions and taking decisions. For example, customers typically understand that fourth-hand gossip is less reliable than first-hand, and that they are more likely to miscomprehend information about a product written in a foreign language. They are also often aware of the incentives that exist for firms serving them; customers put less weight on the dining experiences of Instagram influencers or newspaper reviewers because they understand that firms will have been trying harder to impress them. In the conclusion we provide some discussion how our findings might differ if customers did not understand the frictions they encountered (for example, if they engaged in credulous rather than Bayesian updating when observing noisy signals).

Noise Noise in observations is a common feature of many real world settings; in life we rarely observe the exact values of underlying economic parameters. While modelling noise may make many models more *descriptively* accurate, it is not necessarily parsimonious or illuminating to do so in all models featuring uncertainty. In general, we can identify two main reasons to *explicitly* model noise in observations: firstly, because we may want to study the impact of the noise *itself* (for example, to explore how optimal contracts vary with noise in performance measures); and secondly, as a robustness exercise to understand the generality of results in noiseless models: we can ask, do settings with small noise behave approximately the same as settings with no noise[1]?

"Strategic Information Release on a Communication Network" can be thought of as belonging to the first category; we take a feature of real world information diffusion processes (the 'broken telephone' effect) that has received little attention, and explore how it affects optimal seeding. Seeding problems are interesting precisely *because* information transmission is unreliable; if seeded information were transmitted noiselessly and with certainty throughout a network then on a connected network the sender would have no seeding problem to solve.

We show that if real world transmission processes are noisy, modelling this noise is important in the sense that the sender's seeding problem has significant differences to a seeding problem featuring other forms of unreliable transmission. The existing seeding literature (e.g. Banerjee et al. (2019)) focuses on information being transmitted with probability less than one, while we focus on the accumulation of misunderstandings as messages travel. Both modelling approaches highlight that optimal seeding will be characterised by centrality, but we identify important fea-

[1]The literature on leader-follower games with noisy observations of the leader's action (e.g. Bagwell (1995) is one example where a setting featuring even a tiny amount of noise in observations is qualitatively very different from a setting featuring no noise. Suppose a leader can invest in a costly deterrent for the follower, and would do so if perfectly observed. If the follower imperfectly observes the leader's investment decision, then if they expect the leader to always invest they will assume that if they observe non-investment this is the result of noise rather than a deviation, and therefore act as though the leader invested. Given this, if expected to invest the leader would do strictly better not investing.

tures of seeding problems that are specific to noisy transmission. For example, in a classic diffusion setting with noiseless transmission, a receiver does not care how far a message has travelled, only whether they receive it or not, and so the model with observed and unobserved seeding is identical. In contrast, with noisy transmission, a receiver places less weight on a message that has travelled further, and so with unobserved seeding needs to form a conjecture of the distance a message has travelled, while with observed seeding can adjust their updating rule to ensure they weight a message correctly. In other words, receivers face a completely non-strategic problem in classic diffusion settings, but with noisy transmission their optimal strategy depends upon their (conjecture of) sender strategy; as such, studying noisy transmission can generate significantly different insights to studying random diffusion.

"Noisy Disclosure", in contrast, can be thought of as exploring the robustness of classic verifiable disclosure models to the presence of noise. While we find equilibria featuring familiar sceptical reasoning (with no news interpreted as bad news by a receiver) which converge to the noiseless disclosure rule as noise vanishes, we *also* find equilibria which do not feature sceptical reasoning on seeing no news, and *do not* converge to noiseless disclosure as noise vanishes. Our findings highlight that there are equilibria that exist in disclosure games with a tiny amount of noise which are qualitatively very different to equilibria in completely noiseless disclosure games. For example, an analysis of noiseless (and costless) disclosure games would predict that a receiver can only interpret no news as good news if a sender wants to be *underestimated*; we find with a negligible amount of noise in messages that a sender can want to be *overestimated* and a receiver can still interpret no news as good news in equilibrium.

Networked Learning In settings in which many customers separately form beliefs and take purchase decisions, we model customers as engaging in *networked* learning. We can think of networked learning either as a literal description of a

learning process, or as a formalisation of a more abstract observation structure. In many settings customers do learn only from their neighbours on a social or professional network; for example, customers can observe when neighours or colleagues purchase a new car, but cannot tell when strangers do. More generally, even in settings without an identifiable social network, many observation structures can be *formalised* as networks; for example, we could formalise that a bilingual customer can read online reviews in both of their languages using a network in which customers are linked only to other online customers with whom they share a language.

A common challenge posed to models of networked learning is whether they remain relevant given technological advances have made it much easier for customers to learn from other customers beyond their friendship group (or school, or village, or church, etc.). As discussed earlier, we think of networked learning as being driven at least in part by rational inattention on behalf of customers who face a cost to observing others. The greater the amount of information that is available to a customer, the less realistic it is to assume they are capable of processing all of it in a cheap and efficient manner, and the more likely it is they will benefit from developing a heuristic or filtering mechanism to avoid being overwhelmed. As such, instead of thinking of online information as rendering networked learning obselete, we can think of networked learning as part of the customer's *solution* to the rational inattention problem created by the increasing amounts of information available online.

It is worth noting, however, that treating customers as rationally inattentive may not fully address concerns relating to the availability of online information. While customers may find the *marginal* value of maintaining an additional friendship to not be worth the cost, this does not imply that the *marginal* value of an additional piece of information is small when a customer is uncertain about a decision. In other words, if a customer is, for example, buying a new car, it may not have be proportionate for them to maintain a friendship with a keen motorist, but this does not mean that they do not want to engage in deep research or solicit an opinion

from a friend of a friend on a one-off basis. Phrased another way, if the marginal value to a customer of one additional observation is very high, it seems implausible that a customer with internet access does not find it worthwhile to source one from outside their immediate network.

This criticism has some force (though we note it is not unique to our models). With high-stakes customer decisions, it may be more appropriate to explicitly model the ability of customers to acquire additional information on top of what they observe costlessly. Again, however, we emphasise that there are many circumstances in which customers may view the stakes of a decision as much lower than a firm does; for example, suppose that customers hold a prior that Pepsi and Coke are very similar, and would choose Coke over Pepsi under their prior. If both drinks are priced the same, customers will have a very low willingness to pay to receive additional information about either drink, because even if they discovered that in fact they preferred Pepsi, if both drinks are very similar, the gains from switching would be minimal. In contrast, Pepsi would have very strong incentives to provide customers with information; if under their prior all customers purchase Coke, Pepsi can only gain from providing information, and have very strong incentives to do so as they make zero sales without providing information. More generally, we can reason that because it is easier for information to influence the decision of an agent who views two options as very similar under their prior, it may be that the circumstances in which customers have the *least* incentives to acquire costly information are *precisely* the circumstances in which firms have the *most* incentives to try and influence them via seeding, disclosures, or effort choices.

We now present the papers.

References

Bagwell, Kyle (1995). "Commitment and observability in games". In: *Games and Economic Behavior* 8.2, pp. 271–280. ISSN: 0899-8256. DOI: https://doi.org/10.1016/S0899-8256(05)80001-6. URL: http://www.sciencedirect.com/science/article/pii/S0899825605800016.

Banerjee, Abhijit et al. (2019). "Using Gossips to Spread Information: Theory and Evidence from Two Randomized Controlled Trials". In: *The Review of Economic Studies* 86.6, pp. 2453–2490. ISSN: 0034-6527. DOI: 10.1093/restud/rdz008. eprint: https://academic.oup.com/restud/article-pdf/86/6/2453/30302450/rdz008.pdf. URL: https://doi.org/10.1093/restud/rdz008.

Chapter 2

Strategic Information Release on a Communication Network

Strategic Information Release on a Communication Network

Aidan Smith

August 8, 2022

Abstract

Information quality can deteriorate as information spreads by word of mouth, due to accumulated misunderstandings. We analyse a game in which a sender seeds a communication network with information, which travels through the network to receivers, acquiring noise. If seeding is observed, when receiver preferences are homogeneous the sender wants to provide information to the most central receiver if the sender's preferences are sufficiently aligned with the receivers, else they want to provide information to the least central receiver. With heterogeneous receiver preferences or unobserved seeding, it may remain optimal for the sender to seed the most central receiver even when all receivers on the network are very misaligned with the sender. In a setting with two competing senders, we show that a sender whose preferences are misaligned with receivers may find it optimal to copy the seeding decision of the other sender, even if this implies seeding the most central receiver.

Keywords: Information, Information Transmission, Network Formation, Noise

JEL Classification: D82, D83, D85

1 Introduction

When we decide to share information, we are naturally concerned both with how our audience might react to it, *and* who they in turn might tell. Information spreads through social networks, and so when considering what we would like an agent on a social network to know, we also need to take into account what we would like their neighbours to know (and their neighbours' neighbours, and so on). However, we expect agents located further from the point of information release on a network to receive less reliable information, both because messages may not be guaranteed to reach them, *and* because messages that do reach them will have accumulated all the misunderstandings that occurred *en route* (the 'broken telephone' effect).

This 'broken telephone' effect has received relatively little attention in the literature[1] (while messages going missing with positive probability has been widely explored; see for example Banerjee et al. (2019)). The effect is the following; each time a message is transmitted, any previous misunderstandings are inherited *and* some new misunderstanding may occur. This leads to a *non-strategic* deterioration in information quality as a message travels; as Alice gossips with Bob, Bob may mishear or misunderstand her, meaning when Bob gossips with Charles, Charles inherits the misunderstanding between Bob and Alice, plus any new misunderstandings that occurred in Charles' conversation with Bob. When forming opinions, we generally take this effect into account; when we hear an account of something fourth-hand, we place less weight on it than a first-hand account because we understand it less likely to be accurate. Nonetheless, we are often unable to learn only via first-hand accounts; if Charles is not friends with Alice, Charles is only able to learn from her *via* Bob.

In this paper we model an information diffusion process affected by this 'broken telephone' effect, and study the problem of a sender looking to optimally seed a

[1] Jackson, Malladi, and McAdams (2019) study it in a non-seeding context focusing on its effects on learning for a decision maker.

network with information to influence behaviour. Receivers are located on a network, and a sender commits to a choice of receiver to seed with information. Each receiver on the network observes a noisy message about the state of the world and takes an action, where the amount of noise in the message is determined by the length of the shortest path from the receiver to the sender. Each receiver cares only about their own action, which they want to match with their state-contingent ideal action, while the sender cares about the actions of all receivers, and wants each receiver to take the sender's state-contingent ideal action. When choosing who to seed, a sender thus needs to take into account both how their seeding choice determines the amount of noise in receivers' messages, *and* how aligned receiver preferences are with the sender. We characterise the sender's optimal seeding choice with quadratic loss preferences when both the state and noisy messages are normally distributed.

With quadratic loss preferences, Kamenica and Gentzkow (2011) show that incentives for information provision take a 'bang-bang' form, where above a threshold for preference alignment between sender and receiver the sender would *ex-ante* prefer that receiver to learn the state of the world perfectly, while below the threshold for alignment the sender would prefer *ex-ante* the receiver learn no new information and inherit their prior. In our context, provided the sender's seeding decision is public and receiver preferences are homogeneous, these 'bang-bang' incentives imply that if sender's preferences are sufficiently aligned with receivers, it is optimal for the sender to release information to the most central receiver, while if they are not sufficiently aligned the sender wants to release information to the least central receiver.

For a normal state and normal noise terms we give a closed-form expression for the appropriate centrality measure, which we call 'Persuasion Centrality'. 'Persuasion Centrality' is closely linked to existing measures: if the sender perfectly communicates the state of the world to the seeded receiver, 'Persuasion Centrality' converges to harmonic centrality as the signal-noise ratio goes to zero (noise dom-

inates), while it becomes equivalent to closeness centrality as the signal-noise ratio explodes (signal dominates).

In contrast, if the sender's seeding decision is not public, we show regardless of preference alignment, the sender will prefer to seed the most central receiver (where the appropriate centrality measure is now a function of receiver conjectures). When seeding is unobserved, receiver updating rules do not adjust following deviations by the sender, so seeding a more central receiver does not affect the sensitivity of receiver actions to messages, but does reduce the volatility of their action due to noise, which benefits a risk-averse sender. Hence the sender wants to provide high quality information to all receivers regardless of preference misalignment.

When seeding is observed but receiver preferences are not homogeneous, we show that the relationship between preference alignment and information provision is more complicated. While a sender would seed the least 'Persuasion Central' receiver (minimising aggregate information quality) on a network on which the preferences of all receivers were equally misaligned with the sender, if we increase the misalignment of all but one of the receivers, this can cause the sender to switch to seeding the most 'Persuasion Central' receiver, and so increasing aggregate preference misalignment can *improve* aggregate information provision.

We apply our analysis to the case of two competing senders with heterogeneous preferences. If competing senders observe correlated signals, their optimal seeding problems are linked since, for receivers, a clear message from the second sender is relatively less useful if they also receive a clear message from the first sender. Hence a sender who is misaligned with receivers and wants to release information in a way that maximises posterior variances may face a trade off between taking the same seeding decision as the other sender and seeding the least central receiver. We give examples in which the incentives for a misaligned sender to copy the other sender's seeding decision is sufficiently strong that a misaligned sender would still seed the most central receiver (given the other sender is also doing so).

We discuss an extension of our model in which receivers have co-ordination concerns. Since in our model information transmission is non-strategic, we focus on the case of homogeneous receivers. If receivers have co-ordination concerns then receivers place less weight on messages and more on their prior (since a common prior acts as a public signal which can function as a co-ordination device). We argue that this can soften our 'bang-bang' result, as senders whose preferences are not perfectly aligned with receivers can now influence the sensitivity of actions to the state both by choosing the amount of noise in messages, *and* by making it easier or harder for receivers to coordinate with one another. In some settings a sender may want to seed a receiver of intermediate centrality, because they gain ex-ante from providing receivers with information only if receiver messages are not *too* correlated.

The paper is structured as follows: in the remainder of section 1, we discuss related literature and applications. In section 2 we formally introduce the model. In section 3 we analyse incentives for information provision, introduce our centrality measure, and characterise optimal seeding in settings of both homogeneous and heterogeneous receivers. Section 4 discusses how our results differ in a setting in which receivers do not observe the point of information release. Section 5 applies our analysis to the case of two competing senders. Section 6 discussed how our results would differ if receivers had co-ordination concerns. Section 7 discusses our results, modelling choices, and directions for further research. All proofs are in the appendix.

1.1 Literature

We model the strategic seeding decision of a sender in the presence of a noisy but non-strategic diffusion process. The existing literature on seeding networks with information has largely focused on diffusion processes in which with positive probability a receiver does not pass messages to their neighbours (but when messages are transmitted they are transmitted noiselessly). In these settings, the sender's

seeding problem is the result of the *probability* of successful transmission falling as information spreads further from the seeding point; in contrast, in our setting the sender's seeding problem results from the *informativeness* of a receiver's message falling as the message travels further.

In Banerjee et al. (2013) and Banerjee et al. (2019), the authors study the diffusion of the adoption of microfinance and vaccinations respectively within communities, and explore how the centrality of those seeded with information can both *explain* uptake and *assist* a strategic sender looking to maximise uptake. In a setting in which nodes fail to pass information to neighbours with positive probability, they introduce the notion of 'diffusion centrality', where the diffusion centrality of a node i gives the expected number of times all nodes would hear a piece of information after T rounds of transmission if i were seeded.

Our model differs significantly, both because we model noisy rather than uncertain transmission, and because our objective is to provide a *theoretical* model to study a strategic seeding problem, while Banerjee et al. (ibid.)'s main focus is on *estimating* a model and *testing* the effectiveness of seeding. We also find that centrality plays a key role in characterising optimal seeding when imperfect transmission is due to noise rather than random messaging, but the centrality measure we identify differs in important respects. We assume receivers only pass on gossip to friends who are uninformed, so our centrality measure depends only on the shortest path length; in contrast, Banerjee et al. (2013)'s diffusion process allows multiple messages to reach a receiver, so their centrality measure effectively counts expected number of completed walks shorter than T from a node.

Both centrality notions make different simplifying assumptions on the transmission process for tractability[2]. Diffusion centrality tracks the expected number of messages nodes receive, whereas from a purely informational perspective receivers

[2]Banerjee et al. (2013) derive diffusion centrality as a tractable approximation of an underlying measure which explicitly captures adoption probabilities for an innovation, which they call 'communication centrality'.

do not benefit from any (noiseless) messages beyond the first, while our 'Persuasion Centrality' assumes receivers observe only one message each, although from an informational perspective receivers *would* benefit from receiving multiple noisy messages.

Galeotti and Goyal (2009) also study a model of strategic diffusion, which nests the problem of a profit maximising firm choosing a fraction of customers to directly advertise to, knowing customers will then share information about the product via word-of-mouth to neighbours on an infinite network with known degree distribution. Their firm does not observe the network directly, but may be able to target segments of customers with different degree distributions. They find that if a firm can target customers by out-degree (how many others a customer observes) the firm does best targeting *low* degree customers, because these customers are least likely to learn about the product otherwise via word of mouth, while if the firm can target by in-degree (how many others observe a customer), they do best targeting *high* degree customers who share information more widely.

Our word-of-mouth communication differs in that we model noisy transmission, and allow messages to travel to all receivers connected to the seed; in contrast, Galeotti and Goyal (ibid.)'s transmission process is noiseless, and (as they study an infinite network) the distance messages can travel is capped. We assume that the network is finite and publicly known, but our results on optimal seeding are similar to their results on targeting an unobserved network via in-degree; we find that a sender aligned with receivers does best targeting central receivers, who often also have *high* degrees. We note that their formal framework might offer a fruitful structure within which to generalise a model of noisy diffusion to large, unobserved networks.

In general, settings featuring noisy information transmission due to the 'broken telephone' effect are relatively unexplored within economics. An earlier working paper version of Navarro (2012) features a similar information transmission process

to ours, in which decaying information about product quality spreads from buyers via word of mouth through social networks along shortest paths. However, the analysis is focused on a monopolist choosing both quality and price of a good in the presence of network effects (the final paper subsumes word-of-mouth effects into a broader class of network effects), whereas our model investigates *targeted* information release.

Jackson, Malladi, and McAdams (2019) explicitly model the 'broken telephone' effect as a mechanical noisy information transmission process in a setting with a binary state. Their focus is quite different to ours; they explore how information quality relates to the number of noisy channels a receiver has access to, and focus on the case in which the sender does not observe the path messages have travelled (corresponding to unobserved seeding in our model). They show that learning can be improved by interventions which limit 'depth' (preventing messages from travelling more than a certain distance) or 'breadth' (limiting how many friends an agent can forward messages to), as these can increase the fraction of informative messages an observer sees, even if they reduce an observer's total number of messages.

In contrast our analysis is focused on seeding, and for the majority of the paper we assume receivers understand how far their messages have travelled. As our receivers only observe a single message which travels shortest paths, in general our receivers would be harmed by limiting depth (deleting messages that have travelled too far) or breadth (only letting a certain number of messages pass through a node), because this would not affect the sender's ex-ante incentives for information provision but would weakly increase noise in messages for any given seeding decision. In general in our setting receivers are better off with *more* connected networks, even if they do not directly observe the distance messages have travelled.

Our modelling of noisy information transmission gives our paper connections to work in the information theory and signal coding literature, following Shannon (see for example MacKay (2002)). However, while our model is similar to models

of information transmission via Gaussian channels, our focus is different to typical work within information theory; instead of taking a source and a receiver as fixed and trying to determine the optimal coding or network structure to maximise mutual information, we take the network structure of receivers as fixed, and consider where on that network a sender looking to influence behaviour would prefer to release information.

While we focus on a non-strategic diffusion process with *mechanical* noise, a related literature studies strategic diffusion processes when some agents are (non-strategically) seeded with information, and can communicate with other agents either via cheap talk (Galeotti, Ghiglino, and Squintani (2013), Hagenbach and Koessler (2010)) or via verifiable messages (Squintani (2018), Gieczewski (2022)). Our base model abstracts away from the agent preferences over one another's actions that drive models of strategic transmission, but such preferences could easily be incorporated into a microfoundation of our transmission process.

Galeotti, Ghiglino, and Squintani (2013) studies a setting in which agents observe informative signals about the same component of the state of the world, and want all other agents to take their state dependent ideal action. Their model does not feature an underlying network; instead, the communication network represents which agents send truthful messages to which other agents in equilibrium. They find agents are more willing to share their information to audiences with which they are more aligned, and are more willing to provide information to an audience containing a *misaligned* agent if that misaligned agent receives many other truthful messages. Our analysis of optimal seeding with two senders in section 5 features a similar effect; the more a receiver learns from another sender, the more willing a misaligned sender is to provide them with information.

Gieczewski (2022) studies agents who *are* located on a fixed network, where following random seeding agents can verifiably pass on information to their neighbours across multiple rounds of communication. He shows that full transmission will occur

only if preference misalignment between neighbours is not too large and the set of agents seeded is sufficiently diverse; else transmission will fail because agents will feign ignorance when they observe what they view as bad states. If preferences are not sufficiently aligned, messages about moderate states are transmitted more easily throughout the network in equilibrium, as biased senders have more incentive to not pass on messages in extreme states. While his model features a setting with random seeding, we would expect similar results to hold with public seeding and disclosure featuring noisy transmission; information in certain states will be censored in transmission, because in some states senders prefer to induce the non-disclosure action from a receiver rather than induce a draw from a random action distribution through noisy disclosure.

Our sender's seeding decision is taken without any private information, and can be thought of as choosing posterior distributions for receivers from a very constrained set of possible information structures. As such, it has the flavour of a very constrained information design problem; Galperti and Perogo (2020) studies a related problem in which an information designer who designs signals to reveal to fixed seeding points on a network, following which these signals spread through a network along directed paths, and agents take an action. Like us, they allow information to spread beyond the initial seeding point, although they do not model transmission as noisy. They show a sender may be able to manipulate agents using coded messages; for example, sending half of Charlie's message via the signal Bob passes on and half via the signal Alice passes on allows them to provide Charlie with information without revealing it to Bob or Alice. They consider optimal seeding, and show that a sender is better off when seeds are more influential, which in their model corresponds to creating a less connected information system. This is because if all seeding points are connected by directed paths to all other nodes on the network, the sender is effectively forced to chooses a single public signal; in contrast, if not all nodes are connected via a path to seeding points the sender can achieve

more outcomes by revealing different amounts of information to different agents.

We have a different focus; in their model, if there is a path from each possible seed to each other node on the network, the sender's seeding choice is irrelevant, as the sender provides a public signal regardless; in contrast, due to noisy transmission, our sender faces a trade-off in seeding even with connected networks. One interesting link that arises between our papers occurs when we let receivers have co-ordination concerns in section 6. We show in the presence of co-ordination concerns with homogeneous receiver preferences, optimal seeding may not be characterised by a centrality measure, because the sender's seeding choice affects *both* information quality *and* the ability of receivers to co-ordinate. Galperti and Perogo (2020) allow for arbitrary strategic interactions between receivers, and also find a sender can benefit from manipulating correlation in receiver messages via both seeding and signal design, to make it harder or easier for receivers to co-ordinate. As such, both our papers highlight that senders may be able to benefit from seeding a network in a way that *reduces* correlation between agents' messages.

The problem our receivers face is trying to estimate a random state of the world after observing messages from neighbours on a network. There is a large literature on learning from neighbours on a network, both by Bayesian and non-Bayesian agents: see Golub and Sadler (2016) for a recent survey. One branch of the literature concerns DeGroot learning (following DeGroot (1974)), in which agents announce their beliefs to their neighbours over a number of rounds, and engage in non-Bayesian updating, such that each announcement averages the beliefs they observed from their neighbours in previous rounds. This literature studies which network properties lead beliefs to converge to correct beliefs over time, if agents start by observing private noisy signals. In contrast, we focus on learning via individual noisy messages travelling from a seeding point; as such, an important feature of our setting is that agents *cannot* directly observe their neighbours' beliefs.

Nei (2021) studies a related rational social learning process featuring noisy mes-

saging between networked agents. His aim is to understand the effect of social context on learning; in his model, each agent has an idiosyncratic bias in messages they send, which is not publicly known and captures differences in the ways individuals express themselves ('social context'). For example, a positive bias may reflect that an optimistic individual tends to overstate the underlying state. Agents observe private signals and engage in multiple rounds of communication, announcing their beliefs each round, updating via Bayes' rule both about the signals and biases of others. He shows that agents' ability to filter out noise in messages by comparing messages across rounds and inferring bias terms depends upon the network structure; with agents located on a directed cycle and enough rounds of communication, all agents can learn one another's bias and underlying signals.

In contrast, we study a noisy transmission process in a seeding context, and do not endow agents with private information. We note that our motivation of noisy information transmission as the result of unavoidable misunderstandings in conversations is compatible both with time invariant noise (optimism/pessimism) as in Nei (2021), or time *varying* noise, due to a fundamental flaw in the transmission process. Under both interpretations, the details of our transmission process matter. We model agents as only passing gossip on to neighbours who are not yet informed. Learning with either time invariant or time varying noise would be very different with multiple rounds of belief sharing between agents; with time invariant noise, agents could compare messages that travelled different paths and strip out noise, while with enough rounds of messaging with time varying noise, a receiver will learn the underlying state with arbitrary precision due to the law of large numbers.

We study optimal seeding, focusing on identifying the best single receiver for the sender to seed, as is common in the seeding literature. Stepping back, Akbarpour, Malladi, and Saberi (2020) poss a more general question in a standard random diffusion setting: how many random seeds are worth the same to a sender as a fixed number of optimally chosen seeds? Their question highlights an underlying trade-off

for firms/lobbyists: gathering network data and seeding a network are both costly, so firms face a trade-off between spending resources identifying the most central agents and spending those resources on additional random seeds instead. They find that on many commonly studied networks, seeding would need to be very costly compared to information gathering for a sender to prefer a fixed number of optimal seeds to a larger number of random seeds.

As they note, the intuition behind their results extends easily to other diffusion models. Our model differs in that we have noisy diffusion, but shares the feature that the returns to moving closer to a receiver are convex; moving a seed one step closer to a receiver is more valuable when that receiver is already close. As such, we would expect very similar results to hold in our setting. Akbarpour, Malladi, and Saberi (2020)'s insights are of great practical importance for firms/lobbyists engaging in seeding; however, we note that they do not render the study of *optimal* single seeding irrelevant. In particular, if we wanted to assess the trade-off they identify in a setting of noisy transmission, we would need to *understand* the sender's optimal seeding decision; more generally, a sender may sometimes face an exogenous constraint on seed numbers.

1.2 Applications

Our analysis has applications to lobbying. In a political setting, if a lobbying group has a goal of moving voters' policy positions towards their own ideological bliss point, then when providing information to a political decision maker they would like to take into account both that decision maker's preferences and who that decision maker might share the information with. We can think of our model as capturing the problem of a think-tank looking to cultivate an ongoing relationship with a member of parliament, who they would incentivise (potentially via campaign donations) to disseminate the think tank's research.

Our model can also be applied to information release within organisations. For

example, a firm choosing how to spread information to employees about their policies may face a trade-off between spreading information via lower level managers, who are closer to 'customer facing' workers, and releasing information to higher level managers who may spread the information through the the organisational hierarchy more efficiently at the cost of providing worse quality information to 'customer-facing' workers. Our model is able to easily capture asymmetry in quality of information transmission between two receivers, which may be particularly relevant in organisational settings.

2 Model

The game is played by a single sender S, and M receivers. Receivers are located on a graph $\mathcal{G}$ with $\mathcal{N} = \{R_j | j \in \{1, ..., M\}\}$ as its node set, and commonly known adjacency matrix $A = [a_{ij}]$.

Timing The timing is as follows:

1. The sender publicly chooses a receiver R_l to seed with information.

2. Each receiver R_j receives a noisy message m_j that is informative about a state θ (where the amount of noise in m_j depends upon the distance between R_l and R_j on $\mathcal{G}$).

3. Each receiver R_j simultaneously chooses an action $x_j \in \mathbb{R}$, and payoffs are realised.

Preferences The sender's preferences are given by;

$$u(\boldsymbol{x}, \theta) = -\sum_{j=1}^{M} (x_j - w_s\theta)^2 \tag{1}$$

Receiver preferences are given by:

$$v_j(x_j, \theta) = -(x_j - w_j\theta)^2 \tag{2}$$

In state θ receiver R_j wants to set $x_j = w_j\theta$, while the sender would like R_j to set[3] $x_j = w_s\theta$. We assume without loss of generality that $w_s > 0$.

The parameter w_j is the sensitivity of a receiver's ideal action to the state of the world, and can capture both disagreement on the scale *and* direction of the ideal action. For example, policy positions of politicians who are up for re-election may be more sensitive to information about short-term policy benefits than a sender would like, meaning w_s and w_j may differ in scale, while if θ contains information on a polarising topic the *signs* of w_j and w_s may also differ.

Information Transmission All players share a common prior that $\theta \sim N(\mu, \sigma^2)$, Each receiver R_j receives one message m_j from the sender, with $m_j \sim N(\mu, \sigma^2 + \eta^2 + d_j)$ and $Cov(m_j, \theta) = \sigma^2$, where d_j is the length of the shortest path between R_l and R_j on $\mathcal{G}$.

The underlying information transmission process we aim to model is as follows: the sender chooses a receiver R_l to seed with a noisy signal $s \sim N(\mu, \sigma^2 + \eta^2)$ with $Cov(s, \theta) = \sigma^2$. R_l then passes this signal on to all of their neighbours on $\mathcal{G}$. As the message passes along a link between R_l and R_k, it acquires an additional noise term $\epsilon_{lk} \sim N(0, a_{lk})$ due to misunderstandings in transmission. Each neighbour of R_k then passes the signal on to all of their neighbours *who have not already received a message*, and again the signal acquires additional noise as it is transmitted (within each wave of transmission, we assume conversations happen sequentially, so we do not have two receivers simultaneously trying to talk to a shared neighbour). This process continues until no receiver who has received a message has any neighbours

[3] As the sender commits to a seeding decision before observing θ, they can influence the *variances* but not the *means* of receiver actions. As such, our results are also unchanged if we include additive bias in ideal actions and set R_j and S's ideal actions to be $w_j\theta + b_j$ and $w_s\theta + b_s$ respectively.

left who have not.

Formally, we can write this as:

$$m_j = s + \sum_{\langle k,l \rangle \in \mathcal{E}(\mathcal{P}_j)} \epsilon_{kl}$$

where $\mathcal{P}_j$ denotes a minimal subgraph of $\mathcal{G}$ containing a shortest path[4] from R_l to R_j on $\mathcal{G}$, and $\mathcal{E}(\mathcal{P}_j)$ is its set of edges. We assume all noise terms are independent: $E(\epsilon_{kl}s) = 0$ and if $\langle kl \rangle \neq \langle ij \rangle$, $E(\epsilon_{kl}\epsilon_{ij}) = 0$ $\forall i, j, k, l \in \{1, ..., M\}$).

We assume that each receiver *only* passes a message on to their neighbours if their neighbours have not already received a message from another source. If $\mathcal{G}$ is a tree the content of this assumption is that two receivers do not repeatedly bounce a message back and forth between one; instead receivers understand that the first time they hear a message will be the clearest. If $\mathcal{G}$ is not a tree the content of this assumption is that each receiver only engages in conversations about topics that are new to them; if a friend asks "Have you heard about X?" and a receiver answers "Yes", the conversation ends[5]. We discuss this assumption in greater depth in later sections.

For exposition, we will typically assume that $\mathcal{G}$ is a simple graph, (with symmetric and binary adjacency matrix A); however, our model can accommodate more general graphs. If $\mathcal{G}$ is an unweighted graph, then a message passing between neighbours R_i and R_j acquires a standard normal error term, while if it is weighted, $\epsilon_{ij} \sim N(0, a_{ij})$, so the weight on the edge between R_i and R_j captures how clearly R_i and R_j communicate[6]. Similarly if $\mathcal{G}$ is a directed graph, then if we have $R_{ij} = 0$ but

[4]Note that because neither the sender nor receiver has co-ordinating concerns, it does not matter *which* shortest path a message travels if there are multiple.

[5]Given that a receiver would receive better quality information if they received messages from two (or more) distinct paths, it is important to emphasise that the information transmission process in our model is *non-strategic*; we are trying to capture a plausible gossip process, rather than an optimal information gathering process for receivers. However, if we did want to try to micro-found our gossip process as a communication game it is worth remembering that while our receivers have incentives to *listen* to messages that do not travel shortest paths, they do not have any incentive to *pass on* messages given they only care about their own action.

[6]Note that if edges are weighted the shortest path between R_l and R_k may *not* be the path a

$R_{ji} \neq 0$, this captures that R_i does not pass messages on to R_j, but R_j does pass messages to R_i.

Equilibrium Our solution concept is Perfect Bayesian Equilibrium (appropriately generalised). We focus on pure strategy equilibria. In our setting, a Perfect Bayesian Equilibrium is an updating rule and strategy for the receiver such that $x_j = w_j E(\theta|m_j, d_j)$ (computed via Bayes' rule) and a strategy for the sender comprising a choice of R_l to seed $l \in \{1, ..., M\}$ that maximises $E(u(\boldsymbol{x}, \theta))$ given receiver strategies.

Given the sender has no private information when they choose R_l, we will restrict the receiver to an updating rule $E(\theta|m_j, d_j)$ which depends only upon the content of a message m_j, the distance to the point of seeding d_j, and their prior. In other words, because the sender is uninformed when they choose who to seed, the receiver cannot believe the sender signals anything via off-path seeding choices.

2.1 Remarks

For the remainder of this section we discuss our modelling choices.

We assume the sender commits to a seeding choice without having observed θ to avoid issues of signalling. If the sender took their seeding decision after observing (a noisy signal of) the state, they could potentially bypass noise in transmission by conditioning their seeding decision on the state. In particular, if a sender took their seeding choice after learning the state, as R_l is chosen publicly the seeding decision gives access to M coarse messages, so as $M \to \infty$ a sender could achieve arbitrarily precise communication via their seeding decision.

We also assume the sender is forced to choose a receiver to seed: they cannot choose not to release information. For some parameterisations of our model, the sender would choose the option not to seed if available, but assuming they are

message would take as the result of (up to M rounds of) receivers talking to neighbours sequentially, so we should be careful on interpreting the transmission process with weighted edges.

forced to engage in seeding allows us to study how misaligned senders might make use of noisy transmission to bury information. In many real world settings senders are obliged to commit to an information release policy even when they would rather release nothing: for example, governments may be forced to release policy reports even when they would rather keep them private.

We make the assumption that the sender's seeding choice R_l and the network $\mathcal{G}$ are both publicly observed (we relax this in section 4 to explore how the sender's seeding problem differs if seeding is unobserved). In general assuming that receivers observe the whole communication network will be equivalent to assuming that receivers observe the length of the path their message has travelled, so we could write an equivalent model in which the only thing each receiver observes is the ordered pair $\langle m_j, d_j \rangle$; in other words, receivers do not observe the network but when they hear a piece of gossip are told whether it is first-hand, second-hand, or third-hand etc.

Our assumption that the *sender* observes the network $\mathcal{G}$ is more substantive. It will fit settings where the sender either has access to data on social interactions (for example, firms may have very accurate data on the structure of online social networks), or where a social structure is public knowledge (e.g. the network is an organisational hierarchy, or the social structure of a political party which is reported on widely in the news). If the sender were uncertain about the network structure (or knew the underlying network but believed meetings only occurred randomly between neighbours), we could instead model the sender as maximising *expected* payoff across *possible* networks, and would obtain analogous results rephrased in terms of averages over possible networks (for example, if a sender wanted to minimise the sum of posterior variances they should seed the receiver with the highest *average* centrality across all possible networks).

In our model the sender's information release problem is limited to deciding which single receiver should receive a signal. In particular, we do not allow the sender to

reveal the signal to more than one receiver directly, and instead focus on the tractable single seeding problem. In general seeding problems, the optimal way to seed the network at two locations will not be to seed the sender's first and second best choices[7] in a single seeding problem. Note that aside from tractability issues, introducing multiple seeds brings additional *modelling* difficulties in the presence of the 'broken telephone' effect that are not present in standard diffusion models (for example, if due to multiple seeds Alice and Bob have two conversations about the same topic, it is not clear how we should model the *relationship* between misunderstandings in those two conversations).

We assume the sender cannot strategically manipulate signals before or after observing them both to avoid issues of strategic manipulation of messages to bypass noise, and for tractability. Note that if the sender could choose *any* messaging rule (with or without commitment), then by multiplying their signal by an arbitrarily large constant they could achieve arbitrarily precise communication for any given noise process with finite variance[8]. If the sender chooses a disclosure policy for the signal (with or without commitment) we would expect from Noisy Disclosure that they would engage in partial disclosure, significantly complicating receiver posteriors.

We assume that each receiver receives a single message, because they only pass gossip on to friends who have yet to hear about a topic. In contrast, if a receiver observed a separate message for each distinct path between them and seeded receiver R_l, then if $\mathcal{G}$ has cycles a receiver linked to R_l via several longer paths might receive better quality information than a receiver who is linked to R_l by a single, shorter path.

[7] Our model of two senders in section 5 allows the receiver to observe a message from each sender and if the two senders were to act co-operatively, their problem would be closely linked to a single sender seeding the network at two points.

[8] If the sender committed to a disclosure rule but we restricted the sender to a maximum variance in the message sent, results from the signal processing literature (see, for example, MacKay (2002)) tell us that when all variables are Gaussian, the messaging rule which maximises correlation between messages and the state is simply to send a signal with variance equal to the maximum permitted variance, obtained by amplifying the original signal by a constant.

A model where each receiver passes gossip on to all friends, even if those friends have already heard about a topic, is significantly more complex. If this were the underlying gossip process, we can note that d_j may still be a relevant summary statistic for the information quality of R_j if $\mathcal{G}$ is either approximately tree-like or sufficiently regular (in the sense that if two nodes are connected by short paths, they are also (weakly) more likely to be connected by longer paths). We can also note that many commonly used models of network formation, such as preferential attachment, have the property of generating networks which are locally tree-like, so any cycles are likely to be very large, meaning the second (and third etc.) messages a receiver observes are likely to be very noisy, and so have minimal impact on optimal seeding.

3 Optimal Seeding

We show that we have a 'bang-bang' result for information provision, with the sender wanting to minimise posterior variances for sufficiently aligned preferences, and maximise posterior variances for sufficiently misaligned preferences. If receiver preferences are homogeneous or heterogeneous but unobserved, we define a centrality measure which fully characterises optimal seeding. When receiver preferences are heterogeneous and observed, we show that centrality does not always characterise optimal seeding, and the relationship between preference misalignment and optimal seeding may be complex; higher aggregate preference misalignment can improve equilibrium information quality, while more connected receiver networks can harm equilibrium information quality.

R_j maximises their expected utility by setting $x_j = w_j E(\theta|m_j)$. This gives S an

expected payoff as follows:

$$\begin{aligned} E(u(\boldsymbol{x},\theta)) &= -\sum_{j=1}^{M} E((x_j - w_s\theta)^2) \\ &= -\sum_{j=1}^{M}\left[w_j(w_j - 2w_s)E([E(\theta|m_j)]^2) + w_s^2 E(\theta^2)\right] \end{aligned} \tag{3}$$

since for square-integrable random variables X, Y and a σ-algebra $\mathcal{H}$ we have $E(E(X|\mathcal{H})Y) = E(E(X|\mathcal{H})E(Y|\mathcal{H}))$.

We can immediately see that incentives for information provision take a 'bang-bang' form:

Proposition 1 'bang-bang' information provision. *If $w_j \in (0, 2w_s)$ then S's expected payoff is increasing in $Var(E(\theta|m_j))$, while if $w_j \notin [0, 2w_s]$ S's expected payoff is decreasing in $Var(E(\theta|m_j))$.*

In other words, S strictly prefers R_j to receive full information when preferences are sufficiently aligned ($w_j \in (0, 2w_s)$), while when preferences are misaligned ($w_j \notin [0, 2w_s]$), S strictly prefers R_j to receive minimal information, and would rather they inherited their prior. This result appears in Kamenica and Gentzkow (2011); in their setting, the sender commits to a stochastic messaging strategy *contingent* on the state of the world, and a sufficiently aligned sender should commit to fully revealing the state, while a sufficiently misaligned sender should commit to full garbling[9].

Our first proposition tells us that if a sender believes that a receiver's action choice is too sensitive to the state of the world (for example, because their priorities are too short term), or if any information received pushes receiver actions in the opposite direction to the sender's preferred action, then the sender is best off when the receiver learns nothing about the state of the world. If a lobbying group believe

[9]Rehak and Senkov (2021) extends Kamenica and Gentzkow (2011)'s analysis to cover the case of a sender who experiences quadratic loss in the distance of receiver action from their ideal action, but has an ideal action which may vary non-linearly with θ. They show that whenever the sender's ideal action is non-linear in θ, the sender prefers to provide the receiver with an *intermediate* amount of information; in other words, the sender no longer prefers either full revelation or complete garbling.

politicians are too short-termist when forming policy positions, it may be they do better burying any information on short-term impacts of policies (or not generating the information in the first place).

Note, that if receivers respond to information in the *direction* the sender wants them to, they can never be too *insensitive* to information for a sender to want to provide it to them. In other words, if a sender has information about the long-term benefits of a policy, and receivers are short-termist but agree about what consitutes a benefit, then no matter how little receivers care about the long term, the sender would always rather they receive the information with as little noise as possible.

Turning to our specific setting, R_j's posterior will be:

$$\theta|m_j \sim N\left(\frac{\sigma^2 m_j + (\eta^2 + d_j)\mu}{\sigma^2 + \eta^2 + d_j}, \sigma^2 - \frac{\sigma^4}{\sigma^2 + \eta^2 + d_j}\right) \tag{4}$$

and from the sender's perspective we have:

$$E(\theta|m_j) \sim N(\mu, \frac{\sigma^4}{\sigma^2 + \eta^2 + d_j}) \tag{5}$$

Only the first term in equation (3) can depend on d_j. As our distributional assumptions imply $\frac{\partial Var(E(\theta|m_j))}{\partial d_j} < 0$ we can thus rephrase our first proposition in terms of distances:

Observation 1 'bang-bang' in distances. *If $w_j \in (0, 2w_s)$ then S's expected payoff is decreasing in d_j, while if $w_j \notin [0, 2w_s]$ S's expected payoff is increasing in d_j.*

This observation tells us that, in the context of seeding, other things equal a sender would like to seed a network further away from receivers they are misaligned with and closer to receivers they are aligned with.

Using equations (3) and (5) we can rewrite the sender's general seeding problem

as:

$$\max_{l} -\sum_{j=1}^{M} \frac{w_j(w_j - 2w_s)}{\sigma^2 + \eta^2 + d_j} \tag{6}$$

where $d_j = d(l, j)$ is the distance between R_l and R_j on $\mathcal{G}$ (and $d_l = 0$). We can see that whether the sender should seed R_k will depend both upon how central R_k is on the network *and* on how aligned the preferences of R_k are with the sender compared to other receivers. We first focus on homogeneous receiver preferences to isolate the effect of centrality on seeding, before relaxing this assumption and exploring how heterogeneous receiver preferences can change optimal seeding relative to homogeneous receiver preferences.

3.1 Homogeneous Receiver Preferences

We will assume that all receivers have the same preferences[10] (but continue to allow for preference misalignment between sender and receivers). If $w_j = w \ \forall j$ we can simplify the sender's problem further as follows:

$$\max_{\boldsymbol{l}} -w(w - 2w_s) \sum_{j=1}^{M} \frac{1}{\sigma^2 + \eta^2 + d_j} \tag{7}$$

We can see the sender either wants to maximise or minimise a function which is decreasing in their distance from each individual receiver. As such, we can reason they prefer to seed more central agents if $w \in (0, 2w_s)$, and to seed less central agents of $w \notin [0, 2w_s]$, according to some centrality measure. The relevant centrality measure, which we call 'Persuasion Centrality' (PC), is as follows:

$$PC(n_i) = \sum_{n_j \in \mathcal{N}, j \neq i} \frac{1}{\sigma^2 + \eta^2 + d(n_i, n_j)} \tag{PC}$$

[10]Recall that because the sender cannot influence posterior means in expectation, we could always allow full heterogeneity in *additive* bias in receiver optimal actions and our results will be unaffected. As such, when we refer to full homogeneity in receiver preferences we mean full homogeneity in components of receiver preferences relevant to the sender's problem.

where $\mathcal{N}$ is the node set of a graph $\mathcal{G}$, $n_i, n_j \in \mathcal{N}$, and $d(n_i, n_j)$ gives the geodesic distance between nodes n_i and n_j on $\mathcal{G}$. 'Persuasion Centraility' is a modified form of 'Harmonic Centrality' (we discuss the exact relationship more below).

The following proposition sums up this insight:

Proposition 2. *If $w_j = w \; \forall j$, then if $w \in (0, 2w_s)$, it is optimal for the sender to seed the receiver with the highest 'Persuasion Centrality', while if $w \notin [0, 2w_s]$, it is optimal for the sender to seed the lowest 'Persuasion Centrality'.*

We can interpret the PC value of the receiver linked to as a measure of the quality of information provision. If S seeds the receiver with the highest PC value, they minimise the sum of posterior variances, and so in a sense provide the highest aggregate quality of information possible[11], while if they seed the receiver with the lowest PC value they maximise the sum of posterior variances and so provide the lowest aggregate quality of information possible. The intuition for our proposition follows from the 'bang-bang' incentives for information provision; if all receivers have the same preferences, then if they are sufficiently aligned the sender wants to maximise aggregate information quality; else they want to minimise aggregate information quality.

While the explicit form of 'Persuasion Centrality' does depend on our distributional assumptions, the insight that a sufficiently aligned sender should seed the most central receiver (and a sufficiently misaligned sender should seed the least) holds more generally. For example, if the noise terms acquired by messages in transmission were uniform rather than normally distributed, the specific centrality measure which characterised the sender's seeding problem would differ, but it would remain the case that a sufficiently aligned sender wanted to seed the most central receiver, since doing so minimises aggregate estimation error by receivers. 'Persua-

[11] Note maximising aggregate information quality is *not* the same as maximising aggregate sender utility with heterogeneous preferences. A utilitarian sender who maximised aggregate sender utility should place more weight on reducing the posterior variances of receivers with higher $|w_j|$ values, so we should take care using PC value as a benchmark with heterogeneous receivers; seeding the PC central receiver is *egalitarian* but not necessarily *utilitarian* optimal.

sion Centrality' is the relevant centrality measure in the special case in which either both the state and noise are Gaussian, or the case in which the state is unbounded and receivers restrict themselves to linear estimators of θ.

Persuasion Centrality 'Persuasion Centrality' is closely linked to the existing concepts of harmonic (HC) and closeness centrality[12] (CC):

$$CC(n, \mathcal{G}) = \frac{1}{\sum_{n' \in \mathcal{N} \setminus n} d(n, n')} \tag{CC}$$

$$HC(n, \mathcal{G}) = \sum_{n' \in \mathcal{N} \setminus n} \frac{1}{d(n, n')} \tag{HC}$$

Relative to HC a node's PC value places less importance on the shortest paths from a node, but like HC (and unlike CC), it assigns decreasing *marginal* importance to path length as path lengths grow. We can immediately note that:

Observation 2. *As $\sigma^2 + \eta^2 \to 0$, we have $PC(n) \to HC(n)$ $\forall n \in \mathcal{N}$.*

This observation tells us that if the signal R_l observes contains no base noise (such that $\eta^2 = 0$), then as the signal-noise ratio of messages becomes arbitrarily small ($\sigma^2 \to 0$), such that the transmission noise in messages dominates, the optimal receiver to seed to maximise aggregate information quality will become the harmonic central receiver.

We can relate PC value to closeness centrality as follows:

Proposition 3. *For any connected network $\mathcal{G}$, above some threshold value of σ^2 ranking nodes by PC value produces the same ordering as ranking nodes by closeness centrality.*

In other words, as the signal-noise ratio explodes; 'Persuasion Centrality' will give the same ranking as closeness centrality. One way to interpret this result is that

[12]See Bloch, Jackson, and Tebaldi (2019) for a taxonomy of centrality measures, and sufficient conditions for different centrality measures to agree on trees. Within their taxonomy, all centrality measures discussed in this paper use the neighbourhood (path) nodal statistic with an extended weighting.

as $\sigma^2 \to \infty$, a unit change in d_j induces an arbitrarily small relative change in the distribution of m_j, so it is valid to treat the effect of a change in d_j as approximately linear. As such the optimal seeding choice for an aligned sender becomes the choice that would be optimal by ranking receivers by the (linear) aggregate distance to all other nodes; in other words, the PC ranking matches the CC ranking.

When discussing different centrality measures, it is worth emphasising that in many settings, all of our centrality measures produce very similar rankings of nodes. Rochat (2009) computes Spearman's rank correlations for the HC and CC rankings on Erdos-Renyi random graphs, scale-free graphs, and some real world social networks, and in general finds correlation coefficients very close to 1. HC and CC values are likely to differ most on networks which feature very dense clusters connected by very sparse regions, as harmonic centrality attaches a decreasing marginal importance to path length as path lengths grow. Rochat (ibid.) does find a significant negative correlation between the two centrality measures on a graph featuring two dense clusters linked by a line, where harmonic centrality assigns higher values to nodes in the clusters, while closeness centrality picks out nodes in the centre of the connecting line. We would generally expect the ranking from 'Persuasion Centrality' to be somewhere between the two: like HC it assigns significantly more weight to whether a node is 1 or 2 steps from another node than it does to whether a node is 100 or 101 steps from another node, but compared to HC the convexity in distance is less severe and so the ranking will be closer to CC, which weights distances linearly.

3.2 Heterogeneous Unobserved Receiver Preferences

Optimal seeding is also completely characterised by PC value when receiver preferences are heterogeneous but unobserved. Suppose for each receiver w_j is a random variable drawn[13] from some distribution F with mean μ_w and variance σ_w^2 for all

[13]Since sender preferences are additively separable in receiver actions, we do not need to assume that receiver preferences are drawn independently from one another, only that they are identically distributed.

$R_j \in \mathcal{N}$. Assume w_j is independent of θ and m_j, and is privately observed by R_j. Then:

Proposition 4. *If receiver preference parameters w_j are unobserved but identically distributed with mean μ_w and variance σ_w^2, then if $\sigma_w^2 < w_s^2$ and $\mu_w \in (w_s - \sqrt{w_s^2 - \sigma_w^2}, w_s + \sqrt{w_s^2 - \sigma_w^2})$ the sender prefers to seed the receiver with the highest PC value, else they seed the receiver with the lowest PC value.*

If the average sensitivity of receiver actions to the state of the world is sufficiently close to the sender's preferred sensitivity, *and* variance in receiver sensitivities is sufficiently small, the sender does best providing information to the most central receiver, else they prefer to seed the least central receiver. Note that Proposition 2 is just a special case of this proposition in which $\sigma_w^2 = 0$.

To sketch the intuition, recall that if the sender knows a R_j's preferences, they want to give that receiver full information if $w_j \in (0, 2w_s)$, and no information otherwise. We can interpret this result as saying that, if preference heterogeneity is unobserved, the sender wants to maximise aggregate information quality if the probability of drawing receiver preferences $w_j \in (0, 2w_s)$ is sufficiently high; else they want to minimise aggregate information quality. Hence if either the mean receiver type is too misaligned with the sender *or* if the variance of receiver preferences is too high, the sender believes the probability of drawing misaligned receivers is sufficiently high that the sender does best in expectation minimising aggregate information quality by seeding the least central receiver.

To give the unobserved heterogeneity model an economic interpretation, we can think of a sender who observes the structure of a certain political party or organisation, and knows members are likely to have similar preferences, but does not observing the exact preferences of individuals. If the sender has no information about how preferences correlate with location on a social network[14] then their infor-

[14]In some settings the sender might have some information on correlation between preferences and location, meaning they do not view preferences as identically distributed. In a political setting,

mation seeding decision will be based on the variance of preferences within the party (whether it is a 'broad church') and the alignment of the average party member with the sender.

One implication of our proposition is that regardless of mean alignment, increasing the variance in receiver preferences will always (weakly) push the sender towards providing lower quality information. In other words, if a sender is misaligned with the average position of a party and does not observe individual preferences, admitting members drawn from a wider range of ideologies can only harm the party (in terms of aggregate information quality). Hence if a party does feature diverse preferences among members, such that a sender would seed the least central receiver with unobserved preferences, the party may do better on aggregate by making preference heterogeneity observable. In other words, they might be able to increase aggregate information quality by making public who in the organisation disagrees with whom, so that senders can avoid/target individual members based on their preferences.

3.3 Heterogeneous Observed Receiver Preferences

If w_j values are heterogeneous but observed then we cannot fully characterise optimal seeding using a centrality measure. The relationship between preference alignment and the quality of information provided becomes more complicated; we prove a proposition which illustrates that decreasing preference alignment may *improve* aggregate information provision, while increasing network connectivity may *harm* aggregate information provision:

Proposition 5. *If we have* $M-1$ *receivers with* $w_j = w' + c_j$ *and* 1 *receiver* R^* *with* $w_j = w^*$*, then as* $|w'| \to \infty$*, it is optimal for the sender to seed the receiver*

for example, a sender may know that members of a political party who are on the extreme wings of the party ideologically are also likely to be peripheral in the party's social network.

with the highest PC value if:

$$\min_{\{R_j|w_j=w'+c_j\}} \left\{ \sum_{j=1,j\neq k}^{M-2} \frac{1}{\sigma^2+\eta^2+d(k,j)} \right\} + \frac{1}{\sigma^2+\eta^2}$$
$$> PC(R^*,\mathcal{G}) > \max_{\{R_j|w_j\propto w'\}} \left\{PC(R_j,\mathcal{G})\right\}$$

In other words, as the sender becomes infinitely misaligned with all but one receiver R^* (even if they are still severely misaligned with R^*), if R^* is 'Persuasion Central' the sender finds it optimal to seed R^* provided R^* centrality is not *too* much higher than other receivers. To sketch the proof, we can see that as preference misalignment with all receivers except R^* becomes arbitrarily large, seeding incentives are dominated by the distance to receivers who are not R^*. As such, the sender wants to seed the receiver who is *least central* among the arbitrarily misaligned receivers, even if this means seeding the *most* central receiver on the whole network.

The above proposition implies that (unlike with homogeneous preferences) equilibrium information provision may be better on aggregate with *higher* preference misalignment. Provided R^*'s centrality is not *too* much higher than the other receivers', if we start with a network of receivers equally (very) misaligned with the sender, then the sender would seed the receiver with the lowest PC value, *minimising* information quality. However, if we increase the misalignment of all but the most central of those receivers (at the same rate), then above some threshold the sender's optimal seeding decision may become to seed the receiver with the highest PC value, *maximising* aggregate information quality. In other words, communication may be more efficient when overall misalignment between senders and receivers is larger rather than smaller.

Similarly, we can see our proposition implies that (unlike with homogeneous preferences) equilibrium information provision may become *worse* if we increase network connectivity. Suppose we start with a network of receivers where R^*'s

centrality is not *too* much higher than other receivers', such that the upper bound in our condition just holds, and the sender seeds R^*. Now add a link between R^* and another receiver. If this does not change which arbitrarily misaligned receiver is *least* central among themselves then we will now violate that upper bound because R^*'s centrality has increased, while the upper bound has not. As a result, the sender will no longer find it optimal to seed R^*, because doing so is no longer the best way to provide minimal information to the arbitrarily misaligned receivers.

4 Unobserved Seeding

In our baseline model we assume that receivers observe the seeding decision of the sender (in other words, they always hold correct beliefs about the amount of noise in their messages). In some settings, however, receivers might receive a piece of information via word of mouth without knowing how far it has travelled. In this section we consider a modified game in which the sender's seeding decision is unobserved, and show that incentives for information provision no longer take a 'bang-bang' form. We find if receivers do not know the path information travelled to reach them, then the sender's optimal strategy with homogeneous receivers is to seed the receiver who is most central according to a new centrality measure *regardless* of preference misalignment, where the centrality measure is a function of equilibrium conjectures.

With observed seeding, increasing the noise in receiver messages was only useful to senders because receivers observed seeding, and responded optimally to noisier messages by placing less weight on them when choosing their action. When seeding is unobserved, if a sender provides a receiver with a noisier message than expected, the receiver is unaware of this and continues to best respond to the amount of noise they expected in the message. As the weight receivers place on messages is fixed from the sender's perspective, sending noisier messages than expected induces

larger mistakes from a receiver, and so *increases* rather than *decreases* the expected squared distance of actions from a sender's bliss point. As a result, if the sender has no way to influence the *updating* rules of receivers, regardless of preference alignment the sender does better trying to minimise these mistakes and prefers to provide receivers with less noisy messages regardless of preferences.

Note that settings in which seeding is unobserved are potentially very different to settings in which it is observed. With seeding unobserved, messages correspond to rumours of unknown origins which have taken unknown routes; we do not know where a chain of information transmission started, or who talked to whom prior to the information reaching us (although we form conjectures about both). Senders in games with unobserved seeding may well be much more sinister entities, and our assumption that a sender cannot strategically manipulate messages except via their seeding decision may be much less plausible.

4.1 Modified Game

The timing of the modified game is as follows:

1. The sender privately chooses a receiver R_l to seed with information.

2. Each receiver R_j receives a noisy message m_j that is informative about θ (where the amount of noise in m_j depends upon the distance between R_l and R_j d_{lj} on $\mathcal{G}$).

3. Receivers each simultaneously choose an action $x_j \in \mathbb{R}$, and payoffs are realised.

Distributions are unchanged. We now require that $\mathcal{G}$ be strongly connected.

The modified game differs from the original in that, as the seeding decision is unobserved, a receiver R_j does not take into account the actual value of d_j when updating their beliefs on observing a message. We assume that receivers do *not* observe who passes a message on to them (if they did this would potentially be

informative about who was seeded[15]); instead, we think of messages arriving anonymously in an inbox and the receiver holding a conjecture about how noisy they are (alternatively, it could be that the receiver keeps track of gossip they hear but by the time they take a decision forgets its source).

We will find it useful to define the following centrality measure, which we will call 'Alpha Closeness' centrality (ACC):

$$ACC(n_i, \boldsymbol{\alpha}, \mathcal{G}) = \frac{1}{\sum_{n_j \in \mathcal{N}, j \neq i} \alpha_j d(n_i, n_j)} \tag{ACC}$$

where $\mathcal{N}$ is the node set of a graph $\mathcal{G}$, $n_i, n_j \in \mathcal{N}$, $\boldsymbol{\alpha}$ is an $M \times 1$ vector, and $d(n_i, n_j)$ gives the geodesic distance between nodes n_i and n_j on $\mathcal{G}$. Note that, like closeness centrality, ACC is not well defined on graphs which are not strongly connected.

4.2 Equilibrium with unobserved seeding

We show that in pure strategy equilibria with homogeneous receivers, regardless of preference misalignment the sender wants to provide the best quality information given equilibrium conjectures.

Proposition 6. *If $w_j = w \; \forall j$ (with $w \neq 0$), in any pure strategy equilibrium S must seed the receiver with highest ACC value for some equilibrium belief vector $\boldsymbol{\alpha} > 0$.*

In other words, regardless of preference misalignment, with unobserved seeding the sender always prefers to seed the most central receiver given equilibrium conjectures. Thus even if receiver preferences are extremely misaligned with the sender's preferences, with unboserved seeding the sender still prefers to provide them with 'better' quality information.

[15]Note that our assumptions imply that even R_l does not know that they were chosen by the sender to seed, which might seem implausible. However, note that the message R_l receives still contains noise (with variance η^2), so it will not necessarily be apparent from the *contents* of the message that it came from the sender rather than another receiver. Modifying our model so that R_l did know they were seeded would not significantly impact our results

The vector $\boldsymbol{\alpha}$ captures receiver conjectures, and if $R_l = R^*$ in equilibrium the corresponding equilibrium vector $\boldsymbol{\alpha^*}$ has jth element α_j^* satisfying:

$$\alpha_j^* = \frac{1}{(\sigma^2 + \eta^2 + d(R^*, R_j))^2}$$

We can see the marginal benefit to a sender from reducing distance to a receiver is decreasing in how far away that receiver believes the sender is; in other words, if the receiver puts very little importance on a sender's message, then making that message less noisy than anticipated does not have much of an effect.

We can reason that seeding the 'Persuasion Central' receiver must always be an equilibrium outcome when seeding is unobserved with homogeneous receivers, as if receivers all expect the sender to seed the 'Persuasion Central' receiver, a sender wanting to minimise the sum of posterior variances can definitely do no better than by seeding the 'Persuasion Central' receiver. However, even when a network has only one 'Persuasion Central' receiver, the following example illustrates that there may be multiple equilibria when seeding is unobserved. These equilibria feature 'self fulfilling prophecies', with receivers conjecturing that S will seed a certain receiver and those conjectures then making seeding that receiver optimal.

Example: Self-Fulfilling Prophecies Set $\eta^2 = 1$, and let $\mathcal{G}$ be a line graph, with $M = 5$ and $w_j = 1\ \forall j$. Number receivers by their position on the line. The 'Persuasion central' receiver is thus R_3.

Suppose all receivers expect S to seed R_4, and define $\boldsymbol{\alpha}^4$ as the corresponding vector of receiver conjectures. We thus have:

$$ACC(R_4, \boldsymbol{\alpha}^4, \mathcal{G})^{-1} = \frac{1}{(\sigma^2+1)^2} + \frac{4}{(\sigma^2+2)^2} + \frac{3}{(\sigma^2+3)^2} + \frac{4}{(\sigma^2+4)^2}$$

and:

$$ACC(R_3, \boldsymbol{\alpha}^4, \mathcal{G})^{-1} = \frac{2}{(\sigma^2+1)^2} + \frac{2}{(\sigma^2+3)^2} + \frac{4}{(\sigma^2+2)^2} + \frac{3}{(\sigma^2+4)^2}$$

where $\boldsymbol{\alpha}^4$ has jth element α_j which satisfies:

$$\alpha_j = \frac{1}{(\sigma^2 + 1 + d(R_4, R_j)^2}$$

Hence we can see that $ACC(R_3, \boldsymbol{\alpha}^4, \mathcal{G}) < ACC(4, \boldsymbol{\alpha}^4, \mathcal{G})$ if:

$$\frac{1}{(\sigma^2+1)^2} - \frac{1}{(\sigma^2+3)^2} - \frac{1}{(\sigma^2+4)^2} > 0$$

This holds if σ^2 is sufficiently small ($\sigma^2 \lesssim 4.92834$) and as such, if σ^2 is small then if all receivers expect S to seed R_4, it is optimal for the sender to do so. Hence we have an equilibrium with unobserved seeding in which the receiver with the highest ACC value given equilibrium beliefs is not the receiver with the highest PC value.

One question of interest is whether there is a limit to which seeding decisions we can support with 'self-fulfilling prophecies'. In the above example, seeding R_3 given receivers expect a sender to seed R_4 is optimal for $\sigma^2 \gtrsim 4.92834$. In other words, self-fulfilling prophecies worked only when σ^2 was small. This is intuitive; if the variance of θ is small relative to the amount of noise in messages then incorrect beliefs about d_j will lead to large differences between x_j and $w_j E(\theta|m_j)$, so if the sender does not act as expected receivers will make larger mistakes. If σ^2 is very large relative to noise, however, a receiver who holds incorrect beliefs about d_j will make proportionally smaller mistakes, so the sender is more willing to act unexpectedly if doing so leads to less noisy messages.

The following proposition tells us that the ranking produced by ACC must be independent of conjectures (for any consistent beliefs) and agree with the ranking produced by closeness centrality for σ^2 sufficiently large:

Proposition 7. *As σ^2 becomes large, above some threshold the receiver with the highest $ACC(i, \boldsymbol{\alpha}, \mathcal{G})$ value will also have the highest closeness centrality value for $\boldsymbol{\alpha}$ consistent with any conjectured pure strategy.*

Corollary 1. *As σ^2 becomes large, above some threshold any receiver seeded in a pure strategy equilibrium when seeding is unobserved must have the highest PC value.*

Note that Proposition 3 follows immediately from the above two results.

These results imply that for σ^2 sufficiently high, there exists a unique (pure strategy) equilibrium in the unobserved seeding game whenever there is a unique closeness central receiver. Hence the effect of receiver conjectures in determining the optimal seed must become negligible for a signal-noise ratio sufficiently large.

If we compare the setting in which seeding is observed to the setting in which seeding is unobserved, we can argue that both an aligned and misaligned sender weakly prefer (expected) equilibrium outcomes when seeding is observed to any pure strategy equilibrium outcome when seeding is unobserved. We can see that when seeding is unobserved, if $w_j = w \in (0, 2w_s)\ \forall j$, the pure strategy equilibrium which is optimal for both the sender and (utilitarian) efficient for receivers is for the sender to seed the 'Persuasion Central' receiver. This follows straightforwardly from adding the requirement that in equilibrium we must have $\bar{d}_j = d_j$ to the sender's expected utility maximisation problem. This tells us that, with sufficiently aligned preferences, the sender weakly prefers the equilibrium outcome under observed seeding to any equilibrium outcome under unobserved seeding.

If $w \notin [0, 2w_s]$ then the sender would seed the receiver with lowest PC value if seeding were observed, and the sender-optimal pure strategy equilibrium with unobserved seeding will have a feature of 'self fulfilling prophecy' (if such an equilibrium exists). We know that the sender prefers, ex-ante, that receivers observe no new information, and so would prefer to be able to commit to providing low quality information. When seeding is unobserved, however, they are unable to commit to seeding the least central receiver, and will want to seed the receiver with

highest ACC given receiver conjectures. If there exist multiple equilibria in the unobserved seeding game, the sender will *ex-ante* prefer to play the equilibrium in which they provide the lowest quality aggregate information possible; in other words, the equilibrium in which receivers conjecture they seed the least 'Persuasion Central' receiver.

5 Competing Senders

We now return to a setting in which seeding decisions are observed, and consider the problem of two senders simultaneously seeding the network with correlated signals (where a single noisy message travels to each receiver from *each* sender). Our main question of interest is: when will two senders prefer to seed the same receiver?

We focus on the case of competing senders, whose preferences are misaligned with one another, and who both observe signals that are informative about θ. We show that reducing the amount of noise in a message reduces posterior variance less when a receiver already has access to an informative message. This creates an incentive for aligned senders to take distinct seeding decisions, and for misaligned senders to all seed the same receiver. We derive results for when a misaligned sender prefers to copy the seeding decision of an aligned sender.

5.1 Model with Multiple Senders

The modified game has two senders S_1 and S_2, who simultaneously take their seeding decision. Receivers will set $x_j = w_j E(\theta|\boldsymbol{m_j})$, so sender S_i's objective is given by:

$$\begin{aligned} E(u_i(\theta, \boldsymbol{x})) &= -\sum_{j=1}^{M} \Big[(x_j - w_{S_i}\theta)^2 \\ &= -\sum_{j=1}^{M} \Big[w_j(w_j - 2w_{S_i}) E((E(\theta|\boldsymbol{m_j}))^2) + w_{S_i}^2 E(\theta^2) \Big] \end{aligned}$$

The timing of the game is as in the base model, with seeding decisions taken

simultaneously. Each sender S_i sends each R_j a message m_{ij}, giving the following information structure:

$$\begin{pmatrix} \theta \\ m_{1j} \\ m_{2j} \end{pmatrix} \sim N \left[\begin{pmatrix} \mu \\ \mu \\ \mu \end{pmatrix}, \begin{pmatrix} \sigma^2 & \sigma^2 & \sigma^2 \\ \sigma^2 & \sigma^2 + \eta_1^2 + d_{1j} & \sigma^2 + \rho \\ \sigma^2 & \sigma^2 + \rho & \sigma^2 + \eta_2^2 + d_{2j} \end{pmatrix} \right]$$

Using the standard notation for multivariate normals, we can rewrite this as follows:

$$\begin{pmatrix} \theta \\ \boldsymbol{m_j} \end{pmatrix} \sim N \left[\begin{pmatrix} \mu \\ \boldsymbol{\mu} \end{pmatrix}, \begin{pmatrix} \Sigma_{\theta\theta} & \Sigma_{\theta m_j} \\ \Sigma_{m_j\theta} & \Sigma_{m_j m_j} \end{pmatrix} \right]$$

giving $E(\theta|\boldsymbol{m}_j) = \mu + \Sigma_{\theta m_j}\Sigma^{-1}_{m_j m_j}(\boldsymbol{m}_j - \boldsymbol{\mu})$, and $Var(E(\theta|\boldsymbol{m}_j)) = \Sigma_{\theta m_j}\Sigma^{-1}_{m_j m_j}\Sigma_{m_j\theta}$.

The underlying information transmission process we have in mind is the same as in the base model; the message m_{ij} from sender S_i travels from seeded receiver R_l^i to R_j via a shortest path, acquiring noise. We assume that receivers treat each sender's message as a separate piece of news, so each receiver (on a connected $\mathcal{G}$) will end up observing two messages; one from each sender. We allow the noise in the sender's base signals to be correlated, but for simplicity assume noise added via gossip is not correlated across m_{1j} and m_{2j}:

We can write S_i's problem as follows:

$$\max_{R_l^i} -\sum_{j=1}^{M} w_j(w_j - 2w_{S_i}) \frac{\sigma^4(\eta_1^2 + d_{1j} + \eta_2^2 + d_{2j} - 2\rho)}{(\sigma^2 + \eta_1^2 + d_{1j})(\sigma^2 + \eta_2^2 + d_{2j}) - (\sigma^2 + \rho)^2}$$

As before differentiation will show that we have a 'bang-bang' incentives in distances; other things equal, senders prefer to provide fully informative messages to sufficiently aligned receivers, else they prefer receivers inherit their priors.

We can also see that $\frac{\partial^2 E((E(\theta|\boldsymbol{m_j}))^2)}{\partial d_{1j}\partial d_{2j}} < 0$, and so (for an aligned sender) the benefit to a sender of moving closer to a receiver is lower if that receiver is closer

to another sender[16]. As a result, comparing to the base model, we can see a sender whose preferences are aligned with receivers will in general do worse when the other sender seeds the same receiver as them, and will prefer not to take the same seeding decision of the other sender except on networks where one receiver is significantly more central than the others. In general we can see settings featuring multiple senders may feature multiple equilibria, since the senders now face a co-ordination problem. In particular, other things equal aligned senders now prefer to seed distinct receivers, while misaligned senders now prefer to seed the same receiver.

5.2 Competing senders

We now investigate a setting in which there is one sender whose preferences are aligned with receivers, and one whose preferences are misaligned with receivers, and ask when the misaligned sender would like to copy the aligned sender's seeding decision. We will assume $w_j = w \ \forall j$ and $w \in (0, 2w_{S_1})$, $w \notin [0, 2w_{S_2}]$.

S_1's problem is thus:

$$\max_{R_l^1} \sum_{j=1}^{M} \frac{\sigma^4(\eta_1^2 + d_{1j} + \eta_2^2 + d_{2j} - 2\rho)}{(\sigma^2 + \eta_1^2 + d_{1j})(\sigma^2 + \eta_2^2 + d_{2j}) - (\sigma^2 + \rho)^2} \tag{8}$$

while S_2's problem is:

$$\min_{R_l^2} \sum_{j=1}^{M} \frac{\sigma^4(\eta_1^2 + d_{1j} + \eta_2^2 + d_{2j} - 2\rho)}{(\sigma^2 + \eta_1^2 + d_{1j})(\sigma^2 + \eta_2^2 + d_{2j}) - (\sigma^2 + \rho)^2} \tag{9}$$

S_2 wants to minimise aggregate information quality, and can attempt to do so via two channels; they could seed a less central receiver, so that messages contain more noise

[16]This set up also gives us an insight into settings in which each receiver observes a 'private' signal in addition to their seeding message; note in equilibrium S_1 takes S_2's strategy as fixed, so it does not matter to S_1 whether m_{2j} is an exogenous 'private' signal or a message seeded strategically by S_2. If there is heterogeneity in the variance of private signals, the cross-partial effect will mean other things equal an aligned sender prefers to seed receivers with noisier private signals; if private signals are identically distributed their effect will be make the the variance of conditional expectation 'less convex' in distances, meaning the sender will care more about longer path lengths when seeding.

on aggregate, or they could seed a receiver who is closer to S_1's seeding choice, since a second message is relatively less useful for a receiver who has already observed a clear message. An immediate implication is that if S_2 can *both* reduce the centrality of the receiver they seed *and* seed closer to S_1's seeding choice, they will have a strict incentive to do so; if S_1 seeds R_1^*, then S_2 prefers also seeding R_1^* to seeding any receiver R_j with $PC(R_1^*) \leq PC(R_j)$.

It follows that if receivers do not vary in centralities (for example, a complete receiver network), S_2 always prefers to seed the *same* receiver as S_1 when S_1 plays a pure strategy. Given that S_2's objective is the negative of S_1's objective, this implies that in turn, if all receivers are equally central, S_1 strictly prefers to seed a *distinct* receiver to S_2. As such the game of seeding networks with equally central receivers has a 'matching pennies' structure, where S_2 wants to choose the same action as S_1 and S_1 wants to choose a distinct action, meaning the only equilibria are in mixed strategies.

More generally, we can see we may struggle to obtain a pure strategy equilibrium in settings with competing senders due to the opposition of interests between S_1 and S_2.

Proposition 8. *If $\eta_1^2 = \eta_2^2$ there exists no pure strategy equilibrium in which S_2 and S_1 seed the same receiver when $w_{S_1} \in (0, 2w)$ and $w_{S_2} \notin [0, 2w]$.*

This follows because if S_1 and S_2 seed the network with equally informative signals, their problems are exactly symmetric; S_1 maximises an objective that is identical to the objective S_2 minimises[17].

When signals are not equally informative there may exist pure strategy equilibria in which both senders seed the same receiver; for example, if S_2 has a very uninformative signal (η_2^2 large), S_1 will want to seed the most central receiver regardless

[17]Since mixed strategy equilibria occur because the aligned sender is seeking to avoid seeding the same receiver as the misaligned sender, while the misaligned sender is trying to seed the same receiver as the aligned sender, as in any 'matching pennies' type game, it follows that if seeding decisions were taken sequentially, there would be a significant second mover advantage.

of what S_2 does, so we may have an equilibrium where both seed the most central receiver. In general when signals are not equally informative our only candidate pure strategy equilibria in which both seed the same receiver is when both seed the most central receiver (according to an appropriate centrality measure), since if S_1 and S_2 both seed a receiver who is not the most central, S_1 can switch to seeding a more central receiver, and benefits both by reducing aggregate noise *and* by moving their seeding point away from S_2's.

When we consider settings in which receivers are not all equally central, a misaligned sender will face a trade-off between seeding the least central receiver and seeding the same receiver as the other sender. If a receiver is very isolated, a misaligned sender may do better seeding them than seeding same receiver as the aligned sender.

Example: Star Network For a simple star network, we can show that if $\eta_1^2 = \eta_2^2 = \rho$ and $\mathcal{G}$ is a simple star then for $M \geq 3 + \frac{1}{\sigma^2+\rho}$ there exist pure strategy equilibria in the seeding game in which S_1 seeds the centre and S_2 does not. If $M < 3 + \frac{1}{\sigma^2+\rho}$ there exist no pure strategy equilibria in the seeding game.

Suppose S_1 seeds the centre. Seeding a non-central receiver is optimal for S_2 if:

$$(M-2)\frac{1}{\sigma^2+\eta_1^2+1} + \frac{2}{\sigma^2+\eta_1^2} \leq (M-1)\frac{2}{2(\sigma^2+\eta_1^2)+1} + \frac{1}{\sigma^2+\eta_1^2}$$

Rearranging:

$$\frac{1}{\sigma^2+\eta_1^2} \leq \frac{M+2(\sigma^2+\eta_1^2)}{(\sigma^2+\eta_1^2+1)(2(\sigma^2+\eta_1^2)+1)}$$

This tells us $M \geq 3+\frac{1}{\sigma^2+\rho}$ is needed for seeding a non-central receiver be optimal for S_2 (and symmetrically for seeding the centre to be optimal for S_1). Hence S_1 seeding the centre and S_2 seeding a non-central receiver is a pure strategy equilibrium. If $M < 3 + \frac{1}{\sigma^2+\rho}$ then if S_1 seeds the centre S_2 also prefers to seed the centre (and thus if S_2 seeds the centre S_1 prefers not to), so it follows that there exists no pure

strategy equilibrium.

Note a corollary is that if $M \geq 3 + \frac{1}{\sigma^2+\rho}$ and we had $w_{S_1} = w_{S_2}$ and $w \notin [0, 2w_{S_1}]$ both senders seeding the centre of a star network would be an equilibrium; in other words, with *two* misaligned senders, incentives to reduce the usefulness of messages by taking the same seeding decision can dominate incentives to increase noise in messages by seeding less central receivers.

Our example illustrates that if signals are sufficiently uninformative, incentives for a misaligned sender to release information in the same way as an aligned sender can be strong enough that they lead the misaligned sender to release information to the most central receiver on a network. Releasing all information through the same receiver will often not maximise the sum of receivers' expected payoffs in the presence of noisy transmission, and so if we observe multiple senders seeding correlated pieces of information at the same point on a network, we may worry that at least one of those senders has interests that are not aligned with the receivers.

6 Preferences for Co-ordination

We consider an extension in which we allow receivers to have preferences over one another's actions, and to want to co-ordinate their actions with their neighbours. We focus on the case case of homogeneous receiver preferences, both for tractability and because receivers with heterogeneous preferences are less likely to gossip non-strategically when they have co-ordination concerns.

Note that from a modelling perspective, a significant difference when receivers want to co-ordinate and have aligned preferences is that receivers now have a strict incentive to communicate more clearly, both when listening to messages *and* when sending them. As such, if receivers are able to take steps to reduce noise in communication, or had strategic tools with which to bypass noise, we would expect them to take these steps (in contrast in the base model receivers have incentives to *listen*

well but not to *speak* clearly). We will continue to use the same model of noisy and non-strategic information-transmission. We emphasise that when interpreting the model, the noise in messages should be viewed as the remaining noise once receivers have taken all *proportionate* actions to improve clarity of communication, and that our model may be less appropriate if receivers view the stakes of their decisions as being large.

Our analysis in this section has links to Myatt and Wallace (2019), who study how much weight networked players with co-ordination concerns place on signals which differ in both noise level and correlation between players.

6.1 Model with co-ordination concerns

Sender preferences are unchanged. Receiver R_j's preferences are now as follows:

$$v_j(\boldsymbol{x}, \theta) = -(1-\beta)(x_j - \theta)^2 - \beta \sum_{k=1}^{M} (x_j - x_k)^2 \tag{10}$$

We assume receiver preferences are homogeneous, so it is wlog to set the receiver's ideal action equal to θ. We also assume $\beta \in (0,1)$, so that receivers want to co-ordinate with both the state and one-another, and that β is sufficiently low that actions cannot explode due to strategic complementarity. We continue to assume the same information transmission process, and further assume that if there are multiple shortest paths a message could travel, one is selected according to a commonly known rule.

We model receivers as wanting to co-ordinate their action with *all* other receivers (rather than, for example, their neighbours on $\mathcal{G}$). A large literature on network games studies settings in which networked agents take actions when their neighbours' actions are strategic substitutes or complements (see for example Ballester, Calvó-Armengol, and Zenou (2006)). Our focus here is on capturing *information transmission* through networks, so we use the simplest possible model of co-ordination

concerns; our findings in this setting would be robust to co-ordination concerns being restricted to neighbours on a network. To give an example of a setting which fits our model, members of a political party may benefit from aligning their positions with *all* other party members, even if they have a more limited social circle.

6.2 Equilibrium Outcomes with Co-ordination Concerns

We focus on a linear BNE of the action choice subgame. Suppose R_j has a strategy of the following form: $x_j = \omega_j \mu + (1 - \omega_j) m_j$ where $\omega_j \in [0, 1]$. Let c_{jk} denote the variance of shared noisy gossip terms in m_j and m_k's messages.

Substituting these strategies into receivers' expected payoffs and taking the FOC with respect to ω_j we obtain:

$$1 - \omega_j = \frac{\sigma^2(1-\beta)}{(\sigma^2 + \eta^2 + d_j)(1 + \beta(M-2))} + \frac{\beta}{1 + \beta(M-2)} \sum_{k=1, k \neq j}^{M} (1 - \omega_k) \frac{\sigma^2 + \eta^2 + c_{jk}}{\sigma^2 + \eta^2 + d_j}$$

The FOC finds a maximum if β satisfies $1 + \beta(M-2) > 0$, for which we can see $\beta > 0$ is a sufficient condition (this is equivalent to requiring utility to be concave in a receiver's own action).

We can interpret this term by term. The first term tells us that if we release information further from R_j, then, ignoring any effect this has on other receivers, R_j places more weight on their prior. The second term captures how the optimal weight to place on a message for R_j depends both on the weights other receivers place on *their* messages, and the correlation *between* messages. It tells us that the more weight other receivers place on their messages, the more weight a receiver should place on their own message, and that the more noise in the receiver's message is shared with other receivers, the more weight the receiver should place on their own message.

Given receiver strategies, we can write the sender's payoff as follows:

$$E(u_i) = -\sum_{j=1}^{M} E((\omega_j \mu + (1-\omega_j)m_j - w_s\theta)^2)$$

Note we must have $E(x_j) = \mu$ in equilibrium. The sender's problem is hence equivalent to:

$$\max_{l} \sum_{j=1}^{M} (1-\omega_j)\left[2w_s\sigma^2 - (1-\omega_j)(\sigma^2+\eta^2+d_j)\right]$$

If the sender shares the same bliss point as receivers ($w_s = 1$), if we fix the message distribution the sender is best off when $1-\omega_j = \frac{\sigma^2}{\sigma^2+\eta^2+d_j}$; in other words, if receivers act as in the base model with no co-ordination concerns. We can see that when receivers have co-ordination concerns ($\beta > 0$) we will have $1-\omega_j < \frac{\sigma^2}{\sigma^2+\nu^2+d_j}$; so receivers place *less* weight on their messages than they would in the base model, because they know they all share a common prior and so are less likely to misco-ordinate if they place more weight on their prior. As such, even if the sender has the same full-information bliss point as receivers, we can see that from their perspective receivers will not place enough weight on messages.

Co-ordination concerns for the receivers will affect the sender's optimal seeding decision relative to our base model. For example, if we have $2w_s = 1-\epsilon$, such that receiver preferences in the base model are misaligned with the sender (as $w = 1 \notin [0, 2w_s]$), in the base model the sender would seed the least central receiver, while here they may still seed the *most* central if co-ordination concerns have reduced the attention paid to messages sufficiently.

The sender can now influence receiver behaviour via two channels: they can affect the *amount* of noise in receiver messages, which affects the ability of the receiver to co-ordinate their action with the state; they can also affect the *correlation* of noise between two receivers' messages, affecting the ability of the receivers to co-ordinate their actions with one another. As we show in the following example, the interaction between these two channels can mean a sender sometimes finds it optimal to seed a

receiver of *intermediate* centrality.

Example: Seeding Intermediate Centralities Let's consider $M = 4$, with $\mathcal{G}$ a simple graph with R_4 unconnected to anyone and R_1, R_2, and R_3 arranged in a line, with R_2 in the middle. Set $\sigma^2 = 1$, $\eta^2 = 0$, and $\beta = \frac{1}{2}$. In this parameterisation, we can state the following:

Proposition 9. *In this parameterised example, the sender's optimal seeding choice is:*

- R_4 *if* $w_s \leq \frac{146}{645}$
- R_1 *or* R_3 *if* $w_s \in [\frac{146}{645}, \frac{54134}{176085}]$,
- R_2 *if* $w_s \geq \frac{54134}{176085}$

Hence if $w_s \in [\frac{146}{645}, \frac{54134}{176085}]$ *the sender does not seed the receiver with the highest or lowest PC value in the presence of co-ordination concerns.*

From this example, we can observe it is possible for co-ordination concerns to soften our 'bang-bang' characterisation of optimal seeding. When receivers' have co-ordination concerns, the weight R_j places on a message $1 - \omega_j$ depends not just on m_j's informativeness but also on how likely it is neighbours are also basing their actions on that message. It may thus be the case that a sender does not seed the receiver with the highest *or* lowest PC value, since if they seed the lowest PC value receiver (in our example, R_4), receiver caution due to co-ordination concerns leads receivers on aggregate to *underreact* to information from the sender's perspective, while if they seed the receiver with highest PC value (R_2) receivers find it easy to co-ordinate and so on aggregate *overreact* from the sender's perspective. As such, for intermediate preference misalignment between the sender and receivers the sender may do better seeding a receiver of intermediate centrality (R_1 or R_3), making co-ordination sufficiently difficult that receivers react to messages with a sensitivity that is closest to what the sender would like.

Intuitively, information seeding incentives cease to be 'bang-bang' with co-ordination concerns because without co-ordination concerns the sender's seeding decision only influences receiver action choices by changing posterior variance, while with co-ordination concerns senders also influence receiver action choices by changing posterior covariances. A sender who would like a receiver to place less weight on messages than they would with no co-ordination concerns may not find it worthwhile increasing variance in messages in order to induce this reduced weight, since this may make their own payoff more volatile. However, by releasing information in a way which makes co-ordination harder, they may be able to achieve reduced weight without increasing the volatility of their own payoff. We can see this reasoning will extend easily to settings in which receiver co-ordination concerns are asymmetric (for example, when they only want to co-ordinate with their neighbours on $\mathcal{G}$).

Receiver concerns for co-ordination in this setting *expands* the range of sender preferences for which the sender is willing to seed the most central receiver. As such, there is a sense in which *increasing* co-ordination concerns of receivers can *improve* aggregate information via seeding. If we moved to a setting in which receivers wanted to co-ordinate with their neighbours on a network, one implication of this would be that co-ordination concerns may induce senders to provide information more 'efficiently' when networks are more connected. This has implications for network design; for example, when senders are not too misaligned with receivers, it may be that efficient information release is possible when the network capturing co-ordination concerns is a connected network, but not when it is a star network. As such, in our setting if communication networks were also co-ordination networks (for example, when they are social networks), adding links could increase aggregate information quality when a sender is misaligned, even if none of the new links are used to transmit information. This contrasts existing results in the strategic transmission literature (for example, Squintani (2018)) in which stars are often found to be the most efficient network structures for ideological groups, as they allow maximum

strategic transmission for a given number of edges.

7 Discussion

We have shown that, if information is noisily transmitted through a network of homogeneous receivers (due to the 'broken telephone' effect), a sender with quadratic loss preferences in receiver actions will prefer to seed the most central receiver if preferences are sufficiently aligned, and the least central receiver if preferences are misaligned. We have shown this result depends upon receivers being able to observe the point of information release, and that if receivers do not take into account the actual amount of noise in their messages when forming beliefs, the sender always prefers to seed more central receivers (where centrality is defined relative to receiver beliefs).

We have shown that the presence of even one receiver relatively aligned with the sender on a network where all other receivers are extremely misaligned may be sufficient to cause the sender to switch from seeding the least central receiver to seeding the most central receiver. If the sender is forced to seed one receiver, even one relatively aligned receiver on a network is enough to discipline the information provision of a sender who would prefer all receivers to learn nothing about the state of the world.

In a setting of competing senders who observe correlated signals, we have shown it can be optimal for a sender who is trying to minimise the amount of information provided to receivers to seed the most central receiver, if the other sender also does so. As such, when we interact with multiple senders, knowing that they all choose to release information in the same way is not sufficient to tell us whether their interests are aligned with our own.

It is common in the literature on non-strategic information transmission on networks to abstract away from 'broken telephone' effects, and focus solely on random

diffusion processes. In this paper, in contrast, we have focused on the 'broken telephone' effect but assumed that the diffusion paths of messages are known by the sender. However, it would be easy to accommodate a random routing rule for messages, by allowing the gossip network $\mathcal{G}$ to be drawn from some distribution Γ after the sender has taken their seeding decision. For example, Γ might be formed by taking some underlying social network, and assigning a meeting probability to each edge (it could also accommodate uncertainty about the structure of the *underlying* social network). In this case the sender would take their seeding decision to maximise their expected payoff across all possible gossip networks, so an aligned sender would seed the receiver with the highest expected 'Persuasion Centrality'.

If we think of Γ as a distribution over subgraphs of some original social network $\mathcal{G}'$, expected 'Persuasion Centrality' across subgraphs may end up differing significantly from 'Persuasion Centrality' on the underlying graph. If we have a social network in which meetings are random, when considering the optimal receiver to seed we will pay more attention to immediate neighbours, since information is less likely to reach those further from the point of release. If a social network is divided into cliques, we might expect the optimal person to seed for an aligned sender when meetings are random to be someone who is central in the largest clique, even if they are not 'Persuasion Central' on the underlying social network, since the probability of a message reaching muliple cliques is low.

As we have noted, one limitation of our analysis is our single message assumption. We restricted attention to single messages for tractability, but when information can be transmitted by short rather than lengthy conversations, it may be that in a noisy gossip process receivers pass news on even to friends who have already heard it from another source. As such, there is scope to analyse a more general model, in which a message travels along each path between a sender and a receiver, meaning a receiver is at an informational advantage if they can be accessed by multiple paths *and* may learn more from multiple longer paths than they would from a single short path. A

more general model is less analytically tractable, but may be a fruitful setting for simulation based approaches.

In many settings in which we are concerned about the 'broken telephone' effect in communication, actions are not continuous but discrete (for example, if we consider a one-off referendum, it is plausible a sender looking to influence the vote is concerned only with the actual voting decisions receivers take and not the ideological positions that determined them). As such, there is scope for further work analysing a similar model in a setting of discrete action choice; Jackson, Malladi, and McAdams (2019) model learning via the 'broken telephone' effect with a binary state, where messages can randomly flip in transmission (their model also features other frictions); a similar model may offer a tractable route to studying optimal seeding in discrete choice settings. It may also be easier to incorporate strategic elements to transmission in addition to mechanical noise when actions and messages are discrete.

Our analysis in this paper has been built on our 'bang-bang' result for information provision; we might wonder how far this generalises to other loss functions. We know from Rehak and Senkov (2021) that incentives for information provision are not 'bang-bang' when a sender experiences quadratic loss but has an ideal action that is non-linear in θ. However, we might wonder whether incentives for information provision will be 'bang-bang' for any convex loss function if we hold the sender's ideal action fixed (and linear in θ). In fact we can show the answer is no: a sender with $u(\theta, \boldsymbol{x}) = -\sum_{j=1}^{M} \theta^2(x_j - w_s\theta)^2$ shares the same ideal action with the sender in our base model but for $4w_s\sigma^2 - 5w_j\mu^2 - (w_j - 2w_s)(3\mu^2 + \sigma^2) > 0$, their preferences are not monotone in d_j, and they would prefer R_j to receive an intermediate amount of information.

We would also not expect incentives for information provision to be 'bang-bang' if sender preferences were not separable in loss from each receiver. For example, if senders themselves had preferences over the co-ordination of receiver actions, we would expect this to affect their information release policy. If they cared strongly

about co-ordination between receivers and only weakly about coordinating actions with the state (and receivers did not care about co-ordinating) even if we had $w_s = w_j \ \forall R_j \in \mathcal{N}$, it might be that a sender preferred to seed a less central receiver, to reduce the volatility of receiver actions and hence reduce expected loss via misco-ordination between receivers.

References

Akbarpour, Mohammad, Suraj Malladi, and Amin Saberi (2020). "Just a Few Seeds More: Value of Network Information for Diffusion". working paper. URL: http://dx.doi.org/10.2139/ssrn.3062830.

Ballester, Coralio, Antoni Calvó-Armengol, and Yves Zenou (2006). "Who's Who in Networks. Wanted: The Key Player". In: *Econometrica* 74.5, pp. 1403–1417.

Banerjee, Abhijit et al. (2013). "The Diffusion of Microfinance". In: *Science* 341.6144, p. 1236498. DOI: 10.1126/science.1236498. eprint: https://www.science.org/doi/pdf/10.1126/science.1236498. URL: https://www.science.org/doi/abs/10.1126/science.1236498.

— (2019). "Using Gossips to Spread Information: Theory and Evidence from Two Randomized Controlled Trials". In: *The Review of Economic Studies* 86.6, pp. 2453–2490. ISSN: 0034-6527. DOI: 10.1093/restud/rdz008. eprint: https://academic.oup.com/restud/article-pdf/86/6/2453/30302450/rdz008.pdf. URL: https://doi.org/10.1093/restud/rdz008.

Bloch, Francis, Matthew O. Jackson, and Pietro Tebaldi (2019). "Centrality Measures in Networks". working paper. URL: http://dx.doi.org/10.2139/ssrn.2749124.

DeGroot, Morris H. (1974). "Reaching a Consensus". In: *Journal of the American Statistical Association* 69.345, pp. 118–121. ISSN: 01621459. URL: http://www.jstor.org/stable/2285509 (visited on 08/06/2022).

Galeotti, Andrea, Christian Ghiglino, and Francesco Squintani (2013). "Strategic information transmission networks". In: *Journal of Economic Theory* 148.5, pp. 1751–1769. ISSN: 0022-0531. DOI: https://doi.org/10.1016/j.jet.2013.04.016. URL: https://www.sciencedirect.com/science/article/pii/S0022053113000823.

Galeotti, Andrea and Sanjeev Goyal (2009). "Influencing the influencers: a theory of strategic diffusion". In: *The RAND Journal of Economics* 40.3, pp. 509–532.

Galperti, Simone and Jacopo Perogo (2020). "Information Systems". URL: https://drive.google.com/file/d/1t4hbiMcoh01ldKUmee_VI8ZErmekGhkq/view.

Gieczewski, Germán (2022). "Verifiable communication on networks". In: *Journal of Economic Theory* 204, p. 105494. ISSN: 0022-0531. DOI: https://doi.org/10.1016/j.jet.2022.105494. URL: https://www.sciencedirect.com/science/article/pii/S0022053122000849.

Golub, Ben and Evan Sadler (2016). "Learning in Social Networks". In: *Oxford Handbook of Economic Networks*. URL: https://dx.doi.org/10.2139/ssrn.2919146.

Hagenbach, Jeanne and Frédéric Koessler (2010). "Strategic Communication Networks". In: *The Review of Economic Studies* 77.3, pp. 1072–1099. ISSN: 0034-6527. DOI: 10.1111/j.1467-937X.2009.591.x. eprint: https://academic.oup.com/restud/article-pdf/77/3/1072/18365849/77-3-1072.pdf. URL: https://doi.org/10.1111/j.1467-937X.2009.591.x.

Jackson, Matthew O., Suraj Malladi, and David McAdams (2019). "Learning through the Grapevine and the Impact of the Breadth and Depth of Social Networks". URL: http://dx.doi.org/10.2139/ssrn.3269543.

Kamenica, Emir and Matthew Gentzkow (2011). "Bayesian Persuasion". In: *American Economic Review* 101.6, pp. 2590–2615. DOI: 10.1257/aer.101.6.2590. URL: https://www.aeaweb.org/articles?id=10.1257/aer.101.6.2590.

MacKay, David J. C. (2002). *Information Theory, Inference & Learning Algorithms*. New York, NY, USA: Cambridge University Press. ISBN: 0521642981.

Myatt, David P. and Chris Wallace (2019). "Information acquisition and use by networked players". In: *Journal of Economic Theory* 182, pp. 360–401. ISSN: 0022-0531. DOI: https://doi.org/10.1016/j.jet.2019.05.002. URL: https://www.sciencedirect.com/science/article/pii/S0022053119300493.

Navarro, Noemí (2012). "Price and quality decisions under network effects". In: *Journal of Mathematical Economics* 48.5, pp. 263–270. ISSN: 0304-4068. DOI:

https://doi.org/10.1016/j.jmateco.2012.06.002. URL: http://www.sciencedirect.com/science/article/pii/S0304406812000420.

Nei, Stephen (2021). "Social Context and Social Learning". working paper. URL: https://drive.google.com/file/d/1ZGqP7rMp-e4zeR4VZh8itPKsrmSd2UDC/view.

Rehak, Rastislav and Maxim Senkov (2021). "Form of Preference Misalignment Linked to State-Pooling Structure in Bayesian Persuasion". working paper. URL: https://drive.google.com/file/d/10zSvbeqb-hPjdMF90elxCWR8TxoZ59J1/view.

Rochat, Yannick (2009). "Closeness Centrality Extended To Unconnected Graphs : The Harmonic Centrality Index". In: URL: https://infoscience.epfl.ch/record/200525/files/[EN]ASNA09.pdf;.

Squintani, Francesco (2018). "Networks and Ideology". In: URL: https://warwick.ac.uk/fac/soc/economics/staff/fsquintani/research/networks_ideology.pdf.

A Proofs for Section 3 (Optimal Seeding)

Proposition 1 'bang-bang' information provision. *If $w_j \in (0, 2w_s)$ then S's expected payoff is increasing in $Var(E(\theta|m_j))$, while if $w_j \notin [0, 2w_s]$ S's expected payoff is decreasing in $Var(E(\theta|m_j))$.*

Proof. If $w_j \in (0, 2w_s)$ then equation (3) is increasing in $E([E(\theta|m_j)]^2)$, and if $w_j \notin [0, 2w_s]$ it is decreasing in $E([E(\theta|m_j)]^2)$. We know that $E(E(\theta|m_j)) = \mu$, so $Var(E(\theta|m_j)) = E([E(\theta|m_j)]^2) - \mu^2 \geq 0$. Hence if expected payoffs are increasing in $E([E(\theta|m_j)]^2)$ they are also increasing in $Var(E(\theta|m_j))$. □

Proposition 2. *If $w_j = w \ \forall j$, then if $w \in (0, 2w_s)$, it is optimal for the sender to seed the receiver with the highest 'Persuasion Centrality', while if $w \notin [0, 2w_s]$, it is optimal for the sender to seed the lowest 'Persuasion Centrality'.*

Proof. From (7), if $w \in (0, 2w_s)$ the sender has the following problem:

$$\max_l \sum_{j=1}^{M} \frac{1}{\sigma^2 + \eta^2 + d_j}$$

and if $w \notin [0, 2w_s]$ the sender has the following problem:

$$\min_l \sum_{j=1}^{M} \frac{1}{\sigma^2 + \eta^2 + d_j}$$

We can decompose S's objective function as follows:

$$\sum_{j=1}^{M} \frac{1}{\sigma^2 + \eta^2 + d_j} = PC(R_l, \mathcal{G}) + \frac{1}{\sigma^2 + \eta^2}$$

As the second term takes the same value regardless of which receiver is linked to, we can thus see that if $w \in (0, 2w_s)$ it is optimal for S to seed R_{j^*}, where:

$$R_{j^*} = \underset{R_j \in \mathcal{N}}{\operatorname{argmax}} \, PC(R_j, \mathcal{G})$$

while if $w \notin [0, 2w_s]$ it is optimal for S to seed R_{j*}, where:

$$R_{j*} = \underset{R_j \in \mathcal{N}}{\operatorname{argmin}} PC(R_j, \mathcal{G})$$

Hence if $w \in (0, 2w_s)$, S wants to seed the receiver with the highest PC value, while if $w \notin [0, 2w_s]$, S wants to seed the receiver with the lowest PC value. □

Proposition 3. *For any connected network $\mathcal{G}$, above some threshold value of σ^2 ranking nodes by PC value produces the same ordering as ranking nodes by closeness centrality.*

Proof. This proposition follows immediately from proposition 7 and corollary 1 □

Proposition 4. *If receiver preference parameters w_j are unobserved but identically distributed with mean μ_w and variance σ_w^2, then if $\sigma_w^2 < w_s^2$ and $\mu_w \in (w_s - \sqrt{w_s^2 - \sigma_w^2}, w_s + \sqrt{w_s^2 - \sigma_w^2})$ the sender prefers to seed the receiver with the highest PC value, else they seed the receiver with the lowest PC value.*

Proof. The sender's expected payoff in this setting is:

$$\begin{aligned}
E(u) &= -\sum_{j=1}^{M} E((x_j - w_s\theta)^2) \\
&= -\sum_{j=1}^{M} \left[E(w_j^2)E([E(\theta|m_j)]^2) - 2w_s E(w_j)E(E(\theta|m_j)\theta) + E(\theta^2) \right] \\
&= -\sum_{j=1}^{M} \left[(\sigma_w^2 + \mu_w^2)E([E(\theta|m_j)]^2) - 2w_s\mu_w E([E(\theta|m_j)]^2) + \sigma^2 + \mu^2 \right]
\end{aligned}$$

and, considering only the parts that d_j can affect, the sender's problem is:

$$\max_{\boldsymbol{l}} -\sum_{j=1}^{M} (\sigma_w^2 + \mu_w^2 - 2w_s\mu_w)E([E(\theta|m_j)]^2)$$

Hence our condition for when S would like to seed the receiver with the highest PC value (i.e when they would like to maximise $\sum_{j=1}^{M} E([E(\theta|m_j)]^2)$) is $\sigma_w^2 + \mu_w^2 -$

$2w_s\mu_w < 0$. We can thus see that S will want to seed the receiver with the highest PC value if $\mu_w \in (w_s \pm \sqrt{w_s^2 - \sigma_w^2})$. The width of this interval is decreasing in σ_w^2, and we can see that if $\sigma_w^2 \geq w_s^2$ the interval is degenerate, so if $\sigma_w^2 > w_s^2$ the sender always prefers to seed the receiver with the lowest PC value. □

Proposition 5. *If we have $M-1$ receivers with $w_j = w' + c_j$ and 1 receiver R^* with $w_j = w^*$, then as $|w'| \to \infty$, it is optimal for the sender to seed the receiver with the highest PC value if:*

$$\min_{\{R_j | w_j = w' + c_j\}} \left\{ \sum_{j=1, j\neq k}^{M-2} \frac{1}{\sigma^2 + \eta^2 + d(k,j)} \right\} + \frac{1}{\sigma^2 + \eta^2}$$
$$> PC(R^*, \mathcal{G}) > \max_{\{R_j | w_j \propto w'\}} \left\{ PC(R_j, \mathcal{G}) \right\}$$

Proof. To prove this we will first prove that if R^*'s 'Persuasion Centrality' is below a certain threshold, seeding them is optimal, and then show that this threshold does not rule out R^* being 'Persuasion Central'.

We will consider the argument for $w' \to \infty$ (the argument for $w' \to -\infty$ is analogous).

Firstly, note that we can rewrite the sender's objective function as follows, dividing through by w'^2, as dividing by a positive constant does not affect the maximisation problem:

$$\max_{\iota} - \sum_{j=1}^{M} \frac{w_j(w_j - 2w_s)}{w'^2(\sigma^2 + \eta^2 + d_j)}$$

Seeding R^* gives S the following payoff:

$$\sum_{j=1, j\neq *}^{M-1} \frac{w_j(2w_s - w_j)}{w'^2(\sigma^2 + \eta^2 + d(R^*, R_j))} + \frac{w^*(2w_s - w^*)}{w'^2(\sigma^2 + \eta^2)}$$

while if R_k is a generic receiver with $w_k = w' + c_k$. seeding R_k gives:

$$\sum_{j=1, j\neq k}^{M-2} \frac{w_j(2w_s - w_j)}{w'^2(\sigma^2 + \eta^2 + d(R^*, R_j))} + \frac{w^*(2w_s - w^*)}{w'^2(\sigma^2 + \eta^2 + d(R^*, R_k))} + \frac{w_k(2w_s - w_k)}{w'^2(\sigma^2 + \eta^2)}$$

Now let $w' \to \infty$. The rescaled payoff from seeding R^* tends to:

$$-\sum_{j=1,j\neq *}^{M-1} \frac{1}{\sigma^2+\eta^2+d(R^*,R_j)}$$

while the rescaled payoff from seeding R_k tends to:

$$-\sum_{j=1,j\neq k,*}^{M-2} \frac{1}{\sigma^2+\eta^2+d(R_k,R_j)} - \frac{1}{\sigma^2+\eta^2}$$

So our condition for seeding R^* to be better than seeding R' is:

$$\sum_{j=1,j\neq '}^{M-2} \frac{1}{\sigma^2+\eta^2+d(R_k,R_j)} + \frac{1}{\sigma^2+\eta^2} > \sum_{j=1,j\neq *}^{M-1} \frac{1}{\sigma^2+\eta^2+d(R^*,R_j)}$$

The RHS is just the PC value of R^*. Hence we have an upper bound on R^*'s 'Persuasion Centrality' for seeding R^* to be preferred to seeding R_k. The inequality above needs to hold for all R_k with $w_k = w' + c_k$. Hence for seeding R^* to be optimal we have the following upper bound on R^*'s 'Persuasion Centrality':

$$\min_{\{R_j|w_j=w'+c_j\}} \left\{ \sum_{j=1,j\neq k}^{M-2} \frac{1}{\sigma^2+\eta^2+d(R_k,R_j)} \right\} + \frac{1}{\sigma^2+\eta^2} > PC(R^*,\mathcal{G}) \tag{11}$$

If we consider the LHS of equation (11), we can see that:

$$\sum_{j=1,j\neq l}^{M-2} \frac{1}{\sigma^2+\eta^2+d(k,j)} + \frac{1}{\sigma^2+\eta^2} > PC(R_k,\mathcal{G})$$

for all R'. Hence, it is possible that:

$$\min_{\{R_j|w_j=w'+c_j\}} \left\{ \sum_{j=1,j\neq k}^{M-2} \frac{1}{\sigma^2+\eta^2+d(k,j)} \right\} + \frac{1}{\sigma^2+\eta^2}$$
$$> PC(R^*,\mathcal{G}) > \max_{\{R_j|w_j=w'+c_j\}} \left\{ PC(R_j,\mathcal{G}) \right\}$$

and if this is the case, we have the result that seeding the receiver with the highest

PC value (and minimising aggregate posterior variances) is optimal even when all but one receiver is infinitely misaligned with the sender in preferences. □

B Proofs for Section 4 (Unobserved Seeding)

Proposition 6. *If $w_j = w \; \forall j$ (with $w \neq 0$), in any pure strategy equilibrium S must seed the receiver with highest ACC value for some equilibrium belief vector $\boldsymbol{\alpha} > 0$.*

Proof. In a candidate pure strategy PBE in which R_j expects S to locate themselves $\bar{d}_j$ away, S's expected payoff is given by:

$$E(u(\boldsymbol{x}, \theta) = -\sum_{j=1}^{M} E((w_j E(\theta | m_j, \bar{d}_j) - w_s \theta)^2)$$

Expanding and rearranging, we obtain:

$$E(u(\boldsymbol{x}, \theta) = -\sum_{j=1}^{M} \left[\frac{w_j^2 \sigma^4 d_j}{(\sigma^2 + \eta^2 + \bar{d}_j)^2} + \frac{w_j^2 \sigma^4 (\sigma^2 + \eta^2)}{(\sigma^2 + \eta^2 + \bar{d}_j)^2} + (w_j - w_s)^2 \mu^2 \right.$$
$$\left. + 2 w_j w_s \frac{\sigma^4}{\sigma^2 + \eta^2 + \bar{d}_j} + w_s^2 \sigma^2 \right]$$

Note d_j only enters this expression in one term. As such we can rephrase the sender's problem as:

$$\min_{l} \sum_{j=1}^{M} \frac{w_j^2 d_j}{(\sigma^2 + \eta^2 + \bar{d}_j)^2}$$

If $w_j = w \; \forall j$ and $w \neq 0$, we can rewrite the problem as follows:

$$\min_{i} \sum_{j=1, j \neq i}^{M} \frac{d(R_i, R_j)}{(\sigma^2 + \eta^2 + \bar{d}_j)^2}$$

We can see our objective function is equivalent to $ACC(R_i, \boldsymbol{\alpha}, \mathcal{G})^{-1}$ when $\boldsymbol{\alpha}$ has jth element $\alpha_j = \frac{1}{(\sigma^2 + \eta^2 + \bar{d}_j)^2}$. Note in a pure strategy equilibrium we must have $\bar{d}_j = d_j$ for all $R_j \in \mathcal{N}$. As minimising $ACC(R_i, \boldsymbol{\alpha}, \mathcal{G})^{-1}$ is equivalent to maximising $ACC(R_i, \boldsymbol{\alpha}, \mathcal{G})$, an equilibrium is thus given by a seeding decision with $R_l = R_{j^*}$,

where:

$$R_{j^*} = \underset{j\in\{1,\ldots,M\}}{\operatorname{argmax}}\ ACC(R_j, \boldsymbol{\alpha}^*, \mathcal{G})$$

and $\boldsymbol{\alpha}^*$ has jth element α_j^* which satisfies:

$$\alpha_j^* = \frac{1}{(\sigma^2 + \eta^2 + d(R_j^*, R_j))^2}$$

□

Proposition 7. *As σ^2 becomes large, above some threshold the receiver with the highest $ACC(i, \boldsymbol{\alpha}, \mathcal{G})$ value will also have the highest closeness centrality value for $\boldsymbol{\alpha}$ consistent with any conjectured pure strategy.*

Proof. Firstly, note if R_k has a higher closeness centrality value than R_i by definition we have $\sum_{R_j\in\mathcal{N}, j\neq i} d(R_i, R_j) > \sum_{R_j\in\mathcal{N}, j\neq k} d(R_k, R_j)$. Now, supposing R_k does have a higher closeness centrality value than R_i, compare their ACC values for any conjectured pure strategy. The difference between R_k and R_i's ACC value is given by:

$$\frac{1}{\sum_{R_j\in\mathcal{N}, j\neq i} \alpha_j d(R_i, R_j)} - \frac{1}{\sum_{R_k\in\mathcal{N}, j\neq k} \alpha_j d(R_k, R_j)}$$

where $\boldsymbol{\alpha}$ has jth element α_j which satisfies:

$$\alpha_j = \frac{1}{(\sigma^2 + \eta^2 + \bar{d}_j)^2}$$

Rescale the difference in ACC values by multiplying by σ^4. We can see $\lim_{\sigma^4\to\infty} \alpha_j\sigma^4 = 1$. Hence we can see that that as σ^2 becomes large the difference becomes

$$\frac{1}{\sum_{j\in\mathcal{N}, j\neq i} d(i, j)} - \frac{1}{\sum_{j\in\mathcal{N}, j\neq k} d(j, j)}$$

which is just the difference of closeness centralities. Hence as σ^2 becomes large the ranking produced by the ACC measure in any pure strategy equilibrium must become equal to the ranking produced by the CC measure.

We can further note that as the set of possible conjectures about pure strategies receivers could hold is finite, there will in fact be some finite threshold for σ^2 for any graph, above which the σ^4 term in α_j dominates all terms involving $\bar{d}_j$. □

Corollary 1. *As σ^2 becomes large, above some threshold any receiver seeded in a pure strategy equilibrium when seeding is unobserved must have the highest PC value.*

Proof. Note there is always an equilibrium in the seeding unobserved game in which the sender seeds the receiver with the highest PC value. As such, it follows that if for any valid conjecture the receiver with highest ACC value becomes the closeness central receiver as σ^2 becomes large, it must also be the case that as σ^2 becomes large the receiver with the highest 'Persuasion Centrality' also becomes the closeness central receiver, since the receiver with highest PC value always has highest ACC value for some valid conjecture $\boldsymbol{\alpha}$. Hence in any equilibrium with sufficiently large σ^2, it follows the receiver linked to must have the highest PC value. □

C Proofs for Section 5 (Competing Senders)

Proposition 8. *If $\eta_1^2 = \eta_2^2$ there exists no pure strategy equilibrium in which S_2 and S_1 seed the same receiver when $w_{S_1} \in (0, 2w)$ and $w_{S_2} \notin [0, 2w]$.*

Proof. If signals are equally informative then S_1's problem is the exact inverse of S_2's. Hence whenever it is strictly optimal for S_2 to seed a receiver it cannot be optimal for S_1 to seed that receiver, and pure strategy equilibria in which both seed the same receiver will not exist. □

D Proofs for Section 6 (Preferences for Co-ordination)

Proposition 9. *In this parameterised example, the sender's optimal seeding choice is:*

- *R_4 if $w_s \leq \frac{146}{645}$*

- R_1 *or* R_3 *if* $w_s \in [\frac{146}{645}, \frac{54134}{176085}]$,
- R_2 *if* $w_s \geq \frac{54134}{176085}$

Hence if $w_s \in [\frac{146}{645}, \frac{54134}{176085}]$ *the sender does not seed the receiver with the highest or lowest PC value in the presence of co-ordination concerns.*

Proof. First let's consider seeding R_4. No receiver except R_4 observes a message, so we have $1 - \omega_j = 0$ for $j = 1, 2, 3$. R_4 will then set

$$1 - \omega_4 = \frac{\sigma^2(1-\beta)}{(\sigma^2+\eta^2)(1+2\beta)}$$

Given our parameterisation we have sender payoff of $\frac{-1}{4}\left[\frac{1}{4} - 2w_s\right]$.

Now let's think about seeding R_2, and note we must have $\omega_1 = \omega_3$ and $\omega_4 = 0$:

$$1 - \omega_2 = \frac{\sigma^2(1-\beta)}{(\sigma^2+\eta^2)(1+2\beta)} + \frac{\beta}{1+2\beta}2(1-\omega_1)$$

and

$$1 - \omega_1 = \frac{\sigma^2(1-\beta)}{(\sigma^2+\eta^2+1)(1+2\beta)} + \frac{\beta}{1+2\beta}(1-\omega_1)\frac{\sigma^2+\eta^2}{\sigma^2+\eta^2+1} + \frac{\beta}{1+2\beta}(1-\omega_2)\frac{\sigma^2+\eta^2}{\sigma^2+\eta^2+1}$$

which with our assumptions gives:

$$1 - \omega_2 = \frac{1}{4} + \frac{1}{2}(1-\omega_1)$$

Substituting in we obtain $1-\omega_1 = \frac{3}{13}$ and $1-\omega_2 = \frac{19}{52}$. Seeding payoff is $2w_s\frac{43}{26} - \frac{937}{2704}$

Now let's consider seeding R_1. Given our assumptions we have:

$$\begin{aligned}
1-\omega_1 =& \frac{1}{4}[1+[(1-\omega_2)+1-\omega_3]] \\
1-\omega_2 =& \frac{1}{8}+\frac{1}{8}(1-\omega_1)+\frac{1}{4}(1-\omega_3) \\
1-\omega_3 =& \frac{1}{12}+\frac{1}{12}(1-\omega_1)+\frac{1}{6}(1-\omega_2)
\end{aligned}$$

Hence we have $1-\omega_3 = \frac{25}{172}$, $1-\omega_1 = \frac{29}{86}$, and $1-\omega_2 = \frac{35}{172}$ giving payoff $\frac{59}{43}w_s - \frac{7689}{172^2}$.

Our proposition follows immediately. □

Chapter 3

Noisy Disclosure

Noisy Disclosure

Aidan Smith

August 8, 2022

Abstract

Even when experts are unable to lie, they may be misunderstood. We model a setting of verifiable disclosure in which a privately informed sender chooses between disclosing the state and sending no message to a receiver who observes noisy realisations of messages, due to a language barrier. We show that full disclosure will only occur if the sender is sufficiently biased, and that the receiver-optimal disclosure rule does not feature full disclosure. When sender bias is small and noisy messages do not narrow posterior supports, we show both upper and lower censorship are possible equilibrium disclosure strategies. We show that, ex-ante, a receiver prefers to face an unbiased rather than biased sender, and that noise in message creates a commitment problem for biased senders.

Keywords: Information, Information Transmission, Noise

JEL Classification: D82, D83

1 Introduction

Experts are often misunderstood. Even when monitored by a fact-checking regulator, an expert may be unable to communicate perfectly with an uninformed audience. For example, a reader may skim read a peer-reviewed article, or attempt to read it thoroughly but lack the technical understanding to fully comprehend its contents. Readers are often aware of this possibility of miscomprehension; other things equal, when forming opinions we put less weight on articles we know we may have misunderstood.

Even if experts know they may be misunderstood, we often trust them to tell the truth. Casual investors may not fully comprehend the implications of firm disclosures on stock value, but can trust that financial regulators will prevent firms presenting fake balance sheets. An expert may be restricted to truth telling even when their *sole aim* is to influence the behaviour of decision makers with imperfect comprehension; for example, public health messaging will be fact checked by medical professionals but designed to influence the behaviour of those who lack an in-depth understanding of medicine.

We model the problem of an expert who may be misunderstood as a game of noisy verifiable disclosure, in which a sender, who has preferences over the action of a receiver, privately observes a random state of the world, and can choose whether to 'disclose' the state to the receiver or send no message. If they choose to 'disclose'[1], the receiver observes a noisy signal about the state of the world; if they do not 'disclose', the receiver observes nothing. After observing any disclosures, the receiver updates their beliefs using Bayes' rule and takes an action to maximise their state-dependent preferences. We model the receiver as wanting to match their action to

[1]When discussing a noisy setting, we will adopt the convention that 'disclose' means 'having seen the true state, allow the receiver to observe a noisy signal'; in other words, we say the sender discloses if they choose to send a message containing the state via a noisy channel, even if the receiver does not accurately observe the message. This definition is arguably a fair extension of natural language to our setting; it would also cover settings in which a receiver did not engage in Bayesian updating.

the state of the world, and the sender as experiencing quadratic loss in the distance of the receiver's action from the sender's state dependent ideal action.

We show that if bias in the sender's ideal action is sufficiently small, they will not always disclose in equilibrium. If the receiver's posterior has full support on observing the noisy signal, both lower censorship (disclosing states above some threshold) *and* upper censorship (disclosing states below some threshold) are possible in equilibrium, regardless of the direction of sender bias. For a sender with positive bias, lower censorship equilibria feature sceptical 'unravelling' reasoning (with no-news interpreted as bad news), and converge to full disclosure as noise vanishes, while approximating two-message cheap talk as the signal becomes arbitrarily noisy. In contrast upper censorship equilibria feature a self fulfilling prophecy by the receiver, who assumes that a sender in high states will never disclose and thus believes high signals must be the result of noise, preventing unravelling. These equilibria persist even as noise vanishes, and also approximate two-message cheap talk as the signal becomes arbitrarily noisy.

We analyse expected payoffs in a parameterised example, and show that, generically, when disclosure is noisy a biased sender does not use the receiver's ex-ante preferred disclosure rule. In contrast, if the state and noisy signal have continuous densities and equilibria are unique, we find an unbiased sender *does* use the receiver's ex-ante preferred disclosure rule in equilibrium.

Our results have three key positive implications. Firstly, we show that if the sender is not too biased, we should not expect full disclosure in equilibrium if there is any possibility of misunderstanding on the receiver's behalf. In contrast, in noiseless disclosure models full disclosure is the unique equilibrium outcome; we show this prediction is non-robust to the introduction of noise, both in the sense that full disclosure may not be *the* unique equilibrium for a sender with large bias and that it may not even be *an* equilibrium outcome for a sender with small bias.

Secondly, we show that if we introduce even a small amount of noise into a

disclosure model, it may no longer be the case that a receiver should interpret no-disclosure as 'bad news' in equilibrium. We show that it is possible that a sender who prefers to be *overestimated* will engage in *upper* censorship in equilibrium; in other words, they will only disclose states that are sufficiently low, and the receiver will interpret non-disclosure as evidence the state is high. In other words, in a classic disclosure model it is not just the prediction of full disclosure that is not robust to noise; the prediction that the receiver should always be sceptical on seeing no message is also not robust to even a small amount of noise in messaging. This result is important for inference; we show that the *direction* of sender bias is *not identified* by their equilibrium disclosure rule; in other words, we cannot infer a sender's preferences from their messages. For example, observing that over time a think-tank chose only to report evidence that supported high taxes to a non-expert politician need not be evidence that the think-tank is left wing; it is possible that they actually prefer *lower* taxes than the politician, but cannot achieve these by disclosing given the politician assumes disclosures only occur when the tax rate is too low.

Thirdly, we show that if disclosure is via a noisy channel, the receiver is worse off ex-ante when the sender's bias is of larger magnitude. As such, interventions to align the interests of experts with decision makers can be beneficial for decision makers. This contrasts findings in the literature on costly information acquisition[2] and costly disclosure, in which in general a more biased sender may have a higher willingness to pay to acquire and disclose information, benefiting the receiver. We show that while misunderstandings due to noise act as a kind of *endogenous* cost to disclosure, because this cost is experienced by *both* the sender and receiver, an unbiased sender uses exactly the disclosure rule that is ex-ante optimal for the receiver.

[2]See for example Onuchic (2021), who studies the problem of a sender who has access to two-dimensional private information, can commit to a disclosure rule for one of these dimensions, and faces an information acquisition decision. She gives conditions for receiver payoff to be decreasing in the transparency of the sender's motives, which results because a sender whose motives are more ambiguous can gain more by acquiring information to disclose.

The key natural benchmark for noisy disclosure is classic models of noiseless verifiable disclosure. Milgrom (1981) and Grossman (1981) study a sender with state-independent preferences who can make any true statement about a privately observed state of the world, which is perfectly transmitted to a receiver, and show that if the receiver has a unique ideal action in each state, the state is fully revealed in any equilibrium. Seidmann and Winter (1997) extend this 'unravelling result' to settings featuring state-dependent preferences for the sender, and show that for the state to be fully revealed in any equilibrium it is sufficient that the sender always prefers more extreme actions to the receiver in any state.

The unravelling reasoning is as follows; if a receiver is trying to set their action equal to the state of the world (say, the length of time to wash their hands for that is optimal for personal health), and they know a sender always wants them to take a larger action (because the sender cares not only about the receiver's health but also about the probability they spread disease to others), if the sender produces limited evidence of the efficacy of handwashing, the receiver concludes this is because they have no better evidence to show, since if they did, they would have done better revealing it. Hence, the receiver interprets any lack of message as sceptically as possible, and as a result the sender does best revealing all their information in equilibrium. When the unravelling result holds, if messaging is noiseless the sender's equilibrium strategy maximises a (risk-averse) receiver's expected payoff.

We show that, compared to the noiseless benchmark, introducing noise to a model of verifiable disclosure is both *mechanically* and *strategically* harmful[3] for the receiver (and sender). Noise is *mechanically* harmful because regardless of the sender's disclosure rule, the receiver will not perfectly learn the state in equilibrium, and so will almost always mismatch their action with the state of the world. Noise is *strategically* harmful because it pushes the receiver's (and sender's) set of possible

[3]This contrasts the findings of Blume, Board, and Kawamura (2007) in a cheap talk setting, who show that while noise is mechanically harmful it may be strategically *helpful* for a receiver communicating with a biased sender, because when the receiver reacts less to messages due to noise a sender with intermediate bias is more willing to send a message indicating the state is low.

equilibrium payoffs below the payoffs achievable with commitment. With noisy disclosure, equilibrium outcomes are no longer unique; off-path beliefs are no longer pinned down in equilibrium, creating a co-ordination problem in beliefs between sender and receiver. Further, even if the sender and receiver co-ordinate on the ex-ante receiver (and sender) optimal equilibrium, unlike in the noiseless benchmark generically a biased sender no longer uses a strategy which maximises receiver (or sender) ex-ante expected payoffs in equilibrium; in other words, they still face a commitment problem.

Compared to the noiseless benchmark, introducing noise to the messaging technology affects both incentives to disclose in equilibrium *and* the ex-ante receiver optimal disclosure rule. To see why equilibrium disclosure incentives are affected, note that non-disclosure induces a certain action for the sender, while disclosure induces a distribution over actions. If the sender experiences convex loss in the distance of a receiver action from their ideal action, other things equal they prefer certainty, so in addition to comparing the average actions induced by disclosure and non-disclosure (as they would with noiseless disclosure), with noisy disclosure the sender must also take into account how volatile the distribution of actions induced by disclosure will be. When bias is small relative to the amount of noise in messages, the risk of a noisy message realisation leading to a large overestimate (or underestimate) of the state will deter the sender from engaging in full disclosure in equilibrium.

To see why the ex-ante receiver optimal disclosure rule depends upon the noise in messaging, consider the limiting case as disclosures become arbitrarily noisy; in that case, the receiver will update minimally based on the actual value of a message, but will still draw inference from the fact that a disclosure was made. As such, we can think of the sender as effectively having two coarse messages, non-disclosure and disclosure, and note that the receiver will clearly be better off when both of these messages are utilised with positive probability. In other words, the receiver is best

off when the sender does not disclose in all states. This logic extends even to cases when noise is relatively small; whenever a sender has access to one perfectly observed (coarse) message, in addition to imperfectly observed noisy messages, the receiver will do best when the sender utilises their perfectly observed[4] (coarse) message with positive probability.

For parameterisations of our model in which messaging is very noisy, the other natural benchmark for noisy disclosure is the canonical cheap talk model of Crawford and Sobel (1982). Crawford and Sobel study the problem of a biased sender who privately observes the state of the world, and can send *any* message to a receiver using a sufficiently rich messaging language. In contrast to disclosure models, the sender is unrestricted by truthtelling; as such, if receiver action is strictly increasing in beliefs and the sender has monotone preferences in receiver action, *no* (payoff relevant) information will be transmitted in equilibrium, because *all* sender types will send the message which induces the highest posterior belief for a given updating rule of the receiver. In contrast, if preference misalignment is small (for example, if the sender likes small overestimates but not large overestimates), it is possible for *some* (payoff relevant) information to be transmitted in equilibrium, with the sender partitioning the state space into intervals, and sending the same message for all states within an interval.

Cheap talk is an important benchmark for noisy disclosure because when messaging is very noisy, the receiver learns very little from the *content* of sender messages, but can still update their beliefs based on the decision of the sender to disclose. Since the sender's *decision* to disclose is not restricted by their private information, information transmitted by this decision is effectively cheap talk, and equilibria in

[4]Even if non-disclosure is not perfectly observed, the receiver will still do best when the sender does not disclose with positive probability, provided non-disclosure is observed with sufficient accuracy. Intuitively, if a sender has access to multiple channels of communication with different 'bandwidths' and different noise levels (e.g. email, postal mail, word of mouth, lighting a beacon etc.) the receiver optimal messaging strategy will trade off 'bandwidth' of channels against their noise level, and so may utilise a mix of message types rather than solely using high-noise-high-bandwidth channels or low-noise-low-bandwidth channels.

which all communication is via the disclosure decision resemble cheap talk equilibria. For example, if sender bias is very large and positive, and the receiver's posterior has full support at each noisy message realisation and believes that a sender would only disclose if the state was sufficiently *small*, then the equilibrium outcome would be no-disclosure by any sender type, with no information transmitted in equilibrium, as in cheap talk.

The joint impact of noise on disclosure incentives *and* the ex-ante optimal disclosure rule means that, in general, equilibria in the noisy disclosure game with biased senders will be not feature ex-ante optimal disclosure rules, because the sender faces a commitment problem (similar to cheap talk settings). When choosing whether or not to disclose, a sender only takes into account their expected payoff given the state. However, the updating rule the receiver uses in equilibrium depends on all the states for which the sender discloses. As such, a sender in a low state disclosing may harm senders in high states who disclose, because the receiver is more uncertain on observing any message, and in equilibrium adjusts their updating rule in response. Because noise affects *both* the optimal disclosure rule *and* disclosure incentives, these 'externalities' can lead to too much disclosure in equilibrium when a sender has positive bias and the receiver is sceptical on observing no message.

In contrast, if the sender is unbiased we show that (with some regularity conditions) their equilibrium disclosure rule exactly coincides with their disclosure rule under commitment, because the condition to make a cutoff type indifferent between disclosure and non-disclosure coincides exactly with the first order condition for maximising their expected payoff. When the sender and receiver are perfectly aligned, as a disclosure rule changes the receiver adjusts their updating rule exactly optimally from the sender's perspective. As a result, a sender with commitment power needs only to consider the *direct* effect of the change in disclosure rule on the cutoff type, and not the *indirect* effect via changes in the receiver's updating rule. As such the optimal disclosure rule is characterised solely by matching a marginal type's payoffs

from disclosure and non-disclosure; in other words, making them indifferent.

This paper is structured as follows; in section 2, we set out our model. Section 3 discusses in a general setting what outcomes may be supported in equilibrium. In section 4 we analyse a parametrised example and discuss the receiver's ex-ante expected payoffs. In Section 5 we discuss three possible directions in which our model could be generalised. We reference related literature where appropriate, but defer a full literature review until Section 6. Section 7 concludes. Proofs are in the appendix.

2 Model

A sender S communicates with a receiver R about the state of the world θ via a noisy channel, after which the receiver takes an action x. The sender's preferences over state and action are given by:

$$U_S(x, \theta, b) = -(x - \theta - b)^2 \tag{1}$$

where $b \in \mathbb{R}$ is common knowledge. The receiver's preferences are given by:

$$U_R(x, \theta) = -(x - \theta)^2 \tag{2}$$

The noisy disclosure game is played as follows:

1. S observes θ.
2. S either chooses to disclose or not disclose θ.
3. If S does not disclose θ, R observes nothing (we denote this information set $\emptyset$). If S chooses to disclose, R observes a noisy signal s.
4. R takes an action $x \in \mathbb{R}$.

When the sender discloses, we will write the receiver's strategy as $x(s)$. We will write x_0 for the receiver's action on observing no-disclosure.

Distributions S and R share a common prior that $\theta \in [0,1]$ is distributed according to cdf $F(\theta)$, with $E(\theta) = \mu$. We assume $F(\theta)$ is atomless[5].

s is a noisy signal which is informative about θ. We assume $s \in \mathbb{R}$ is atomless, and that for $\theta_1 < \theta_2$, $F(s|\theta_2)$ first order stochastically dominates $F(s|\theta_1)$, which implies that $E(\theta|s, I)$ will be increasing in s for any conjecture I the receiver makes about the sender's strategy. An example of the sort of noise process we are interested in modelling is when $s = \theta + \epsilon$, where ϵ is a mean-zero unimodal error term satisfying $E(\theta\epsilon) = 0$.

Solution Concept We are interested in what outcomes and expected payoffs can be achieved in Perfect Bayesian Equilibrium (appropriately generalised to a continuous setting). Informally, a Perfect Bayesian Equilibrium (henceforth equilibrium) is defined as a strategy for each player, and beliefs at each information set, such that both players are best responding to the strategy of the other player given beliefs, and beliefs at all information sets are derived by Bayes' rule wherever possible.

Formally, let D be a Borel subset of $[0,1]$, which represents the states in which the sender discloses. A pure strategy equilibrium will comprise a conjecture $\hat{D}$ for the receiver, an updating rule for the receiver, a strategy for the receiver, and a disclosure set D for the sender. The receiver's updating rule satisfies the following:

- If $\hat{D}$ has positive measure and signal realisation s is such that $f_{\theta|s}(\theta) > 0$ for some $\theta \in \hat{D}$, the receiver believes the state has density $\hat{f}_{\theta|s,\theta\in\hat{D}}(\theta) = \frac{f_{\theta|s}(\theta)}{\int_{\theta\in\hat{D}} f_{\theta|s}(\theta)}$. If there is no $\theta \in \hat{D}$ such that $f_{\theta|s}(\theta) > 0$, following signal realisation s the sender assigns zero probability to any type such that $f_{\theta|s}(\theta) = 0$,

[5]We restrict attention to the case of a bounded state for the sake of comparison with noiseless disclosure games; in a noiseless disclosure game with $b > 0$ and $\theta \in (-\infty, \bar{\theta})$, Perfect Bayesian Equilibria will fail to exist. Given θ has bounded support above and below, further restricting attention to the unit interval is wlog.

- If $\hat{D}$ does not have positive measure, following signal realisation s the sender assigns zero probability to any type such that $f_{\theta|s}(\theta) = 0$, and for a signal realisation s such that $f_{\theta|s}(\theta) > 0$ for some $\theta \in \hat{D}$, assigns all probability to $\theta \in \hat{D}$.

- If $\hat{D}'$ has positive measure, on observing no disclosure the receiver believes the state has density $\hat{f}_{\theta|\theta\in\hat{D}'}(\theta) = \frac{f_\theta(\theta)}{\int_{\theta\in\hat{D}'} f_\theta(\theta)}$.

- If $\hat{D}'$ does not have positive measure, on observing no disclosure the receiver assigns all probability to $\theta \in \hat{D}'$.

The receiver's strategy is such that at each signal realisation, $x(s)$ maximises the receiver's expected payoff given their beliefs as specified above, and x_0 maximises their expected payoff given their beliefs when they observe no disclosure.

The sender's strategy comprises a set of states in which they disclose D which is a Borel subset of $[0, 1]$, such that $\theta \in D$ iff $\int U(x(s), \theta, b) f_{s|\theta} ds > U(x_0, \theta, b)$.

In equilibrium, the receiver's conjecture $\hat{D}$ satisfies $D = \hat{D}$.

Modelling Assumptions We assume the sender is unable to successfully lie about the state, despite the fact that our receiver is implicitly unable to fact-check the sender's messages themselves. The assumption that the sender will only consider true disclosures in equilibrium will be justified if with sufficiently high probability any disclosures are checked by a regulator who can detect lies and has the power to punish the sender. This would be true of a sender writing in a fact-checked publication, or a firm which faces a sufficiently high probability of having to justify their disclosures in court.

To understand why our receiver does not employ the lie-detection technology that is implicitly available to the regulator, it is best to think of the receiver as facing an underlying multi-dimensional rational inattention problem, and choosing in advance what level of attention to devote to understanding each element of a multi-dimensional state. For example, most economists choose not to study chemistry

to graduate level, and in doing so accept they will imperfectly comprehend articles about research in chemistry. As such, when they read an article about chemistry research, they would expect to understand it less clearly than a Chemistry PhD might; in other words, they are aware that they observe a noisy signal. We are interested in modelling settings in which *at the point of interaction* neither the sender nor the receiver has any available actions to reduce the amount of noise present in the communication technology; the economist lacks time to learn graduate level Chemistry before reading the article.

Given we motivate our question by discussing a receiver with imperfect comprehension, it is worth emphasising that communication between sender and receiver in our model is a *rational* phenomenon which results from communicating via an unavoidably noisy channel. In other words, in our model imperfect comprehension is a feature of the communication technology, rather than a description of the receiver's sophistication. In particular, our receiver understands Bayes' rule; they may not be able to understand the exact contents of an article written in a foreign language, but they can (correctly) infer that if they see no article, the expert must have deemed it not worth writing one.

We focus on a Bayesian receiver, but note that in some settings, it might be more appropriate to assume a receiver with imperfect comprehension is also credulous; they take noisy messages at face value, and fail to account for the possibility they might have misunderstood. The sender's disclosure problem with a Bayesian receiver is harder than their problem with a credulous receiver, which will be a simple non-strategic choice-under-uncertainty problem. Comparing this problem to the problem of finding equilibrium disclosure rules with a rational receiver, we can see that disclosure with a credulous receiver will be riskier, since the receiver will fail to account for noise in messages and so their actions will be more volatile. However, the receiver's credulity could still benefit the sender in very low probability states, since the receiver will fail to consider that a given message is more likely to result

from a lot of noise in a high probability state than from only a small amount of noise in a very low probability state.

We have assumed that after observing the state the sender faces a binary choice between disclosing θ via a noisy channel or not disclosing (in the terminology of Hagenbach and Koessler (2017), we model a sender with a *simple* language whose disclosures are imperfectly observed), rather than allowing the sender to disclose any true statement about θ. This restriction would be without loss of generality if disclosure were noiseless[6], but *is* with loss of generality when disclosure is via a noisy channel. We make this restriction to allow for a tractable treatment of noise.

Modelling noise appropriately when the sender can report multiple possible values of θ is significantly more challenging, because the sender can now attempt to convey information to the receiver via the *size* of their message. To illustrate the problem, consider the following strategy for the sender: report 1 possible value for $\theta \in [0, \frac{1}{n})$, 2 possible values for $\theta \in [\frac{1}{n}, \frac{2}{n})$, and so on, up to n possible values for $\theta \in [\frac{n-1}{n}, 1]$. If the number of possible values the sender reports is not perturbed by noise in messaging, then as $n \to \infty$ the sender can achieve arbitrarily precise messaging with the above strategy regardless of the noise process. Modelling a noise process which perturbs both the content of messages *and* the size of messages is a significant challenge[7]; to avoid such complications, we restrict the sender to either a single valued message or no message. In many verifiable disclosure settings this restriction is present; for example, a lobbyist may commission a fact-checked piece of research which they are unable to edit but can choose not to disseminate. We

[6]With noiseless disclosure a receiver who knows the sender has positive bias should be as sceptical as possible after observing any imprecise claim by the sender. For example, if a University trying to impress applicants claims to be ranked in the top 11 in the country, a receiver should infer that the university is ranked 11th, since a university which was ranked higher would do better claiming to be ranked in the top 10. Hence claiming to be ranked in the top 11th for the university is equivalent to disclosing that their rank is *exactly* 11th.

[7]Blume, Board, and Kawamura (2007) do model a noise process which perturbs both content and size of messages in a cheap talk setting, by assuming that messages are either transmitted accurately with positive probability, or are draws from a noise distribution that is independent of the state. In contrast our treatment of noise allows for much greater dependence of the message distribution on the state, at the cost of restricting the sender to single valued messages.

provide a longer discussion of how one might use different treatments of noise to accommodate set-valued messages in section 5.

3 Equilibria

In this section, we discuss the possible outcomes in equilibrium in a general setting. We show if noise terms are unbounded equilibria featuring no disclosure always exist, and if bias is sufficiently large, full disclosure equilibria will also exist. We characterise when incentives to disclose are monotone in the state for a fixed conjecture of the receiver, and give conditions for there to exist a unique equilibrium in the class of lower censorship equilibria (disclose states *above* a threshold) and in the class of upper censorship equilibria (disclose states *above* a threshold). We find that for small biases, both lower and upper censorship equilibria may be possible for a given parameterisation. We characterise equilibrium disclosure rules in the limits as noise disappears or explodes, and show that while (with positive bias) lower censorship equilibria converge to full disclosure as noise vanishes, upper censorship equilibria do not. As noise explodes we show that equilibria in our setting approximate equilibria in a cheap talk model.

If S discloses, R's best response is to set $x(s) = E(\theta|s, \hat{D})$, where $\hat{D}$ is the receiver's conjecture of the sender's disclosure set, while if S does not disclose, R's best response is to set $x_0 = E(\theta|\theta \in \hat{D}')$.

If S does not disclose, their payoff is given by:

$$U_S(x, \theta, b) = -(x_0 - \theta)^2 - 2b(\theta - x_0) - b^2$$

while if they disclose, their expected payoff is given by:

$$U_S(x, \theta, b) = -E[(x(s) - \theta)^2 - 2b(\theta - x(s))|\theta] - b^2$$

Thus, the condition for S to strictly prefer disclosure can be written as:

$$E[(x_0 + x(s) - 2(\theta + b))(x_0 - x(s))|\theta] > 0$$

Equivalently:

$$2\left(\frac{x_0 + E(x(s)|\theta))}{2} - (\theta + b)\right)(x_0 - E(x(s)|\theta)) > Var[x(s)|\theta] \tag{3}$$

Note if $Var[E(\theta|s)|\theta] = 0$ our condition for disclosure reduces to asking whether the receiver's average action following non-disclosure lies closer to the sender's ideal action than their average action following disclosure, as in standard noiseless disclosure games or cheap talk[8]. For $Var[E(\theta|s)|\theta] > 0$ we can interpret the condition as saying that the average receiver action following disclosure, given θ, must not only be closer to the sender's ideal action than the receiver's non-disclosure action, but must be *sufficiently* closer than the receiver's non-disclosure action, to compensate for the risk a sender faces in disclosing.

Equation (3) tells us that if $Var[x(s)|\theta]$ does not vary too much with θ, the sender's equilibrium disclosure rule will involve not disclosing for an interval of states sufficiently close to x_0, and disclosing otherwise. In general, whether we can support a specific disclosure rule for the sender in equilibrium will depend upon the joint distribution of θ and s. However, in extreme cases we can characterise possible equilibria with minimal structure on θ and s.

3.1 Full Disclosure and No Disclosure

If the sender's bias is sufficiently large, we can construct equilibria in which *any* receiver action following disclosure is preferred by the sender to the receiver's non-disclosure action, regardless of the joint distribution of θ and s.

[8]Note that $Var[E(\theta|s)|\theta] = 0$ for $\theta = s$ and $Var[E(\theta|s)|\theta] = 0$ for θ and s independent; with no noise our model is standard verifiable disclosure, while for noise arbitrarily large we approximate two message cheap talk.

Proposition 1. *For there to exist an equilibrium in which the sender discloses for all* $\theta \in [0,1]$ *it is sufficient that* $|b| \geq \frac{1}{2}$*, and necessary that* $b \neq 0$*.*

Noisy disclosure is *risky* for the sender, generating a distribution over actions $x(s)|\theta$, compared to the certain action x_0 (and therefore certain payoff) induced by non-disclosure. As such, if the sender has small bias, in some states they will prefer non-disclosure to disclosure *even though* they prefer the average action induced by disclosure, because non-disclosure is less risky. If the sender has large enough positive bias, however, then *any* action they can induce by disclosing will be preferred to inducing x_0 by not disclosing, and so they will strictly prefer disclosure regardless of the risk (this is analagous to saying that a risk-averse utility maximiser prefers lottery A to lottery B if lottery A first-order stochastically dominates lottery B).

$|b| > \frac{1}{2}$ is sufficient but not necessary for full disclosure equilibria to exist; for particular distributions of θ and s we may also have full disclosure for smaller bias terms. We can think of $\frac{1}{2}$ as the *lowest* value of $|b|$ for which full disclosure equilibria are guaranteed to exist for *any* joint distribution of θ and s[9].

In full-disclosure equilibria, the sender's decision to disclose communicates nothing to the receiver in equilibrium, because if the sender could signal to the receiver via their disclosure decision, a sender with large positive bias would want to take the disclosure decision which signals the state is as high as possible in all states. As such, we may be able to construct no-disclosure equilibria by a similar reasoning to full-disclosure equilibria, with (with positive bias) all sender types attempting to signal a large state by not disclosing, and the receiver interpreting disclosure as a signal of a low state:

Proposition 2. *For non-disclosure for all* $\theta \in [0,1]$ *to be an equilibrium strategy for the sender, it is sufficient that:*

[9]Let $\{\theta_i\}_{i=1,2,...}$ be a sequence of random variables, such that $F(\theta_{i+1})$ first order stochastically dominates $F(\theta_i)$, and $\lim_{i\to\infty} F(\theta_i) = 0$ for all $\theta_i < 1$. Then for any $b < \frac{1}{2}$ and $s = \theta + \epsilon$ for unbounded ϵ, there exists some $k \in \mathbb{N}$ such that for $i > k$, a full disclosure equilibrium does not exist.

- *The conditional state distribution $\theta|s$ has full support for all signal realisations s,* ***or****:*

- *For all θ and s we have $\theta|s \in [s-e, s+e]$, $E(s|\theta) = \theta$, and:*

  - $\diamond$ *$b \geq \mu$, $1-\mu < e$, and if $e > \frac{1}{2}$, $s|\theta$ is log-concave,* ***or****:*

  - $\diamond$ *$b \leq -(1-\mu)$, $\mu < e$, and if $e > \frac{1}{2}$, $s|\theta$ is log-concave.*

To sketch the intuition; there are two ways we could try and support non-disclosure for all states in equilibrium. The first (trivial) way, if $\theta|s$ has full support, is to set $x(s) = \mu$ for all s; in other words, have the receiver conjecture that the only type who would ever disclose[10] is the mean type $\theta = \mu$. Given this conjecture, both disclosure and non-disclosure induce action μ with certainty, so no sender type has a strict incentive to disclose.

The second way would be to have the receiver conjecture that any disclosure message they receive must result from the lowest (highest) type possible when the sender has positive (negative) bias. If $\theta|s$ has full support, then given this conjecture all sender types will prefer non-disclosure if $|b| > \frac{1}{4}$.

We can see that if $\theta|s$ does *not* have full support, we will be unable to use the first (trivial) construction for non-disclosure equilibria. However, we may still be able to use the second construction if the bounds on the support of $\theta|s$ are not *too* tight relative to sender bias. For example, suppose that $x_l(s)$ is the most sceptical updating rule a receiver can use. If using this rule gives an average disclosure action in state $\theta = 1$ of $E(x_l(s)|\theta = 1) < \mu$, then $b \geq \mu$ will guarantee the sender will not disclose in any state. Our second condition in the proposition gives *sufficient* (but not necessary) conditions for the most sceptical updating rule a receiver can use to

[10]This might seem like an implausible equilibrium construction that it would be desirable to refine away. However, an equilibrium of this form survives several commonly used refinements in the signalling literature; in this equilibrium all sender types (except $\theta = \mu - b$) could gain from disclosure for *some* possible best responses by the receiver, and the set of possible receiver strategies for which different sender types would profit from deviation are generally not ranked by set-inclusion, so refinements in the tradition of the Intuitive Criterion or Divinity will have little bite.

induce a biased sender to prefer non-disclosure.

In contrast, if the bounds on the support of $\theta|s$ are too tight, then some types can gain from deviating to disclosure for any valid off-path beliefs for the receiver. For example, if $s|\theta \in [\theta - e, \theta + e]$, then $e \leq \frac{1-\mu}{2}$ and $b > -\frac{1-\mu}{2}$ will guarantee that if the receiver conjectures the sender will not disclose and interprets all signal realisations as coming from the lowest type possible, sender type $\theta = 1$ will strictly prefer to disclose.

Our results show that (for $\theta|s$ not bound too tightly) if bias is sufficiently large, we can support both full *and* no-disclosure in equilibrium. We can interpret this insight as follows: when bias is sufficiently large, the *signalling* effect of the disclosure *decision* on receiver actions dominates the effect of message *contents* in determining the sender's incentives to disclose. A similar result obtains in the literature on counter-signalling in the presence of private information for the receiver. For example, Harbaugh and To (2020) find that if the sender can costlessly choose a verifiable message from a finite set, and the receiver has access to a noisy private signal about the sender's type, the weakest senders who can send a particular verifiable message will have the most incentive to do so because they should expect the noisy signal to be least favourable. In equilibrium, this can lead to the receiver interpreting non-disclosure as a positive signal of sender type, meaning there are parameterisations for which both non-disclosure and full disclosure can be supported in equilibrium because the signalling value of the messaging decision dominates.

3.2 Monotone Incentives

Moving beyond extreme cases, we can see from (3) that in general sender incentives to disclose will depend upon $E(x(s)|\theta)$ and $Var[x(s)|\theta]$. As noted above, if $Var[x(s)|\theta]$ does not vary too much with θ, the sender will not disclose for states sufficiently close to x_0, and will disclose otherwise. In this section, we make this observation precise by deriving sufficient conditions on θ and s for incentives to disclose

to be monotonic in distance from x_0. A reader happy that $Var[x(s)|\theta]$ not varying too much with θ is a reasonable assumption can safely proceed to later sections, in which we characterise the form of partial disclosure equilibria.

To see why a type θ' preferring disclosure or non-disclosure need not imply whether $\theta'' > \theta'$ prefers disclosure or non-disclosure, consider the following example: suppose the receiver believes the sender follows a strategy of the form disclose if $\theta \geq \underline{d}$, don't disclose if $\theta < \underline{d}$, for $\underline{d} \in (0, 1)$. Let $\theta' \geq \underline{d}$ and suppose $s = \theta + \epsilon$ and $\epsilon \in [-e, e]$. Now note that in equilibrium the incentives of type $\theta'' > \theta'$ to disclose depend on the distribution of θ on the interval $(\theta' + e, \theta'' + e]$ via the receiver's updating rule, while the incentives of type θ' do not, since θ'' shares message realisations with these types, while θ' does not. If types in this interval are very high probability, it could be that even if type θ' prefers to disclose, type θ'' does not, since the average action θ'' induces from disclosure is a lot higher than the average action induced by type θ' disclosing, leading them to prefer x_0 with certainty to the action distribution from disclosure.

Understanding when, for a fixed conjecture $\hat{D}$ of the receiver, incentives to disclose are monotone will be important for characterising equilibria. We will use the following definition.

Definition 1. *For a fixed $x(s)$ and x_0, define the sender's **incentives to disclose** as the difference between their expected disclosure and non-disclosure payoff:*

$$E[U_S(x(s), \theta, b)|\theta] - U_S(x_0, \theta, b) = E[(x_0 + x(s) - 2(\theta + b))(x_0 - x(s))|\theta]$$

*Incentive to disclose are **monotone** if this difference is monotone in θ.*

Note that for equilibrium, we will only need to know how incentives to disclose for a type compares to incentives to disclose for an *indifferent* type. This means, for example, that for $\hat{D} = [\underline{d}, 1]$ to be an equilibrium disclosure set we do not need to require that incentives to disclose are increasing in θ everywhere. In particular, we

can note that if $\hat{D} = [\underline{d}, 1]$ *and* the support of $\theta|s$ does not depend upon s, no type $\theta \leq x_0 - b$ will ever disclose in equilibrium and no type $\theta \geq 1 - b$ will ever not-disclose in equilibrium, regardless of the distributions of θ and s. As such, we need to place *no* conditions on incentives to disclose in these regions to verify whether $\hat{D}$ can be an equilibrium disclosure set.

The following proposition gives sufficient conditions for incentives to disclose to be monotone in state θ for states sufficiently far from the non-disclosure action:

Proposition 3. *Define the difference between signal s and state θ as $s - \theta = \epsilon$. Suppose the following conditions hold:*

- *For a conjectured disclosure set $\hat{D}$, the receiver's beliefs satisfy $\frac{\partial E(\theta|s,\hat{D})}{\partial s} \in [0, 1]$.*
- *For all θ, changes in the density of $\epsilon|\theta$ as θ changes $|\frac{\partial f_{\epsilon|\theta}(\epsilon)}{\partial \theta}|$ are sufficiently small.*

Then incentives to disclose are increasing in θ for $\hat{D} = [\underline{d}, 1]$ if $\theta + b > x_0$ and ***either*** *$\theta \in \hat{D}$* ***or*** *$s|\theta$ has full support.*

A symmetric result holds for $\hat{D} = [0, \bar{d}]$.

To sketch the intuition behind this result, consider a setting in which the receiver expects the sender to disclose for $\hat{D} = [\underline{d}, 1]$. Assuming the sender's ideal action $\theta + b$ is above the action they induce by non-disclosure, let's ask when senders in higher states find disclosure relatively more appealing.

The sender observing a higher state has two implications for their disclosure incentives; firstly, their ideal action moves further away from the action induced by non-disclosure, so the payoff they would receive from non-disclosure falls. Secondly, because higher signals are more likely in higher states, the distribution of actions they induce by disclosing is shifted upwards.

For disclosure to become relatively more appealing as the state rises, we want to check that a rise in the state does not shift *too much* probability towards actions

following disclosure that the sender disprefers to the non-disclosure action. In other words, if the sender likes small overestimates (or small underestimates), increasing the state will make disclosure more appealing provided it does not move *too much* probability towards very large overestimates of the state, such that the sender would now find an underestimate of x_0 more appealing.

Relating this to our proposition, we can see that requiring $|\frac{\partial f_{\epsilon|\theta}(\epsilon)}{\partial \theta}|$ be sufficiently small will imply that the *distribution* of signals induced by disclosure does not increase too quickly as the state rises, while requiring $x'(s) \in [0,1]$ will imply that the *actions* induced by each signal do not increase too quickly as signals increase. We emphasise that our proposition gives only *sufficient* conditions for incentives to disclose to be monotone (and that monotone incentives themselves are sufficient but not necessary for equilibria).

It may not be clear how often the sufficient conditions in our proposition will be satisfied. On this, we can reason that a uniform state plus an independent single peaked noise term should satisfy $x'(s) \in [0,1]$, provided the *slope* of the density of the noise term is not too large. We can further reason that if we strictly satisfy this with a uniform state, we should be able to satisfy it with a state sufficiently close to uniform, provided the density does not vary too much or too fast with the state, such that the receiver's posterior mean is increasing in s but not *too* responsive.

More generally, we can ask whether the sufficient conditions in our proposition appear reasonable restrictions on a model of noisy disclosure. Note that since the proposition *defines* $\epsilon = s - \theta$, there is no intrinsic substantive assumption in writing the signal in terms of additive noise. Assuming that for a conjectured $\hat{D}$ we have $x'(s) \in [0,1]$ is a joint restriction on ϵ, θ; it requires that the posterior mean cannot increase too quickly with the signal. We can think of this restriction as saying the informativeness of a signal cannot vary too quickly with the realisation; the posterior mean induced by signal s must not be too much smaller than the posterior mean induced by signal $s + \delta$. In the context of modelling misunderstandings in

communication, it does not appear too restrictive to consider noise processes with this property[11] on the equilibrium path; for example, it seems plausible a reader with imperfect comprehension reading two very similar articles will update their beliefs in a similar way for both.

If the support of $s|\theta$ depends upon θ, then assuming that for conjectured $\hat{D}$ we have $x'(s) \in [0,1]$ also restricts beliefs *off* the equilibrium path. We note, however, that it is easy to find plausible noise specifications compatible with $x'(s) \in [0,1]$ for off-path s:

Remark 1. *If for all θ we have $s-\theta \in [\underline{\epsilon}, \bar{\epsilon}]$, then for $\hat{D} = [\underline{d}, \bar{d}]$ an off-path updating rule which sets $E(\theta|s) = s + \underline{\epsilon}$ for off-path $s < \underline{d} - \underline{\epsilon}$ and $E(\theta|s) = s - \bar{\epsilon}$ for off-path $s > \bar{d} + \bar{\epsilon}$ satisfies $x'(s) \in [0,1]$ for off-path s.*

In other words, a sufficient condition to find an updating rule such that $x'(s) \in [0,1]$ for off-path s is that the *support* of additive noise does not depend upon θ. If the receiver then believes the state to be the highest possible state compatible with a signal for unexpectedly low signals (and the lowest possible compatible state for unexpectedly high signals), for off-path signals the receiver action will respond exactly one-for-one with s, which is compatible with the sufficient conditions in our proposition[12].

Assuming that $|\frac{\partial f_{\epsilon|\theta}(\epsilon)}{\partial \theta}|$ is sufficiently small implies that if we think of the signal as comprising the state θ plus some additive noise ϵ, that noise cannot be *too* correlated with the state. Note that if ϵ and θ are independent then $\frac{\partial f_{\epsilon|\theta}(\epsilon)}{\partial \theta} = 0$, so we can interpret this condition as a *weakening* of an independence assumption; the probability of given noise draws *can* vary as θ varies, provided it does so sufficiently smoothly. Requiring that possible misunderstandings in state θ are not too

[11] This assumption *would* be violated by a signal which (noisily) reported what side of a threshold a state was; for example, if the signal (noisily) reported a pass-fail test on the sender's underlying type, senders who are just below the threshold for passing and do not want a large overestimate may find disclosure less appealing than senders in very low states who do not expect to pass.

[12] Note, however, that our proposition *only* characterises incentives to disclose when, for $\hat{D} = [\underline{d}, 1]$, $\theta + b > x_0$. In a setting with bounded support for $\theta|s$, if we have $b < x_0$ we *also* need to check incentives to disclose for types $\theta + b < x_0$ to characterise an equilibrium.

dissimilar to possible misunderstandings in state $\theta + \delta$ for δ small does not appear a restrictive assumption in general settings of noisy disclosure.

We have given sufficient (but not necessary) conditions for incentives to be monotone in disclosure, and argued these conditions are plausible assumptions in general settings of noisy disclosure. For the remainder of the paper we will make the following assumption on monotonicity of incentives and focus on characterising the sender's disclosure thresholds in partial disclosure equilibrium:

Assumption 1. *The distributions of θ and s are such that:*

- *Incentives to disclose are increasing in θ for $\hat{D} = [\underline{d}, 1]$ and $\theta > x_0 - b$.*
- *Incentives to disclose are decreasing in θ for $\hat{D} = [0, \bar{d}]$ and $\theta < x_0 - b$*

for some consistent updating rule for the receiver.

This assumption is sufficient for us to evaluate incentives to disclose in candidate equilibria featuring upper or lower censorship. We know, for example, that if $\hat{D} = [\underline{d}, 1]$ and receiver posteriors have full support for all s then all $\theta \leq x_0 - b$ always prefer non-disclosure, and for $\theta > x_0 - b$ as θ increases disclosure becomes more appealing. Thus to find a lower censorship equilibrium it is sufficient to find a cutoff $\underline{d}$ such that when $\theta = \underline{d}$ the sender is indifferent between disclosure and non-disclosure.

3.3 Partial Disclosure

While we can provide some sufficient conditions for full and no-disclosure equilibria in a general setting, characterising partial disclosure equilibria in a non-parameterised model is more difficult. From equation (3) we can see that if the distributions of θ and s are appropriately well-behaved (in the sense that $Var[x(s)|\theta]$ does not vary *too* quickly with θ) we should have non-disclosure for an interval of states[13] and disclosure for all other states. This gives three possibilities for partial disclosure equilibria: *lower censorship* equilibria featuring disclosure for a set

[13]It is easy to construct pathological counterexamples; suppose $D = [\underline{d}, 1]$ is an equilibrium disclosure set for $b = 0$, $\theta \sim U[0, 1]$ and some signal s, and write $\mu_0 = E(\theta|\theta < \underline{d})$. Now consider

$D = [\underline{d}, 1]$, *upper censorship equilibria*, featuring disclosure for a set $D = [0, \bar{d}]$, and *interior censorship* equilibria, featuring disclosure for a set $D = [d_1, d_2]$, for $\underline{d}, \bar{d}, d_1, d_2 \in (0, 1)$. Note that characterising lower censorship equilibria is sufficient to charactersise upper censorship equilibria, since if $D = [\underline{d}, 1]$ is an equilibrium disclosure set for state θ, signal s, and bias b, $\tilde{D} = [0, 1 - \underline{d}]$ is an equilibrium disclosure set for bias $-b$ and state and signal $\tilde{\theta}$ and $\tilde{s}$, where the joint distribution of $\tilde{\theta}$ and $\tilde{s}$ is the mirror image of the joint distribution of θ and s.

In general, we would expect equilibria featuring interior censorship to be hard to support, except in perfectly symmetric settings[14], since compared to a cutoff rule we add an additional condition to the equilibrium fixed point problem (as for interior censorship two sender types must be indifferent between disclosure and non-disclosure compared to only one type with upper or lower censorship). As such, we will focus our attention on lower (and upper) censorship partial disclosure equilibria; while interior censorship equilibria may exist in asymmetric settings, they are less likely to be robust to perturbations, or to be compelling predictions.

Earlier we argued that full-disclosure equilibria exist for sufficiently large biases. We can link this to a claim about partial disclosure equilibria:

Proposition 4. *If the distributions of θ and s have continuous densities, the support of $\theta|s$ does not depend upon s, and Assumption 1 holds, there exists an equilibrium in which the sender discloses if $\underline{d}(b) \leq \theta$, where there exist $\underline{b} < 0 < \bar{b}$ such that $\underline{d}(b) = 0$ for $b > \bar{b}$ and $\underline{d}(b) \in (0, 1)$ for $\underline{b} < b < \bar{b}$].*

This proposition tells us we can think of full disclosure equilibria as a limiting case of lower-censorship equilibria when bias is sufficiently large. Above some threshold for bias, regardless of the noise in messages, if the receiver conjectures that the sender

a new signal s', where $s' = \theta$ if $\theta \in [\mu_0 - \delta, \mu_0 + \delta]$ and $s = s'$ otherwise. For δ sufficiently small, $D \approx [\mu_0 - \delta, \mu_0 + \delta] \cup [\underline{d}, 1]$ will be an equilibrium disclosure set with signal s', with non-disclosure for the union of two disjoint intervals.

[14] If $b = 0$ *and* θ and s are symmetrically distributed we should be able to support an interior censorship equilibrium since symmetry implies that the interval of non-disclosure will be symmetric about $\frac{1}{2}$, but this is likely to be a knife-edge case.

discloses above some cutoff, the only cutoff that we can support in equilibrium is $\underline{d} = 0$. As we reduce the sender's bias, however, we will eventually reach a value of bias such that a sender observing $\theta = 0$ is indifferent between disclosure and non-disclosure, due to the risk of large overestimate from disclosure. Reducing bias further will lead to senders in a neighbourhood of $\theta = 0$ *strictly* preferring non-disclosure, giving us a lower censorship equilibria. On an intuitive level, above this threshold for bias the sender does not view noise in messages as a cost to disclosure, so acts as they would in a costless disclosure game, while below this threshold the sender does view the possibility of large misunderstandings as a cost, and so is deterred from disclosing in low states (where the 'cost' of disclosure relative to non-disclosure is highest).

When the support of $\theta|s$ does not depend upon s, we do not need to check whether as a lower censorship equilibrium threshold grows, beyond some point the lowest types not disclosing prefer to switch back to disclosure. In contrast, if we did *not* assume that the support of $\theta|s$ does not depend upon s, to guarantee a lower censorship equilibrium exists we would also need to verify that we can set off-path beliefs such that a sender observing $\theta < x_0 - b$ does not prefer to disclose.

Note our proposition tells us that b need not be positive in lower censorship equilibrium. It is possible that even if a sender has negative bias, if a receiver anticipates non-disclosure only for small states, the sender may prefer to act consistently with this; in particular, if noise is unbounded, $b < 0$ and the receiver anticipates all types $\theta < \underline{d}$ will not disclose, no type $\theta < \underline{d}$ can induce a lower action than $\underline{d}$ by disclosing.

Monotone incentives with continuous densities tells us we can find an equilibrium cut-off. In general, however, this is *not* sufficient to tell us this cut-off is unique. Our next result gives sufficient conditions for uniqueness of an equilibrium lower-censorship disclosure rule.

Proposition 5. *Fix the distributions of s, θ, and the bias b. There exists at most* ***one*** $\underline{d} \in [0,1]$ *such that $D = \hat{D} = [\underline{d}, 1]$ is an equilibrium disclosure set if the*

following conditions holds:

- *Assumption 1 is true.*
- $\frac{\partial E(\theta|\theta \notin \hat{D})}{\partial \underline{d}} < 1$
- $\frac{\partial E(\theta|s,\theta \in \hat{D})}{\partial \underline{d}} < 1$ *for* $E(\theta|s, \theta \in \hat{D}) > \underline{d} + b$
- $\frac{\partial E(\theta|\theta \notin \hat{D})}{\partial \underline{d}} < \frac{\partial E(\theta|s,\theta \in \hat{D})}{\partial \underline{d}}$ *for* $E(\theta|s, \theta \in \hat{D}) \leq \underline{d} + b$.

This result tells us that if incentives to disclose are monotone for a fixed receiver conjecture, *and* the receiver's action at any given information set does not react too much to a change in conjectured cutoff, there will be a unique equilibrium cutoff above which the sender will disclose for a given parameterisation.

To sketch the proof, we ask when increasing the receiver's conjecture of a threshold $\underline{d}$ will increase the sender's incentives to disclose when they observe $\theta = \underline{d}$. We can decompose the effect of increasing conjectured threshold $\underline{d}$ on incentives to disclose into two parts; the effect of a change in $\underline{d}$ on the distribution of $x(s)|\theta = \underline{d}$ holding receiver strategy fixed, and the effect of a change in $\underline{d}$ on receiver strategy.

This first effect is positive whenever incentives to disclose are increasing in the state (for a fixed receiver conjecture), so if Assumption 1 holds this effect will be positive (for $\theta + b > x_0$).

The second effect will depend upon both the change in x_0 as a result of a change in $\underline{d}$ and the change in $x(s)$ as result of a change in $\underline{d}$. As conjectured cutoff type rises, their ideal action $\underline{d} + b$ also rises, as does the no-disclosure action x_0 and the action induced by disclosure for any given signal $x(s)$. What our proposition says, roughly, is that if the cutoff type's ideal action rises by more than the actions they can induce by non-disclosure or disclosure rise, then this second effect also makes disclosure more appealing at the cutoff rises.

How often will we satisfy these sufficient conditions for unique equilibria (again,

we emphasise that we derive sufficient[15] rather than necessary conditions)? We can note the following:

Remark 2. *If f_θ and $f_{\theta|s}$ are log-concave for all s, $\frac{\partial x_{\underline{d}}(s)}{\partial \underline{d}}, \frac{\partial x_{\underline{d}}}{\partial \underline{d}} < 1$*

Log-concavity[16] is a property of many commonly studied unimodal distributions used in economics, and is a relatively standard assumption on a state θ in the signalling literature. The assumption that $f_{\theta|s}$ is log-concave for all s is slightly stronger, but is easy to satisfy; for example, it is sufficient that $s = \theta + \epsilon$ for θ and ϵ log-concave and independent.

Note that log-concavity alone is *not* sufficient for uniqueness[17] of equilibrium in our setting, because we also need $\frac{\partial x_{\underline{d}}(s)}{\partial \underline{d}} > \frac{\partial x_{\underline{d}}}{\partial \underline{d}}$ for $\underline{d} + b \geq x_{\underline{d}}(s)$. Requiring $\frac{\partial x_{\underline{d}}(s)}{\partial \underline{d}} > \frac{\partial x_{\underline{d}}}{\partial \underline{d}}$ for $\underline{d} + b \geq x_{\underline{d}}(s)$ tells us that for signal realisations low enough to induce an action below $\underline{d} + b$, increasing the conjectured threshold $\underline{d}$ must increase beliefs following those signal realisations more than it increases beliefs following no message. On an intuitive level, this says that type $\underline{d}$ must be relatively more likely for low signal realisations than it would be under non-disclosure, so removing that type from the disclosure set leads to a larger change in beliefs for low signal realisations than the change in beliefs from adding it to the non-disclosure set[18].

[15]Note for example that a necessary property of equilibrium is that the average receiver action following disclosure must be closer to $\underline{d}+b$ than $x_{\underline{d}}$; as such, if a cut-off does not satisfy this property, we can see it doesn't matter for uniqueness of equilibrium whether disclosure does become less appealing as $\underline{d}$ rises. For example, we know no $\underline{d} + b \geq 1$ can be an equilibrium cut-off, regardless of the effect of a marginal change in $\underline{d}$ on $\underline{d}$'s incentives.

[16]Bagnoli and Bergstrom (2005) derive general results on log-concave distributions. Szalay (2013) shows that assuming log-concave distributions can give uniqueness of equilibria (for a given number of partitions) in a cheap-talk setting, by similar reasoning to our uniqueness result.

[17]To obtain uniqueness in a cheap talk setting, Szalay (ibid.) does not need an additional requirement equivalent to $\frac{\partial x_{\underline{d}}(s)}{\partial \underline{d}} > \frac{\partial x_{\underline{d}}}{\partial \underline{d}}$ on top of log-concavity because in cheap talk the cut-off type indifferent between two messages necessarily has the high message induce an action *above* their ideal action. In contrast with noisy disclosure, there may be some signal realisations following disclosure that induce an action *below* the sender's ideal action, even when the sender is indifferent.

[18]For example, suppose the state is uniform. Then $\frac{\partial x_{\underline{d}}}{\partial \underline{d}} = \frac{1}{2}$, and we can argue that if $s = \theta + \epsilon$ for symmetric, independent and unimodal ϵ this will be the case. Note that for $s = \frac{\underline{d}+1}{2}$ we must have $x(s) = s$ by symmetry, and *on average* an increase in $\underline{d}$ must lead to an *average* rise in $x(s)$ of $\frac{1}{2}$. As for $s = \frac{\underline{d}+1}{2}$ a rise in $\underline{d}$ must shift $x(s)$ up by *exactly* $\frac{1}{2}$ proportionally, and for $s < \frac{\underline{d}+1}{2}$ $\underline{d}$ is more likely while for $s > \frac{\underline{d}+1}{2}$ $\underline{d}$ is less likely, we should have $\frac{\partial x_{\underline{d}}(s)}{\partial \underline{d}} > \frac{1}{2}$ for $s < \frac{\underline{d}+1}{2}$ and $\frac{\partial x_{\underline{d}}(s)}{\partial \underline{d}} < \frac{1}{2}$

We might also wonder how appealing our sufficient conditions are as *descriptions* of a noisy disclosure setting. The requirement that changing the conjectured cutoff $\bar{d}$ does not induce too large a change in a receiver's updating rule will be satisfied provided the density of the state does not vary too much in the neighbourhood of the cutoff and noise in messages $s - \theta$ is not too correlated with θ. In other words, our conditions are plausible in noisy disclosure settings where changing the receiver's conjecture does not induce too large a change in sender incentives to disclose; for example, when priors are sufficiently smooth.

When changing a conjectured cutoff *can* induce a large change in updating rule, lower censorship equilibria may not be unique, because there exists a co-ordination problem between sender and receiver in disclosure strategy. For example, if θ has (a continuous approximation of) a discrete distribution, we would not in general expect unique equilibria[19].

3.4 Limiting Cases

When disclosure is noiseless ($\theta = s$), our model of noisy disclosure corresponds exactly to a classic verifiable disclosure model. In contrast, when disclosure becomes so noisy that the noisy signal s is uncorrelated with the state ($F(s|\theta) = F(s)$), our model of noisy disclosure corresponds exactly to classic cheap talk with two possible messages (disclose and not-disclose).

Given these two benchmarks, a natural question to ask in our setting is: do equilibrium strategies converge to noiseless verifiable disclosure strategies as noise vanishes, and to cheap talk as noise explodes? In this section we show that the answer to the latter question is yes, but that the answer to the former question

for $s > \frac{d+1}{2}$.

[19]If the state's density has several peaks, then the receiver's updating rules will vary significantly depending upon whether those peak types lie inside or outside the disclosure region. For example, a small increase in disclosure threshold which resulted in an additional peak falling in the non-disclosure region could induce a large increase in non-disclosure action, making non-disclosure *more* appealing for the cutoff type as the threshold increased locally.

may be no; even with vanishingly small noise in messaging, equilibrium disclosure strategies may be bounded away from disclosure strategies in a noiseless disclosure game.

To allow for easy comparability with Crawford and Sobel (1982)'s leading example of cheap talk, we will assume for this section that $\theta \sim U[0,1]$. This gives the following two benchmarks (adopting the convention that the sender discloses when indifferent):

Observation 1. *If $\theta = s$, for $b > 0$ ($b < 0$), the unique equilibrium disclosure set is $D = [0,1]$, with unique receiver equilibrium strategy $x(s) = \theta$ and $x_0 = 0$ ($x_0 = 1$).*

Observation 2. *If $F(s|\theta) = F(s)$, the following equilibrium disclosure sets can be supported:*

- $D = [0,1]$
- $D = \emptyset$
- *For $|b| > \frac{1}{4}$, $D = [\frac{1}{2} - 2b, 1]$ (with $x(s) = \frac{3}{4} - b$ and $x_0 = \frac{1}{4} - b$)*
- *For $|b| > \frac{1}{4}$ $D = [0, \frac{1}{2} - 2b]$ (with $x(s) = \frac{1}{4} - b$ and $x_0 = \frac{3}{4} - b$)*

To capture equilibrium disclosure rules varying as the *amount* of noise in messages varies, let $s = \theta + \sigma\varepsilon$ for $\varepsilon \in R$ log-concave and symmetric[20] and independent of θ. This ensures equilibria will be unique. An increase in σ represents a messaging technology becoming more noisy. We can then explore how equilibrium disclosure rules behave as $\sigma \to 0$ and as $\sigma \to \infty$.

As a notational convention, we will write $\lim_{\sigma \to a} D(b) \in \{A, B\}$ to mean that as $\sigma \to a$, there exists an equilibrium disclosure set that converges to A *and* an equilibrium disclosure set that converges to B.

[20] ε log-concave implies that we satisfy Assumption 1, so we know if $\frac{1}{2} < \frac{x(s)}{\underline{d}}$ for $x(s) \leq \underline{d} + b$ for each σ there is at most one $\underline{d}(b)$ that solves our indifference condition. ε symmetric tells us that (since ε is unimodal and log-concave and θ is uniform), $\frac{\partial x(s)}{\underline{d}} \geq \frac{1}{2}$ for $x(s) \leq \frac{1+\underline{d}}{2}$, so we also satisfy our uniqueness conditions.

Proposition 6. *For $\theta \sim U[0,1]$ and $s = \theta + \sigma\varepsilon$ for $\varepsilon \in \mathbb{R}$ log-concave, symmetric, and independent of θ, as $\sigma \to 0$, for $b > 0$ the following can be supported as limiting equilibrium disclosure sets* $\lim_{\sigma\to 0} D(b)$

- *If* $b \in (0, \frac{1}{4})$, $\lim_{\sigma\to 0} D(b) \in \{[0,1], [1-4b, 1], \emptyset\}$
- *If* $b > \frac{1}{4}$, $\lim_{\sigma\to 0} D(b) \in \{[0,1], \emptyset\}$.

A symmetric result holds for negative bias.

Our proposition tells us that as noise vanishes ($\sigma \to 0$), if the support of ε remains unbounded, *both* partial disclosure *and* non-disclosure remain equilibrium outcomes. In other words, it is not just that noise in messaging creates multiple equilibria; these equilibria *persist* even as noise vanishes[21], meaning sending $\sigma \to 0$ is not enough to recover the unravelling result that holds when $\sigma = 0$.

We can see that, for positive bias lower-censorship equilibria converge to the unique equilibrium in the noiseless verifiable disclosure problem. This is intuitive; these feature the same sceptical reasoning on the part of the receiver as in the noiseless game; if the sender does not disclose, this is interpreted as the sender having nothing good to reveal.

In contrast with positive bias upper-censorship equilibria *do not* converge to full disclosure as $\sigma \to 0$, because they do not feature the same sceptical 'unravelling' reasoning as with positive bias. To see this, note that upper-censorship equilibria with positive bias are equivalent to lower-censorship equilibria with negative bias; in these equilibria senders who see the lowest states in the noiseless game may have the greatest incentive to disclose, but can be blocked from doing so in the noisy game by a 'self fulfilling prophecy' on the part of the receiver; if the receiver believes senders who see the lowest states never disclose, they are unable to gain by doing so.

[21] This is similar to a finding in the literature on leader-follower games (for example, Bagwell (1995)), where all first-mover advantage to a leader may be lost if their move is imperfectly observed by the follower; even if the probability of being incorrectly observed vanishes, what matters is that follower beliefs have full support at each information set.

Turning to the case of arbitrarily large noise, we have:

Proposition 7. *For $\theta \sim U[0,1]$ and $s = \theta + \varepsilon$ for $\varepsilon \in \mathbb{R}$ log-concave, symmetric, and independent of θ, as $\sigma \to \infty$, for $b \geq 0$ equilibrium disclosure sets converge to:*

- *If $b \in [0, \frac{1}{4})$, $\lim_{\sigma\to\infty} D(b) \in \{[0, \frac{1}{2} - 2b], [\frac{1}{2} - 2b, 1], \emptyset\}$*
- *If $b \geq \frac{1}{4}$, $\lim_{\sigma\to\infty} D(b) \in \{[0,1], \emptyset\}$*

Again a symmetric result holds for negative bias.

We can see for arbitrarily large noise, equilibrium disclosure sets *do* converge to a (subset of) cheap talk messaging rules. As noise becomes large, all the receiver can effectively infer from their signal is that the sender chose to make a disclosure, in which case the content of any given message is meaningless, and the sender's problem becomes analogous to a cheap talk problem with only two possible messages.

Note that in cheap talk, with $F(s|\theta) = F(s)$ we have two ways to construct a *garbling* equilibrium (with no information transmitted); full-disclosure and no-disclosure. In contrast, with noisy disclosure and small bias, only no-disclosure can be supported in equilibrium as $\sigma \to \infty$. This is because even as $\sigma \to \infty$, disclosure still remains *more* risky for the sender than non-disclosure, and so a sender type observing $\theta = x_0 - b$ still strictly prefers non-disclosure (as do their neighbours). As such, we cannot (necessarily) approximate every *disclosure set* in a pure cheap talk model as $\sigma \to \infty$.

Nonetheless, note we *can* approximate any *payoff* achievable via cheap talk with two messages as $\sigma \to \infty$. As such, there is a meaningful sense in which we can say that arbitrarily noisy disclosure *is* cheap talk, while arbitrarily noiseless disclosure is *not* verifiable disclosure. Phrased another way, our limiting results suggest that cheap talk models *are* robust to introducing a tiny amount of 'hard' information into messages, while verifiable disclosure models are *not* robust to introducing a tiny amount of noise into messages.

A positive implication of these limiting results is that with noisy disclosure we can draw very limited inference from observing the *form* of a disclosure rule in noisy disclosure, since upper censorship, lower censorship, full-disclosure and no-disclosure are all compatible with both positive and negative bias for at least *some* parameterisation. Even if we can directly observe disclosure *sets*, we would also need to observe the distributions of θ and s to identify the direction of sender bias; for example, $D = [\frac{3}{4}, 1]$ could occur for very small σ and a sender with bias $b = \frac{1}{16}$, or for very large σ and a sender with bias $b = -\frac{1}{8}$. This inference problem is of *practical* importance; for example, in lobbying settings there may be good reasons for regulators and watchdogs to want to understand the agendas of lobbyists seeking to influence policy[22]; we show their ability to do so is severely limited by the presence of *any* amount of noise in messaging.

4 Efficiency

Introducing noise to messages in a disclosure setting affects both equilibrium disclosure rules and the receiver's ex-ante optimal disclosure rule (which in turn can affect receiver preference over sender bias *ex-ante*). In this section we analyse equilibria in a parameterised examples with a uniform state and additive uniform noise[23], and compare equilibrium disclosure rules to the ex-ante receiver (and sender) optimal disclosure rule.

Before analysing the parameterised models, we first state a simplifying result for the characterisation of ex-ante efficiency (henceforth efficiency), by showing that (following Crawford and Sobel (1982)), given the receiver is Bayesian and best re-

[22]Note that due to the self-fulfilling prophecy feature of equilibria, we do not need to require that *receivers* understand the sender's *bias* to play equilibrium, provided they understand the sender's *strategy* and the distributions of s and θ. For example, a new minister may be told in handover that expert A will send them a report if pollution is too high; if their updating rule reflects this it is not necessary for them to observe expert A's bias to play a lower censorship equilibrium.

[23]We focus on additive uniform noise as it gives an analytically tractable model in which we can solve for equilibrium; simulations with a uniform state and normally distributed noise produce equilibria which do not differ qualitatively and feature the same comparative statics.

sponding to the sender's strategy, maximising the receiver's ex-ante expected payoff and the sender's ex-ante expected payoff are equivalent:

Lemma 1. *The messaging rule which maximises the sender's ex-ante expected payoff also maximises the receiver's ex-ante expected payoff provided the receiver is best responding.*

Our focus is on whether efficient strategies can be supported in equilibrium *for a given noise distribution.* However, we can also make a comparison between noiseless and noisy disclosure games. The unravelling result tells us that with noiseless communication we would have $-E((x-\theta)^2) = 0$, which is the upper bound on the receiver's payoff for any information structure. As such, an immediate corollary of the above lemma is that when we compare a noisy disclosure setting to a noiseless disclosure setting, other things equal expected payoffs are higher in the noiseless setting. While Blume, Board, and Kawamura (2007) show that, if communication is via cheap talk, a biased sender and receiver might prefer (ex-ante) to communicate via a noisy channel, we can see with verifiable information both sender and receiver would like to commit to noiseless transmission[24]. This is unsurprising; in a noiseless disclosure game, we can support full revelation in equilibrium, whereas even if the receiver could choose the sender's strategy, they could not achieve full revelation in a noisy disclosure game.

Going forwards we will phrase our claims about efficiency from the perspective of the receiver's ex-ante expected payoff, and discuss the receiver's ex-ante preference over sender bias for a fixed signal distribution. In other words, we are interested in whether noise is not just *mechanically* but *strategically* harmful.

[24]Recall that our aim is to model settings in which noise is *unavoidable*, for example due to a language barrier (either literal or technical). While a native English speaker consulting a native French speaking expert might have a higher expected payoff from the interaction if they spoke French themselves, it is unlikely it is feasible for them to learn at short notice.

4.1 Uniform State, Uniform Noise

Let $\theta \sim U[0,1]$, and let $s = \theta + \epsilon$ for $\epsilon \sim U[-e, e]$, with $E(\epsilon\theta) = 0$. We will focus on disclosure sets of the form $D = [\underline{d}, 1]$.

We begin by computing posterior distributions for conjectured cutoff $\underline{d}$. We can see that we will need to treat the cases $\underline{d} + e < 1 - e$ and $\underline{d} + e > 1 - e$ separately. If $\underline{d} + e < 1 - e$, the highest message that could result from the lowest type disclosing is lower than the lowest message that could result from the highest type disclosing. In other words, messages drawn from the middle of the support of s help to narrow down the support of θ in R's posterior. In contrast if $\underline{d} + e > 1 - e$ messages in the middle of the support of s could have originated from any sender type in $[\underline{d}, 1]$ and so do not narrow down the support of R's posterior.

First, assume that $\underline{d} + e < 1 - e$. Computing the distribution of s given the disclosure rule, we find the following density:

$$f_s = \begin{cases} \frac{s+e-\underline{d}}{2e(1-\underline{d})}, & \text{for } s \in [\underline{d} - e, \underline{d} + e] \\ \frac{1}{1-\underline{d}}, & \text{for } s \in [\underline{d} + e, 1 - e] \\ \frac{1+e-s}{2e(1-\underline{d})}, & \text{for } s \in [1 - e, 1 + e] \end{cases}$$

This gives conditional density of $\theta|s$:

$$f_{\theta|s} = \begin{cases} \frac{1}{s+e-\underline{d}}, & \text{for } s \in [\underline{d} - e, \underline{d} + e], \theta \in [\underline{d}, s + e] \\ \frac{1}{2e}, & \text{for } s \in [\underline{d} + e, 1 - e], \theta \in [s - e, s + e] \\ \frac{1}{1+e-s}, & \text{for } s \in [1 - e, 1 + e], \theta \in [s - e, 1] \end{cases}$$

Now consider the case in which $\underline{d} + e > 1 - e$:

$$f_s = \begin{cases} \frac{s+e-\underline{d}}{2e(1-\underline{d})}, & \text{for } s \in [\underline{d} - e, 1 - e] \\ \frac{1}{2e}, & \text{for } s \in [1 - e, \underline{d} + e] \\ \frac{1+e-s}{2e(1-\underline{d})}, & \text{for } s \in [\underline{d} + e, 1 + e] \end{cases}$$

and:

$$f_{\theta|s} = \begin{cases} \frac{1}{s+e-\underline{d}}, & \text{for } s \in [\underline{d} - e, \underline{d} + e], \theta \in [\underline{d}, s + e] \\ \frac{1}{1-\underline{d}}, & \text{for } s \in [1 - e, \underline{d} - e], \theta \in [\underline{d}, 1] \\ \frac{1}{1+e-s}, & \text{for } s \in [1 - e, 1 + e], \theta \in [s - e, 1] \end{cases}$$

Note that this tells us that when $\underline{d}+e > 1-e$ the contents of a message $s \in [1-e, \underline{d}+e]$ are completely uninformative; it is equally likely to have come from any of the sender types who disclose, and so the receiver learns as much as they would if they did not see the contents of the message (but knew one had been sent).

With some algebra we can compute the efficient and equilibrium cutoff rules analytically. The following proposition gives the efficient disclosure rule in the class[25] of lower censorship equilibria:

Proposition 8. *If the sender's disclosure set is given by $D = [\underline{d}, 1]$, ex-ante expected payoffs are maximised when $D^* = [\underline{d}^*, 1]$, where:*

- *If $e \leq \frac{3}{6+2\sqrt{3}}$ then $\underline{d}^* = \frac{2\sqrt{3}}{3}e$*
- *If $e > \frac{3}{6+2\sqrt{3}}$ then $\underline{d}^*$ solves $3(2\underline{d}^* - 1) + \frac{(1-\underline{d}^*)^3}{e} = 0$*

We can also calculate the *unique* equilibrium disclosure rule when the receiver conjectures the sender only discloses states above some threshold:

[25] It is not true in general that lower censorship equilibria are always superior to equilibria featuring interior censorship; if noise in communication is very small, then following interior censorship there will be minimal proabability of mistaking disclosure in a high state for disclosure in a low state, so splitting the disclosure region in two may benefit the receiver compared to disclosing only on some interval. Nonetheless, given interior censorship will be much harder to support in equilibrium, optimal lower censorship arguably remains a relevant welfare benchmark.

Proposition 9. *Suppose the sender's disclosure set takes the form* $D(b) = [\underline{d}(b), 1]$ *or* $D(b) = \emptyset$*. For each possible value of* $-\frac{1}{4} < b$*, there is one value of* $D(b)$ *that can be supported in equilibrium, where:*

- *If* $b \geq \frac{e}{3}$ *and* $e \leq \frac{1}{2}$, $\underline{d}(b) = 0$
- *If* $b \geq \frac{3e-1}{12e-3}$ *and* $e > \frac{1}{2}$, $\underline{d}(b) = 0$
- *If* $b < \frac{e}{3}$ *and* $\underline{d}(b) \leq 1 - 2e$, $\underline{d}(b) = 2(\sqrt{b^2 + \frac{e^2}{3} - be} - b)$
- *If* $-\frac{1}{4} < b < \frac{3e-1}{12e-3}$ *and* $\underline{d}(b) > 1 - 2e$, $\underline{d}(b)$ *solves:*

$$1 + 3\underline{d}^2 + 6\underline{d}e + 3b(4e + 2\underline{d} - \underline{d}^2 - 1) = 3e + 3\underline{d} + \underline{d}^3$$

- *If* $b \leq -\frac{1}{4}$, $D(b) = \emptyset$ *or* $D(b) = [0, 1]$

A symmetric result will hold when the receiver expects the sender to only disclose for states sufficiently low: if $D(b) = [\underline{d}(b), 1]$ is an equilibrium disclosure set, given the symmetry of our setting $\tilde{D}(b) = [0, 1 - \underline{d}(-b)]$ is also an equilibrium disclosure set.

On comparative statics, we can see that if the receiver believes the sender only discloses for states above some threshold, the sender will disclose weakly more often when they are more biased, and weakly less often when messaging is noisier. If bias b is sufficiently large *relative* to the amount of noise in messages (captured by e), the sender will always disclose in equilibrium, because the probability the sender receives an action from disclosure they disprefer to the non-disclosure action of 0 is sufficiently low. As the amount of noise becomes large *relative* to bias, the sender will want to disclose in fewer states, because in low states disclosing risks a large overestimate due to noise, compared to a small and certain miss-estimate from non-disclosure. As $e \to \infty$, we can see that our thresholds $\underline{d}(b) \to \frac{1}{4} - 2b$ (matching our result in the previous section), which correspond to the two-message cheap talk partitions.

When $b \leq -\frac{1}{4}$ our result tells us that if the receiver attempts to form a conjecture that the sender discloses above some threshold and does not disclose below (and supports this conjecture by believing off-path that unexpected disclosures must have come from the highest state possible), the unique disclosure rule consistent with this conjecture is non-disclosure. In other words, if the sender wants sufficiently big *underestimates*, and the receiver insists on making the largest estimate possible following any disclosure, the sender will prefer to never disclose. Note that non-disclosure is *not* the unique disclosure rule for $b \leq -\frac{1}{4}$; we can see that an immediate corollary of our result is that if the sender's disclosure set takes the form $D(b) = [0, \bar{d}(b)]$, full disclosure with $\bar{d}(b) = 1$ will be an equilibrium disclosure rule for $b \leq -\frac{1}{4}$.

The following corollary follows immediately from our last two propositions:

Corollary 1. *If $b = 0$, the disclosure set $D^* = [\underline{d}^*, 1]$ can be supported in equilibrium and is ex-ante optimal in the class of lower censorship equilibria, where:*

- *If $e \leq \frac{3}{6+2\sqrt{3}}$ then $\underline{d}^* = \frac{2\sqrt{3}}{3}e$*
- *If $e > \frac{3}{6+2\sqrt{3}}$, $\underline{d}^*$ solves $3(2\underline{d} - 1) + \frac{(1-\underline{d})^3}{e} = 0$*

In other words, evaluating the sender's equilibrium disclosure rule at $b = 0$ tells us that if the sender's preferences are perfectly aligned with the receiver's, the sender uses the receiver optimal lower censorship disclosure rule. It follows immediately that ex-ante, the receiver's expected payoff will be decreasing in $|b|$; with a uniform state and uniform noise, both sender and receiver benefit from having more closely aligned preferences.

We have found that in a setting with a uniform state and uniform noise, an unbiased sender's unique equilibrium disclosure rule is efficient. We might wonder how far this insight generalises. We can see that in the uniform-uniform case whether a cutoff rule $\underline{d}$ constitutes an equilibrium strategy is determined solely by the distribution $\theta|\theta \in [0, \underline{d} + 2e]$. In other words, whether sender type $\underline{d}$ prefers to disclose or

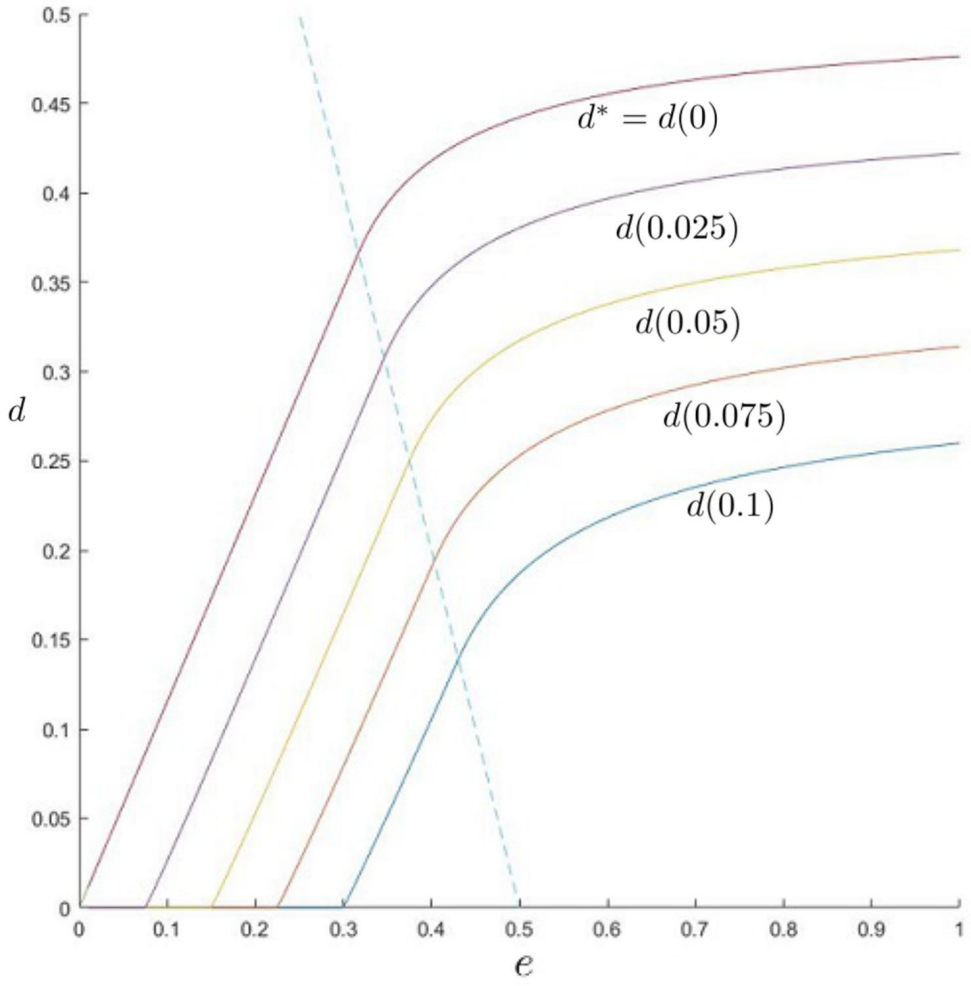

Figure 1: Efficient and Equilibrium Cutoff Rules: Uniform State, Uniform Noise

not disclose depends only on the receiver's actions at information sets that can be reached from sender type $\underline{d}$[26]. In contrast, the optimal lower censorship cutoff depends upon the full prior; as such, we can see that in general finding an equilibrium with an unbiased sender does *not* imply that we have an optimal cutoff; our next proposition makes precise the relationship between optimal cutoffs and equilibria with an unbiased sender.

[26]It follows that if a cutoff rule $\underline{d}$ constitutes an equilibrium strategy for the sender, it will remain an equilibrium strategy following a change in the distribution of $\theta|\theta > \underline{d} + 2e$ (provided that incentives to disclose remain monotone). We could thus rearrange the probability of types $\theta > \underline{d} + 2e$ (or add more high types $\theta > 1$) without altering the equilibrium strategy.

4.2 Preferences over Bias

We can show that when ex-ante optimal disclosure thresholds are characterised by a first order condition, this coincides exactly with the equilibrium indifference condition for an unbiased sender:

Proposition 10. *For θ and s with continuous densities, if the receiver expects the sender to use the optimal lower censorship disclosure rule $D^* = [d^*, 1]$, a sender observing d^* is indifferent between disclosure and non-disclosure iff $b = 0$.*

To sketch the reasoning; we can see that if we move a marginal type from disclosure to non-disclosure, this has two effects; firstly, if the sender observes that type, they now experience the loss associated with non-disclosure rather than the loss associated with disclosure; secondly, for all types, the updating rules the sender faces change, because the receiver is now less certain about the state on observing non-disclosure, and more certain about the state on observing disclosure. Our proof shows that these second effects are equal to 0, such that efficiency is characterised by setting the loss of the cutoff type from disclosure equal to their loss from non-disclosure.

This follows by *envelope theorem* style reasoning; for a fixed disclosure rule, the receiver's updating rule minimises the sender's ex-ante expected loss, because a sender's ex-ante objective is equivalent to the receiver's expected payoff. As such, the receiver's updating rule responds optimally to a change in disclosure cutoff; their average action following disclosure always matches the mean type who discloses, and their average action following non-disclosure always matches the mean type who does not disclose. As a result, when considering how a change in cutoff affects the sender's expected payoff, we only need to consider the direct effect on the cutoff type's expected payoff, because the optimality of the receiver's strategy implies that the indirect effect is 0.

Note this proposition equates an indifference condition with a first order condi-

tion; it does *not* say in general that equilibria with an unbiased sender are efficient. If the FOC is not sufficient for an optimum, there may be multiple equilibria, only one of which is efficient. Similarly if the indifference condition is not sufficient for an equilibrium, it may be that incentives to disclose are not monotone, such that for the efficient disclosure rule, some type other than the cutoff wants to deviate. As such, while equivalence of the indifference condition and FOC is a general result, it does not allow us to conclude that a receiver prefers an unbiased sender.

However, we know if we add Assumption 1 and assume $\theta|s$ has full support, satisfying an indifference condition is sufficient for equilibrium, while Proposition 5 tells us when there exists unique solution to the first order condition/indifference condition. As such, an immediate corollary is:

Corollary 2. *For θ and s with continuous densities, if Assumption 1 and the conditions in Proposition 5 hold, if the sender's disclosure set takes the form $D = [\underline{d}, 1]$, ex-ante the receiver prefers to face a sender with $b = 0$.*

Our result *also* tells us (without any additional assumptions) that an ex-ante optimal disclosure rule *cannot* be implemented in equilibrium with a biased sender. The inefficiency of equilibrium strategies with a biased senders is the result of a lack of commitment power for the sender; if the sender could publicly commit to a strategy before beginning the game, then for any bias level the sender would commit to disclosure set $D = [\underline{d}^*, 1]$. However, in the absence of commitment power, if the receiver anticipated a disclosure set $D = [\underline{d}^*, 1]$, for $b > 0$ a sender observing $\theta = \underline{d}^*$ would in fact strictly prefer to disclose. This arises because when considering what to do in state $\underline{d}^*$, a sender with commitment power takes into account how disclosing in that state affects the updating rule they will face in all other states following either disclosure or non-disclosure, while without commitment power a sender only considers what benefits them most in state $\underline{d}^*$ when sending the message. A biased sender with commitment power would thus accept that on average they will induce actions that are b away from their ideal action; an unbiased sender attempts to do

better, but in equilibrium is unsuccessful.

Preference for a Biased Sender We have argued that in a well-behaved continuous setting with monotone incentives, the receiver ex-ante prefers to face an unbiased sender. However, this does not preclude more general settings in which the receiver may ex-ante prefer to face a biased sender. Our result equates an indifference condition with a first order condition, but without additional assumptions does not tell us the optimal cutoff can be implemented in equilibrium. For example, it could be that there is one solution to our indifference condition which is ex-ante optimal but cannot be supported in equilibrium, *and* another solution which is not ex-ante optimal but can be supported in equilibrium.

To sketch how we might construct a counterexample in which a receiver ex-ante preferred a biased sender, suppose that noise is bounded. For some distributions, we may find that although at the ex-ante optimal lower censorship cutoff the cutoff type is indifferent, a type observing $\theta = 0$ strictly prefers to deviate to disclosure with an unbiased sender. In contrast, if the sender had larger bias, there may exist a cutoff for them that is *close* to the ex-ante optimal cutoff, which makes the cutoff type indifferent *and* deters deviation by a type observing $\theta = 0$. As such, a receiver may be better able to approximate their ex-ante optimal cutoff[27] with a *biased* rather than *unbiased* sender for specific distributions.

We can also note that our result is only valid when optimal cutoffs are characterised by a first order condition; in other words, when expected payoff is continuously differentiable in cutoffs. There are settings featuring continuously distributed states and signals in which this is not satisfied: in particular, if the state's *density* has discontinuities, optimal cutoffs may not satisfy a first-order condition if they occur at a jump in the state's density. Incentives to disclose will remain continu-

[27]Note, however, that as lower censorship may not be the ex-ante optimal disclosure rule for the sender in the class of *all* partial disclosure rules, here may exist other forms of disclosure rule which outperform partial disclosure in expectation and which *are* better approximated with an unbiased sender.

ous with discontinuous density (though not continuously differentiable) but are also much less likely to be monotone, meaning indifference may not be sufficient for an equilibrium. As such, while our indifference condition would remain necessary for an equilibrium by continuity of disclosure incentives, it would not be necessary for an optimal cutoff, so if there exists a cutoff at a jump in the state's density where an unbiased sender is not indifferent but expected payoffs are maximised, it might be that a receiver would ex-ante prefer a biased sender.

5 Extensions

We have studied a model in which a sender can either disclose the state exactly or send no message, the *decision* of the sender to disclose is perfectly observed, and noise in transmission only affects the *contents* of a disclosure. We could also consider a model in which the decision of the sender to disclose a particular form of message or not disclose was *not* perfectly observed. If, following disclosure, there is a positive probability the receiver does not observe a message (for example, because they do not ever come across the article written by the sender), the receiver no longer knows with certainty that no message was sent when no message was observed. If messages could be sent unintentionally (for example, if the sender accidentally forwards an email), the receiver would no longer be able to draw inference about the *support* of the state solely from the fact a message was sent.

In this section we briefly discuss how both of these model features would affect our results. We also discuss how our model would differ if we allowed the sender to make any true statement about θ.

5.1 Unreliable Channel

Firstly, consider a setting in which messages can go missing with positive probability (for example, if they travel to the receiver via unreliable intermediaries, such as

Myerson's famous carrier pigeon). Suppose now that if a sender sends a message, it is observed by the receiver with probability $1 - p$ (and with probability p they observe $\emptyset$)[28]. In this new model, if a receiver does not observe a message, they do not know if this is because no message was sent or because a message has gone missing.

Our equilibrium indifference condition is unchanged (with probability p disclosure and non-disclosure give the same payoff, so all that is relevant for the sender's decision is their expected payoffs given a message would arrive if sent). The updating rule $E(\theta|s)$ is also unchanged; however, $x_0 = E(\theta|\emptyset)$ will now be a function of p.

We can see that full disclosure equilibria will be impossible if messages can go missing. If a receiver anticipates full disclosure, $E(\theta|\emptyset) = \mu$, since $\emptyset$ is only reached if a message goes missing. But there are always sender types who prefer μ with certainty to the action distribution induced by disclosure, regardless of b (note this point does not depend on the presence of noise). As such, a positive measure of types will strictly prefer not to disclose.

It is more difficult to predict the exact shape of equilibrium partial disclosure rules with missing messages. If we take an equilibrium cutoff rule under the base model, and ask if we can support it as an equilibrium in the missing message model, we can note that for the cutoff type, the expected payoff from disclosure is unchanged, while the action induced by non-disclosure will be greater than it was in the base model. If the rise in action induced by non-disclosure is small (for example, because p is large), then we would expect non-disclosure to become more appealing, and so would expect equilibrium cutoff rules to feature more suppression.

[28] An alternative approach would be to add an additional information set, at which a receiver knows that a message was sent but does not know its contents (for example, because an email attachment has corrupted). We can see that compared to our base model equilibrium cutoff disclosure rules will feature less disclosure in this setting, since for a type who was indifferent between disclosing and not disclosing in our base model in a candidate equilibrium, the expected payoff from disclosure has fallen due to the possibility of faulty transmission while the payoff from non-disclosure is unchanged. Unlike when messages simply go missing, in this setting full disclosure equilibria are still possible for sufficiently biased senders.

Nonetheless, in general the effect of missing messages is ambiguous; in general, the necessary conditions for a unique equilibrium cutoff to exist will be stricter than the conditions for this cutoff to be strictly increasing in probability of missing messages. If types who disclose under the candidate cutoff rule are relatively more likely than types who do not disclose, the non-disclosure action could potentially increase by a significant amount as p rises, for a given cutoff rule. For example, consider a setting with $b \approx 0$, θ uniformly distributed but with an atom at $\theta = 1$, and $s|\theta \in [\theta - e, \theta + e]$. The existence of the atom does not affect the disclosure payoffs for the cutoff type of a lower censorship disclosure rule provided $\underline{d} < 1 - 2e$, but if the atom is sufficiently large, it is possible that for a fixed $\underline{d}$, $E(\theta|\emptyset) > \underline{d} + 2e$ when $p > 0$. If this were the case and the receiver expected low types not to disclose, disclosure could become more appealing for low types when $p > 0$ (while non-disclosure becomes more appealing for higher types). As a result, it is possible that for highly skewed priors, we may have equilibrium cutoffs which shift down as p rises; if the prios is sufficiently skewed, it may be that we cannot support lower censorship at all even for (small) positive bias.

5.2 Unexpected Messages

We might also be interested in whether our results are affected by the prospect of accidental messages. Suppose now that if a sender chooses not to disclose, the receiver observes m_0 with probability p and s with probability $1 - p$. If the receiver observes m_0, they can still infer that the sender intended to send no message, but if the receiver observes s, they do not know whether this is the result of intentional disclosure or not.

We can see that if a full disclosure equilibrium existed in the base model, it will continue to exist in the unexpected message setting, since the sender's disclosure condition remains unchanged. In contrast, as all information sets will now be on the equilibrium path (regardless of the support of $s|\theta$), we can see that no-disclosure

equilibria will fail to exist, since (if $b \geq 0$) a sender observing $\theta = 1$ will strictly prefer disclosure to μ with certainty.

In the case of partial disclosure, if we take an equilibrium cutoff rule in which the sender discloses only high states in our base model, and consider a type who was previously indifferent between disclosure and non-disclosure, we can see that their non-disclosure payoff remains unchanged, while the distribution of actions induced by disclosure must shift downwards. For any given signal realisation that could result from the cutoff type disclosing, we can argue that the receiver's posterior mean must be weakly lower with unexpected messages, since on observing a low signal the receiver is now unsure whether this is the result of an intentional message with a negative noise realisation or an unintentional message. As such, the action distribution induced in the base model must FOSD the action distribution induced in the unexpected message model.

From this we can conjecture that we take an equilibrium cutoff rule in our base model and consider the unexpected message setting, if the receiver expects the sender to use the same cutoff rule, the type which was previously indifferent between disclosure and non-disclosure will now strictly prefer disclosure, and so have an incentive to deviate in a candidate equilibrium. We would thus expect that for parameterisations in which equilibrium cutoff rules are unique, equilibria will feature more disclosure when unexpected messages can arrive.

5.3 Larger Message Space

One restrictive assumption in our model is that the sender chooses either no disclosure or to disclose the state exactly. A more standard assumption in canonical verifiable disclosure models is that the sender is permitted to make any true statement about the state of the world; in other words, they can report any subset of the state space containing the true state. We have restricted our analysis to single valued reports to benefit from the tractability afforded by being able to model noise

parametrically, while allowing noise to matter in equilibrium.

In a setting of additive noise, if the sender could report any subset containing the true state, even if the receiver observed a noisy realisation of values in that subset, the sender could still effectively bypass the noise using the cardinality of the subset. For example, if the sender reported n possible values of θ, and the receiver observed each of those n possible values plus a noise term, the sender could communicate via n rather than the possible values themselves. This would give them a countably infinite message space, and so they could achieve arbitrarily precise communication in the presence of additive noise with disclosures of the form: "here are n possible values of θ", where n picks out some rational number of given proximity to θ. A model featuring communication of this form in equilibrium arguably fails to capture the misunderstandings via language barriers which motivated our investigation of noisy disclosure.

If we wanted to generalise our model to a setting of noisy communication in which the sender could make any true statement about the state, we would thus need a different treatment of noise. One approach for a fully general treatment would be to define a stochastic noise mapping $N(m, \theta) : 2^{[0,1]} \times [0, 1] \to 2^{[0,1]}$ (where $m \in 2^{[0,1]}$ is the subset of $[0, 1]$ reported by the sender containing the true θ), which maps the sender's message, given the true state, to a subset of the support of θ. This general setting will be much less tractable, and the exact details of the sender's optimal strategy would depend on the properties of this noise mapping (we could think of our base model as mapping all $m \in [0, 1]$ to the value of $F_s(s)$, and mapping all $m \notin [0, 1]$ to $\emptyset$).

To build intuition about how the general setting would differ from our model, it is instructive to consider the example discussed above, in which the cardinality of messages is preserved by the mapping N. It is important to note that while a sender could achieve a noiseless communication language by communicating via the cardinality of their message, communication via cardinality is cheap talk. For that

reason, much of our intuition from a cheap talk model carries over; the sender can use the cardinality of the subset reported to send a cheap, coarse message about the state, which will determine the support of a receiver's posterior mean, on top of which, unlike in cheap talk, the receiver can draw inference from the contents of the subset reported, which will determine the value of their posterior mean.

If we consider sender-optimal equilibria, we can note the following: for small bias terms, the sender will utilise messages of different cardinalities, but for non-zero bias we will not see fully revealing equilibria, since this would require all information to be conveyed via cardinalities, which are cheap talk. For large biases, as all sender types prefer high actions, it cannot be that one cardinality of message used on-path induces a distribution of actions which FOSDs another cardinality of message; we know if communication were solely cheap talk, the unique equilibrium outcome for large bias would be garbling. In our setting, when the sender has a large bias term, in equilibrium the receiver can draw no meaningful inference from the cardinality of message sent, only from its content.

Returning to a fully general setting, we can carry over our intuition from the cardinality example as follows; when senders are trying to overcome noise in communication, they are likely to do better when we add additional dimensions to their messaging language. If they are perfectly aligned with a receiver, they will want the receiver to draw most inference from the dimensions of the messaging language which are least perturbed by the noise process[29]. If they are not perfectly aligned with the receiver, there may be a trade off in communication; since the least noisy dimensions of the messaging language may be cheap talk and so lack credibility, while the dimensions which are disciplined by truth telling may be more perturbed by the noise process. In a general setting, we would expect equilibrium disclosure strategies to combine cheap talk communication via different dimensions of the mes-

[29]If the noise process affects different dimensions of the messaging language independently, the receiver-optimal messaging strategy will utilise all of these dimensions, but the receiver will draw most inference from the least noisy dimensions.

saging language with 'hard' communication via the contents of messages, where the sender-optimal equilibrium strategy trades off the noisiness of dimensions against the credibility of messages utilising them (note that we know from Blume, Board, and Kawamura (2007) that if the sender's bias is large, it may be that the sender-optimal strategy actually utilises *noisier* dimensions of the messaging language for cheap talk messages, since when communication is via noisier channels deviations may be easier to deter.).

6 Related Literature

Within the literature on disclosure games, we contribute to a branch which seeks to explore when the unravelling result fails to hold. Much of the literature has focused on settings in which full disclosure remains the commitment solution for an unbiased sender, and so a failure of unravelling in equilibrium represents a source of inefficiency; in contrast, in our model, the commitment disclosure rule for an unbiased sender does not feature full disclosure, and partial disclosure can benefit both sender and receiver. As such, we contribute a new insight to the literature; a failure of unravelling may be not only an equilibrium outcome, but ex-ante desirable for the *receiver* when communication is via a noisy channel.

When the receiver is uncertain about whether the sender is informed, Dye (1985) shows that with a biased sender equilibria are not fully revealing. If the sender always discloses when they are informed, the receiver should inherit their prior when no message arrives, which motivates a biased sender to not disclose for a positive measure of states. Similarly, if messages went missing with positive probability, a biased sender would not choose full disclosure in equilibrium. Like Dye, our model of noisy disclosure highlights that the existence of a 'null message' (not disclosing) in communication games becomes important if communication is via an unreliable channel. However, in our setting, the existence of this null message may *benefit*

rather than harm the receiver, because it may help the sender overcome the friction in communication, rather than exploit it.

Mezzetti (2020) shows that unravelling can also fail in equilibrium when receivers are uncertain about the preferences of a sender. For example, if receivers are uncertain about whether a sender's bias is positive or negative, if the sender reports a range of values for the state they do not know whether to interpret this as a sender with positive bias hiding a low state, or a sender with negative bias hiding a high state. As a result, senders can gain from failing to fully disclose bad news (from their perspective) in equilibrium because the receiver is uncertain why the sender has not disclosed the exact state. Our model features known sender preferences but noisy disclosure and restricts senders to disclosure or no-disclosure, but because of noise in messages in our model multiple sender types are still pooled at message realisations in equilibrium. An important distinction is that in our model senders cannot opt out of this pooling because the noise is unavoidable, while in Mezzetti (ibid.) a sender can always deviate to noiselessly disclose the state exactly. As a result, while being pooled at message realisations may benefit *some* sender types in our model, this pooling *harms* other sender types, while in Mezzetti (ibid.) pooling on messages unambiguously benefits the sender and harms the receiver.

Jovanovic (1982) and Verrecchia (1983) show that if disclosure is costly, unravelling can fail in equilibrium with a fully biased sender. This insight extends naturally to a setting with unbiased senders, because regardless of bias, if the difference in action induced by disclosure compared to non-disclosure is small enough, the sender will not find paying the disclosure cost worthwhile. Our model can be thought of as extending this insight further by endogenising a cost to disclosure; instead of messages being inherently costly, the cost of disclosure is the risk of a large noise realisation inducing an action worse than the non-disclosure action. Our endogenous costs to disclosure are in general not constant but vary in a natural way with the sender's conjectured strategy; the more the sender discloses, the more costly

disclosure is for marginal types due to the greater possibility of misunderstandings.

Our results highlight the importance of the *form* of disclosure cost for policy and welfare analysis. Jovanovic (1982) shows that mandatory disclosure policies can harm both senders and social welfare when disclosure is costly. We share the insight that mandatory disclosure policies may harm the sender, but we *also* show that mandatory noisy disclosure can be harmful to *receivers*. If costs to disclosure are purely monetary, then a receiver would still prefer a sender to use full disclosure, and may prefer a fully biased sender who discloses more often. In contrast, when the cost to disclosure is the possibility of misunderstandings, this is experienced by both sender and receiver, and so the receiver optimal disclosure rule is not full disclosure, and the receiver may benefit less from the sender having large bias.

Our results are also highlight that in disclosure settings featuring frictions, comparative statics on sender bias depend upon the exact nature of the friction; in our model, the probability a sender discloses in equilibrium is increasing in sender bias, and observing a sender who discloses less frequently is evidence that the sender has a low level of bias if communication is noisy (similar results hold in settings of costly disclosure). In contrast, in a noiseless model, if there is positive probability a sender is uninformed and the sender's bias is small, increasing sender bias makes disclosure *less* likely. Understanding what drives non-disclosure is thus important for empirical work seeking to draw conclusions about preference misalignment from data on past communications between an expert and a decision maker.

The question of how noise in signals may influence communication and benefit agents has been explored in many settings outside of disclosure games. For example, Espinosa© Ray (2022) consider a setting which can be interpreted as the following game: a biased sender has two possible types, and would like to convince a receiver that she is the high type. The sender is committed to disclosing a noisy signal of her type, but can privately choose the variance of an additive noise term in the noisy signal she reveals (subject to a constraint which imposes a minimum variance on the

noise term). The receiver takes a binary action (hire or fire) and prefers only to hire the high type. With unbounded noise terms in equilibrium the low type chooses a noisier signal than the high type, and the receiver only hires for intermediate signal realisations (since very large signal realisations are more likely to come from a low type with a large noise term than a high type with a smaller noise term). In contrast, we fix the variance of noise but allow senders to choose whether to disclose their signal to the receiver (before observing the noise realisation). A common insight from both settings is that with positive bias, conditioning on sender type, larger noise terms are better for low sender types and smaller noise terms are better for high sender types.

Fishman and Hagerty (2003) study a disclosure problem in which with positive probability a receiver cannot process the *contents* of a disclosure, but can draw inference from observing that *some* disclosure was made. They focus on a setting with only two possible states and a fully biased sender who sells a product to customers and explore how the *probability* customers can comprehend disclosures affects equilibrium disclosure rules. In contrast, we assume the receiver's comprehension ability is common knowledge, and explore how sender bias level determines equilibrium disclosure rules. Our set-ups are quite different, but both models illustrate that the presence of some receivers with imperfect comprehension can break unravelling; if it is possible for some receiver types to draw incorrect inference from disclosure, we can have equilibria where with positive probability receivers do not disclose because they prefer the non-disclosure action with certainty to the possibility of disclosing and being misunderstood.

Quigley and Walther (2022) study the disclosure problem of a sender when the receiver has access to some outside information (a private noisy signal of sender type), when sender preferences are state independent. In an earlier draft, they show that allowing the receiver to observe an additional noisy signal of sender type can lead to reverse unravelling when additional disclosures are costly; very high sender

types are confident the receiver's outside information will reveal them to be high, so find disclosure too costly, while very low types do not disclose because doing so will only reveal them to be low types for sure. Intermediate sender types pay the cost of disclosure because they are too worried about the noisy signal leading the receiver to mistake them for a low type. Quigley and Walther (2022) focus on an information design problem in which a designer chooses the distribution of outside information, and show the designer is effectively constrained to provide a signal that is sufficiently informative that it crowds out disclosures by the sender.

Our model is quite different to theirs; our sender's preferences are state dependent, and our sender chooses whether or not to disclose a noisy signal to a receiver with no outside information. However, their results suggest an interesting direction for future work; exploring settings in which senders can choose the amount of noise in messages at a cost (in their setting, we can reinterpret outside information as the result of costless noisy disclosure, on top of which costly noiseless disclosure is possible). If costly noiseless disclosure were available in our setting, it would be first used by extreme types (who are worst off as a result of noisy disclosure), and we might expect equilibria of the following form (for small biases); for low states, the sender does not disclose, for intermediate states, they engage in noisy disclosure, and for high states, they engage in costly noiseless disclosure.

Noisy transmission in communication games has been explored in the context of cheap talk in Blume, Board, and Kawamura (2007). They consider a setting in which a privately informed sender observes the state of the world and can then costlessly send any message to an uniformed receiver via a noisy channel. In comparison to the seminal noiseless analysis of Crawford and Sobel (1982), they show that, fixing sender bias, the possibility of the receiver observing a message drawn from a noise distribution can *improve* expected equilibrium payoffs, and lead to more informative communication in equilibrium. We are interested in analysing similar questions in a setting of *verifiable* information: one in which a sender is restricted to choosing a

true message or no message. We differ in our approach to modelling noise; in Blume, Board, and Kawamura (2007), the receiver either observes the message sent by a sender or a message drawn from the noise distribution, where the noise distribution does not depend on the original message sent. In contrast, by restricting the sender's message space, we model noisy messages which *do* depend on the original message sent; this is arguably more intuitive in a setting of verifiable information as the messaging language is more likely to have a well defined intrinsic meaning. Our results differ significantly from theirs; in our setting, noise is always harmful in equilibrium, both via its direct effect and because it creates a commitment problem for the sender.

We motivate our study of noisy communication as a way to capture (broadly construed) language barriers in communication. Blume and Board (2013) also model communication in the presence of language barriers in a common interest setting. They explore a setting in which players have private information about their own language competence and communication is cheap talk. Because we are interested in communicating about *hard* information in the presence of language barriers, our approach is quite different; we model a language with *inherent* meaning with a receiver who has imperfect literacy. We abstract away from attempts by the sender to use language which is less likely to be misunderstood and assume the noise in messages is unavoidable; for example, because disclosures have to be written in legalese even when this makes them unclear to the receiver. The communication game we model is also more hierarchical than theirs, in the sense that we treat our sender as an *expert* who is known not just to have private information but also to have superior comprehension to the receiver; in contrast, their set-up allows for a sender with private information to have weaker linguistic competence than the receiver.

Andrews and Shapiro (2021) study a model of scientific communication in which an analyst privately observes evidence, and makes a report to receivers who may

be misaligned in either priors or preferences. Like us, an application of their model is the *strategic* problem of a sender communicating potentially complicated scientific evidence to a receiver. They focus on the commitment solution to the sender's problem rather than the sender's disclosure problem, and explore how optimal communication strategies differ depending upon whether the sender is trying to provide useful information to a decision maker, or directly recommending a decision, when the receiver may have private information or differing preferences. In contrast, in our model the sender understands the receiver's preferences and is only uncertain about what realisation of s they might observe following disclosure. As such, we explore a different feature of scientific communication; their sender is trying to provide useful information to an audience where they are uncertain about how the audience will respond to it, while our sender is trying to manipulate the decision of a non-expert receiver via disclosures given the receiver has imperfect comprehension.

We focus on a model in which both sender and receiver have quadratic preferences to study when a sender benefits from disclosing private information via a noisy channel. In a setting with quadratic preferences Kamenica and Gentzkow (2011) and Rehak and Senkov (2021) both explore under what conditions *ex-ante* a sender would like to *commit* to pooling states in their messaging strategy. Kamenica and Gentzkow (2011) show that if the sender's ideal action is linear in the receiver's ideal action, a sender either wants to commit to full or no disclosure, depending on multiplicative bias. Their results highlight than in our setting, partial disclosure is part of the commitment solution to the sender's disclosure problem only because of the noise in messages; absent noise, full disclosure would be the commitment solution.

Rehak and Senkov (2021) shows that this insight does not generalise beyond linear misalignment in ideal actions; if the sender's ideal action is non-linear in the receiver's ideal action, partial disclosure may also play a role in the commitment solution to the sender's disclosure problem with *no noise*. In other words, with non-

linear misalignment, partial disclosure may be used by the sender to deliberately *withhold* some information from the receiver, rather than as a coarse way to transmit *more* information than is possible with full disclosure. By focusing on linear misalignment, we isolate the signalling effect of non-disclosure.

Sequential games in which the second mover observes a noisy signal of the first mover's signal have also been studied in the context of noisy leader-follower games, following Bagwell (1995) (see also Maggi (1999) and Bhaskar (2009)). We differ from this literature in two ways; firstly, in our model payoffs depend only on the the state of the world (which is the sender's private information) and the second mover's action (the receiver's action), and secondly, the set of possible actions for the first mover (sender) depends on the state of the world. As such, our focus is on the 'signaling' concerns this literature abstracts away from.

7 Conclusions

We have analysed the effects of introducing noise to observations in a verifiable disclosure setting, and shown this can create a multiplicity of equilibria and lead to inefficient communication in equilibrium. Our results suggests that, in settings in which a sender is constrained to make either truthful reports or no reports, whether or not a receiver observes those reports accurately is important. If the receiver observes those reports with error, then the classic unravelling result may fail to hold; we would predict that a sender with sufficiently small bias will engage in partial disclosure. If disclosure is noisy then the sender's bias *is* payoff relevant in equilibrium; if the receiver is sceptical on seeing no message, more biased senders disclose more often (and *too* often from the receiver's perspective). Our results also show that, like in cheap talk settings, unless a sender is unbiased, a receiver will be better off if the sender can commit to a disclosure rule when disclosure is noisy.

One implication of the receiver optimal messaging rule featuring partial disclo-

sure is that, unlike with noiseless disclosure, partial disclosure policies can be welfare *improving*. For example, a scientific community which adopts the convention of only publishing results at a certain statistical significance level may improve the decision making of non-expert policy makers compared to a setting in which all results are always published. Policy makers with low statistical literacy can learn more when they know that seeing a finding published implies significant evidence against the null hypothesis than when they know all findings are always published, and are left to try and interpret the strength of the evidence themselves. While we do not suggest that our model explains why partial disclosure rules in scientific research are adopted, our results do suggest that a beneficial side effect may be improved communication with non-expert decision makers.

Our findings contrast the findings of Blume, Board, and Kawamura (2007) in cheap talk settings. While for a fixed messaging strategy adding noise to messages is always *mechanically* harmful, in a cheap talk setting, Blume, Board, and Kawamura (ibid.) show that noise may be *strategically* beneficial in equilibrium when consulting a biased expert. In contrast, we show that in a setting of verifiable disclosure, noise is *strategically* harmful with a biased sender, as equilibrium strategies in a noisy game may not maximise ex-ante expected payoffs taking the noise distribution as fixed. Our differing results here are unsurprising; when communicating with a biased sender, the cheap talk noiseless benchmark is garbling (no information transfer) while the disclosure noiseless benchmark is unravelling (full information transfer), so it is intuitive that noise cannot possible make things worse in a cheap talk setting, or make things better in a disclosure setting.

We have argued that mechanical obstacles to information transmission in a disclosure setting, such as language barriers, are not just harmful in themselves but also create further strategic problems. Frictions in transmission may prevent unravelling and lead to inefficient communication even with verifiable information, creating coordination and commitment problems between sender and receiver. When experts

can be misunderstood, knowing they are not lying does not imply they are acting in our best interests.

References

Andrews, Isaiah and Jesse M. Shapiro (2021). "A Model of Scientific Communication". In: *Econometrica* 89.5, pp. 2117–2142. DOI: 10.3982/ECTA18155. URL: https://ideas.repec.org/a/wly/emetrp/v89y2021i5p2117-2142.html.

Bagnoli, Mark and Ted Bergstrom (2005). "Log-Concave Probability and Its Applications". In: *Economic Theory* 26.2, pp. 445–469. ISSN: 09382259, 14320479. URL: http://www.jstor.org/stable/25055959 (visited on 07/07/2022).

Bagwell, Kyle (1995). "Commitment and observability in games". In: *Games and Economic Behavior* 8.2, pp. 271–280. ISSN: 0899-8256. DOI: https://doi.org/10.1016/S0899-8256(05)80001-6. URL: http://www.sciencedirect.com/science/article/pii/S0899825605800016.

Bhaskar, V. (2009). "Commitment and observability in a contracting environment". In: *Games and Economic Behavior* 66.2. Special Section In Honor of David Gale, pp. 708–720. ISSN: 0899-8256. DOI: https://doi.org/10.1016/j.geb.2008.08.003. URL: http://www.sciencedirect.com/science/article/pii/S0899825608001693.

Blume, Andreas and Oliver Board (2013). "Language Barriers". In: *Econometrica* 81.2, pp. 781–812. DOI: https://doi.org/10.3982/ECTA9183. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA9183. URL: https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA9183.

Blume, Andreas, Oliver Board, and Kohei Kawamura (2007). "Noisy talk". In: *Theoretical Economics* 2.4. URL: https://EconPapers.repec.org/RePEc:the:publsh:263.

Crawford, Vincent P. and Joel Sobel (1982). "Strategic Information Transmission". In: *Econometrica* 50.6, pp. 1431–1451. ISSN: 00129682, 14680262. URL: http://www.jstor.org/stable/1913390.

Dye, Ronald A. (1985). "Disclosure of Nonproprietary Information". In: *Journal of Accounting Research* 23.1, pp. 123–145. ISSN: 00218456, 1475679X. URL: http://www.jstor.org/stable/2490910.

Fishman, Michael J. and Kathleen M. Hagerty (2003). "Mandatory versus Voluntary Disclosure in Markets with Informed and Uninformed Customers". In: *Journal of Law, Economics, and Organization* 19.1, pp. 45–63. ISSN: 87566222, 14657341. URL: http://www.jstor.org/stable/3554972.

Grossman, Sanford J. (1981). "The Informational Role of Warranties and Private Disclosure about Product Quality". In: *The Journal of Law & Economics* 24.3, pp. 461–483. ISSN: 00222186, 15375285. URL: http://www.jstor.org/stable/725273.

Hagenbach, Jeanne and Frederic Koessler (2017). "Simple versus rich language in disclosure games". In: *Review of Economic Design* 21.3, pp. 163–175. URL: https://EconPapers.repec.org/RePEc:spr:reecde:v:21:y:2017:i:3:d:10.1007_s10058-017-0203-y.

Harbaugh, Richmond and Theodore To (2020). "False modesty: When disclosing good news looks bad". In: *Journal of Mathematical Economics* 87.C, pp. 43–55. DOI: 10.1016/j.jmateco.2018.10. URL: https://ideas.repec.org/a/eee/mateco/v87y2020icp43-55.html.

Jovanovic, Boyan (1982). "Truthful Disclosure of Information". In: *The Bell Journal of Economics* 13.1, pp. 36–44. ISSN: 0361915X. URL: http://www.jstor.org/stable/3003428.

Kamenica, Emir and Matthew Gentzkow (2011). "Bayesian Persuasion". In: *American Economic Review* 101.6, pp. 2590–2615. DOI: 10.1257/aer.101.6.2590. URL: https://www.aeaweb.org/articles?id=10.1257/aer.101.6.2590.

Maggi, Giovanni (1999). "The Value of Commitment with Imperfect Observability and Private Information". In: *The RAND Journal of Economics* 30.4, pp. 555–574. ISSN: 07416261. URL: http://www.jstor.org/stable/2556065.

Mezzetti, Claudio (2020). "Manipulative Disclosure". In: 1250. URL: https://ideas.repec.org/p/wrk/warwec/1250.html.

Milgrom, Paul R. (1981). "Good News and Bad News: Representation Theorems and Applications". In: *The Bell Journal of Economics* 12.2, pp. 380–391. ISSN: 0361915X. URL: http://www.jstor.org/stable/3003562.

Onuchic, Paula (2021). "Information Acquisition and Disclosure by Advisors with Hidden Motives". working paper. URL: https://paulaonuchic.com/s/InfAcq10.pdf.

Quigley, Daniel and Ansgar Walther (2022). "Inside and Outside Information". working paper. URL: https://danielquigley.weebly.com/uploads/8/4/0/8/84081230/insideoutsideinfo_2020_01_27_jf_revision.pdf.

Ray Debraj© Espinosa, Francisco (2022). "Too Good To Be True? Retention Rules for Noisy Agents". Forthcoming, American Economoc Journal: Microeconomics. URL: https://debrajray.com///wp-content/uploads/2022/01/EspinosaRayAEJ.pdf.

Rehak, Rastislav and Maxim Senkov (2021). "Form of Preference Misalignment Linked to State-Pooling Structure in Bayesian Persuasion". working paper. URL: https://drive.google.com/file/d/10zSvbeqb-hPjdMF90elxCWR8TxoZ59J1/view.

Seidmann, Daniel J. and Eyal Winter (1997). "Strategic Information Transmission with Verifiable Messages". In: *Econometrica* 65.1, pp. 163–169. ISSN: 00129682, 14680262. URL: http://www.jstor.org/stable/2171817.

Szalay, Dezso (2013). "Strategic Information Transmission and Stochastic Orders". working paper. URL: https://www.econ1.uni-bonn.de/pdf-files/strategic-information-transmission-and-stochastic-orders.

Verrecchia, Robert E. (1983). "Discretionary disclosure". In: *Journal of Accounting and Economics* 5, pp. 179–194. ISSN: 0165-4101. DOI: https://doi.org/10.

1016/0165-4101(83)90011-3. URL: https://www.sciencedirect.com/science/article/pii/0165410183900113.

A Proofs for Section 3 (Equilibria)

Proposition 1. *For there to exist an equilibrium in which the sender discloses for all $\theta \in [0,1]$ it is sufficient that $|b| \geq \frac{1}{2}$, and necessary that $b \neq 0$.*

Proof. To see the first claim holds, consider a candidate equilibrium in which $x_0 = 0$ and R believes all sender types disclose. If $b > \frac{1}{2}$ it follows that all sender types strictly prefer any distribution of actions on $[0,1]$ with strictly positive variance to $x = 0$ with certainty. Since non-disclosure induces $x = 0$ with certainty, while for any $\theta \in [0,1]$ disclosure induces an action distribution with support in $[0,1]$ with strictly positive variance, it follows all types prefer to disclose.

If $b < -\frac{1}{2}$ then if $x_0 = 1$, and R believes all sender types disclose, by symmetry all sender types strictly prefer to disclose.

To see the second claim holds, let $b = 0$, and consider a candidate equilibrium in which $x_0 = 0$ and R believes all sender types disclose.

The sender's equilibrium disclosure condition becomes:

$$-(E(x|\theta) - 2\theta)E(x|\theta) > Var[x|\theta]$$

Note we can rewrite the RHS as $Var[x(s)|\theta] = E[x(s)^2|\theta] - [E(x(s)|\theta)]^2$. Hence our condition is:

$$2\theta E(x(s)|\theta) > E[x(s)^2|\theta]$$

Now consider the condition evaluated at $\theta = 0$. The LHS is zero while the RHS is strictly positive, so the sender has a strict preference for non-disclosure. As both sides of the condition are continuous in θ at $\theta = 0$ and the preference is strict for $\theta = 0$, it also follows that there exists some $\delta > 0$ such that the sender prefers non-disclosure if $\theta < \delta$. □

Proposition 2. *For non-disclosure for all $\theta \in [0,1]$ to be an equilibrium strategy for the sender, it is sufficient that:*

- *The conditional state distribution $\theta|s$ has full support for all signal realisations s,* ***or****:*

- *For all θ and s we have $\theta|s \in [s-e, s+e]$, $E(s|\theta) = \theta$, and:*

  - $\diamond$ *$b \geq \mu$, $1-\mu < e$, and if $e > \frac{1}{2}$, $s|\theta$ is log-concave,* ***or****:*

  - $\diamond$ *$b \leq -(1-\mu)$, $\mu < e$, and if $e > \frac{1}{2}$, $s|\theta$ is log-concave.*

Proof. For the first claim, consider a candidate equilibrium in which $x_0 = E(\theta) = \mu$ and S does not disclose for all $\theta \in [0,1]$. No information set at which R observes a message is on the equilibrium path. As the support of ϵ is unbounded, R is free to believe that all messages observed at off-path information sets originate from the same sender type. If R believes following an off-path message that $\theta = x_0$ is the only type who could have deviated (i.e $E_R(\theta|s) = \mu \ \forall s \in \mathbb{R}$), then if the sender sends no message they induce $x = \mu$, while if they send a message they also induce $x = \mu$. As all S types are indifferent between messaging and sending no message, always sending no message is an equilibrium strategy for S.

For the second claim, define $x_l(s)$ to be the most sceptical off-path updating rule the sender can use. Now note if $s = \theta + \epsilon$ for mean-zero $\epsilon \in [-e, e]$, then we have $E(x_l(s)|\theta = 1) = 1 - e + E(\epsilon|\epsilon \geq e-1, \theta)P(\epsilon \geq e-1|\theta)$. Suppose $b \geq \mu$, so that a sufficient condition for non-disclosure is that disclosure induces a lower average action for all θ. This will be hardest to satisfy for $\theta = 1$.

Hence we want $E(\epsilon|\epsilon \geq e-1, \theta)P(\epsilon \geq e-1|\theta) < \mu - (1-e)$. If $e \leq \frac{1}{2}$ then this is just $1 - \mu < e$, so a sufficient condition is $1 - \mu < e \leq \frac{1}{2}$ and $b \geq \mu$.

Now suppose we increase e. We can see that if increasing e above $\frac{1}{2}$ reduces $E(x_l(s)|\theta = 1)$, then $1 - \mu < e$ and $b \geq \mu$ remains a sufficient condition for non-disclosure. If $e > \frac{1}{2}$ then the average action from disclosure must be larger than $1-e$, but as e rises $1-e$ itself becomes smaller so the overall change in average disclosure action as e rises is ambiguous. If $s|\theta$ is log-concave then $\frac{\partial E(\epsilon|\epsilon \geq e-1, \theta)}{\partial e} < 1$ so the fall in $1 - e$ dominates, so we have the following sufficient condition: if $1 - \mu < e$ and

$b \geq \mu$ then for log-concave noise ϵ we have no-disclosure in equilibrium.

A symmetric argument holds for negative bias. □

Proposition 3. *Define the difference between signal s and state θ as $s - \theta = \epsilon$. Suppose the following conditions hold:*

- *For a conjectured disclosure set $\hat{D}$, the receiver's beliefs satisfy $\frac{\partial E(\theta|s,\hat{D})}{\partial s} \in [0,1]$.*
- *For all θ, changes in the density of $\epsilon|\theta$ as θ changes $|\frac{\partial f_{\epsilon|\theta}(\epsilon)}{\partial \theta}|$ are sufficiently small.*

*Then incentives to disclose are increasing in θ for $\hat{D} = [\underline{d}, 1]$ if $\theta + b > x_0$ and **either** $\theta \in \hat{D}$ **or** $s|\theta$ has full support.*

Proof. Let's begin by writing incentives to disclose:

$$E[U_S(x(s), \theta, b)|\theta] - U_S(x_0, \theta, b) = x_0^2 - E[x(s)^2|\theta] + 2(E[x(s)|\theta] - x_0))(\theta + b)$$

Now writing out in integral terms:

$$x_0(x_0 - 2(\theta + b)) - \int x(s)(x(s) - 2(\theta + b)) f_{s|\theta}(s) ds$$

Let's compute the difference between incentives to disclose for type $\theta + \delta$ and type θ. With some rearranging:

$$2\delta \left[\left(\int x(s) f_{s|\theta+\delta}(s) ds - x_0 \right) - 2 \int \frac{x(s)(x(s) - 2(\theta + b))}{\delta} (f_{s|\theta+\delta}(s) - f_{s|\theta}(s)) ds \right] \quad (4)$$

We are computing the difference in expected value of net disclosure payoff across two different signal distributions. Note that if $s - \theta = \epsilon$ and ϵ is independent of θ, we have $f_{s|\theta}(s) = f_\epsilon(s - \theta)$ and $f_{s|\theta+\delta}(s) = f_\epsilon(s - \theta - \delta)$. As such, an equivalent way to compute this difference in expected value is to take the difference in expected payoff *for each ϵ realisation* and compute expectations across ϵ. In other words, for any

function $y(s)$ and ϵ and θ independent, we can perform the following transformation:

$$\int y(s)(f_{s|\theta}(s) - f_{s|\theta+\delta}(s))ds = \int (y(\theta+\delta+\epsilon) - y(\theta+\epsilon))f_\epsilon d\epsilon$$

Now suppose θ and ϵ are not independent. We can still apply a similar logic to our transformation, but we will now have an additional remainder term. Before we fixed a noise realisation ϵ (whose probability didn't depend on θ), and took the difference in receiver actions for that noise realisation between state $\theta+\delta$ and state θ respectively.

If we do the same now, we also need to take into account that the probability of a given noise realisation now depends on θ. So for realisation ϵ we need to compare $x(\theta+\delta+\epsilon)$ at probability $f_{\epsilon|\theta+\delta}$ to $x(\theta+\epsilon)$ at probability $f_{\epsilon|\theta}$. We have:

$$\int y(s)(f_{s|\theta} - f_{s|\theta+\delta})ds = \int \Big([y(\theta+\delta+\epsilon) - y(\theta+\epsilon)]f_{\epsilon|\theta} + [f_{\epsilon|\theta+\delta} - f_{\epsilon|\theta}]y(\theta+\delta+\epsilon) \Big) d\epsilon$$

Let's divide this term by δ:

$$\left[\int \left(\frac{y(\theta+\delta+\epsilon) - y(\theta+\epsilon)}{\delta} f_{\epsilon|\theta} + \frac{f_{\epsilon|\theta+\delta} - f_{\epsilon|\theta}}{\delta} y(\theta+\delta+\epsilon) \right) d\epsilon \right]$$

We have:

$$\lim_{\delta\to 0}[.] = \left[\int y'(\theta+\epsilon)f_{\epsilon|\theta} + \frac{\partial f_{\epsilon|\theta}}{\partial\theta} y(\theta+\epsilon)ds \right] \tag{5}$$

Now let's return to equation (4). We can see the limit of equation (4) as $\delta \to 0$ will be 0, but we are interested in whether we approach this limit from above or below.

Computing the limit of (4) divided by 2δ, and substitute in equation (5) with $y(s) = x(s)(x(s) - 2(\theta+b)$ and $y'(s) = 2x'(s)(x(s) - (\theta+b))$:

$$\int \Big([x(\theta+\epsilon)(1-x'(\theta+\epsilon)) + x'(\theta+\epsilon)(\theta+b)]f_{\epsilon|\theta} + \frac{\partial f_{\epsilon|\theta}}{\partial\theta} x(\theta+\epsilon)[x(\theta+\epsilon) - 2(\theta+b)] \Big) d\epsilon - x_0$$

Inspecting the first term we can note that if $x' \in [0,1]$, then for each value of ϵ this

first term computes a convex combination of $x(\theta+\epsilon)$ and $\theta+b$, and then averages these convex combinations with respect to $\epsilon|\theta$.

We can thus note that for this first term to be larger than x_0 it is sufficient that it is the average of convex combinations of two terms which are both larger than x_0. Hence a sufficient condition for this first term to be larger than x_0 is that $\theta+b>x_0$ and $x(s)>x_0\ \forall s$ that could be induced by disclosure in state θ (and vice versa for this term to be less than x_0).

What this tells us is that if $\theta+b>x_0$ and $x(s)>x_0\ \forall s$ that could be induced by disclosure in staet θ, then incentives to disclose will be strictly positive if either our second term is positive, or if our second term is sufficiently small. A sufficient condition for our second term to be small is that $|\frac{\partial f_{\epsilon|\theta}(\epsilon)}{\partial\theta}|$ is bounded above sufficiently tightly.

It remains to ask for which θ we will have $\theta+b>x_0$ and $x(s)>x_0$ for all signals that could be reached by disclosure. We can see that a sufficient condition for $x(s)>x_0$ for all possible message realisations is that $\theta\in\hat{D}=[\underline{d},1]$, so that all messages a sender could send are on the equilibrium path. If we add the condition that the support of s is constant for all θ then there are no off-path messages so $x(s)>x_0$ holds regardless of the state.

Hence a sufficient condition for incentives to disclose to be increasing in θ is that the receiver conjectures the sender engages in lower censorship in equilibrium, $s-\theta=\epsilon$ for $|\frac{\partial f_{\epsilon|\theta}(\epsilon)}{\partial\theta}|$ sufficiently small, θ and ϵ satisfying $x'(s)\in[0,1]$, and $\theta+b>x_0$ *and* either $s|\theta$ has full support or the receiver expects θ to disclose.

Analogously a condition for incentives to disclose to be decreasing in θ is that the receiver conjectures the sender engages in upper censorship in equilibrium, $s-\theta=\epsilon$ for $|\frac{\partial f_{\epsilon|\theta}(\epsilon)}{\partial\theta}|$ sufficiently small, θ and ϵ satisfying $x'(s)\in[0,1]$, and $\theta+b<x_0$, *and* either $s|\theta$ has full support or the receiver expects θ to disclose. □

Proposition 4. *If the distributions of θ and s have continuous densities, the support of $\theta|s$ does not depend upon s, and Assumption 1 holds, there exists an equilibrium*

in which the sender discloses if $\underline{d}(b) \leq \theta$*, where there exist* $\underline{b} < 0 < \bar{b}$ *such that* $\underline{d}(b) = 0$ *for* $b > \bar{b}$ *and* $\underline{d}(b) \in (0,1)$ *for* $\underline{b} < b < \bar{b}$*].*

Proof. In our first proposition we have argued that full-disclosure equilibria exist for sufficiently large biases. We will argue as follows; if the receiver believes the sender discloses states only above some threshold, and we start from a large bias, and reduce it, at some point the equilibrium disclosure threshold must rise above 0.

Let's fix a parameterisation, and suppose the receiver conjectures that the sender engages in full-disclosure with $x_0 = 0$. Starting from a large positive bias, we will ask how equilibrium disclosure rules will behave as we reduce b. Given Assumption 1 holds, for full disclosure equilibria to exist with large positive bias it is sufficient to check that a sender observing $\theta = 0$ strictly prefers disclosure, which is the case if:

$$\frac{E[(x(s))^2|\theta = 0])^2}{2E[x(s)|\theta = 0])} \leq b$$

We can see that if bias is above this threshold, full-disclosure equilibria must exist, while if bias is strictly below this threshold, full-disclosure equilibria cannot exist, because a sender observing $\theta = 0$ (or θ in a neighbourhood of 0 of positive measure) strictly prefer non-disclosure given full-disclosure is expected.

It remains to ask whether lower-censorship equilibria exist, given that full-disclosure equilibria do not. Restrict attention to $D = [\underline{d}, 1]$. We can see that if $\underline{d} = 0$ incentives to disclose are strictly negative for bias below our threshold. We can also see that for $\underline{d} = 1$, if $-\frac{1}{4} < b$, incentives to disclose are strictly positive.

If the distributions of θ and s have continuous densities then incentives to disclose will be continuous in $\underline{d}$, and so the intermediate value theorem tells us that there must be some $\underline{d} \in (0,1)$ that satisfies our equilibrium indifference condition when $-\frac{1}{4} < b < \frac{E[(x(s))^2|\theta=0])^2}{2E[x(s)|\theta=0])}$.

It remains to check that this cutoff can be supported in equilibrium. Assumption 1 tells us that incentives to disclose are increasing for $\theta > x_0 - b$. Full support for

$\theta|s$ tells us that we do not need to consider off-path beliefs, so with no sender type $\theta \leq x_0 - b$ can gain from disclosure if $x_0 - b \leq \underline{d}$, as doing so gets them an action above $\underline{d}$ which they disprefer to x_0. It follows that full support for $\theta|s$ and Assumption 1 tell us that if $-(\underline{d} - x_0) \leq b$ and $-\frac{1}{4} \leq b$, if $\underline{d}$ solves our indifference condition it can be supported in equilibrium.

As such, we can reason as follows: if we start with large bias and reduce it, below some threshold full-disclosure equilibria must cease to exist, but under the conditions in the proposition lower censorship equilibria *will* exist, provided bias does not become *too* negative. □

Proposition 5. *Fix the distributions of s, θ, and the bias b. There exists at most* ***one*** $\underline{d} \in [0,1]$ *such that $D = \hat{D} = [\underline{d}, 1]$ is an equilibrium disclosure set if the following conditions holds:*

- *Assumption 1 is true.*
- $\frac{\partial E(\theta|\theta \notin \hat{D})}{\partial \underline{d}} < 1$
- $\frac{\partial E(\theta|s, \theta \in \hat{D})}{\partial \underline{d}} < 1$ *for* $E(\theta|s, \theta \in \hat{D}) > \underline{d} + b$
- $\frac{\partial E(\theta|\theta \notin \hat{D})}{\partial \underline{d}} < \frac{\partial E(\theta|s, \theta \in \hat{D})}{\partial \underline{d}}$ *for* $E(\theta|s, \theta \in \hat{D}) \leq \underline{d} + b$.

Proof. Our proof strategy is to show that if a solution to the sender's equilibrium indifference condition exists, it is unique.

Consider a conjectured disclosure set $\hat{D} = [\underline{d}, 1]$. Let's write $x_{\underline{d}}(s)$ for the updating rule based on conjectured cut-off $\underline{d}$, and $x_{\underline{d}}$ for the non-disclosure action.

Suppose the sender observes state $\theta = \underline{d}$. Their incentives to disclose are given by:

$$E[U_S(x_{\underline{d}}(s), \underline{d}, b)|\theta = \underline{d}] - U_S(x_{\underline{d}}, \underline{d}, b) = x_{\underline{d}}^2 - E[x_{\underline{d}}(s)^2|\theta = \underline{d}] + 2(E[x_{\underline{d}}(s)|\theta = \underline{d}] - x_{\underline{d}}))(\underline{d} + b)$$

In integral terms:

$$\int (x_{\underline{d}} - x_{\underline{d}}(s))(x_{\underline{d}} + x_{\underline{d}}(s) - 2(\underline{d} + b)) f_{s|\underline{d}}(s) ds$$

Necessary conditions for an equilibrium cutoff $\underline{d}^*$ are that incentives to disclose are equal to 0, $\underline{d}^* > x_{\underline{d}} - b$, and incentives to disclose are increasing in θ for $\theta > x_{\underline{d}} - b$.

Our argument will be as follows; we know a valid candidate equilibrium cutoff must satisfy $\underline{d} > x_{\underline{d}} - b$. If we take incentives to disclose in state $\underline{d}$ (for $\underline{d} > x_{\underline{d}} - b$), we can ask how they vary with $\underline{d}$ when we take into account that in equilibrium as $\underline{d}$ changes the receiver's updating rules also change. If for all valid candidate equilibrium cutoffs, increasing $\underline{d}$ (keeping receiver conjecture $\hat{\underline{d}} = \underline{d}$) increases incentives to disclose, it follows there must be at most one $\underline{d}$ such that if the receiver conjectures $\hat{\underline{d}} = \underline{d}$, incentives to disclose equal 0. The proposition follows.

Let's write:

$$y(\underline{d}, s) = (x_{\underline{d}} - x_{\underline{d}}(s))(x_{\underline{d}} + x_{\underline{d}}(s) - 2(\underline{d} + b))$$

If we compare incentives to disclose with cutoff $\underline{d} + \delta$ to incentives to disclose with cutoff $\underline{d}$, we can write the change in disclosure incentives as follows:

$$\int \Big(y(\underline{d} + \delta, s) f_{s|\underline{d}+\delta}(s) - y(\underline{d}, s) f_{s|\underline{d}}(s) \Big) ds$$

We can rewrite this as:

$$\int \Big(y(\underline{d} + \delta, s)[f_{s|\underline{d}+\delta}(s) - f_{s|\underline{d}}(s)] + f_{s|\underline{d}}(s)[y(\underline{d} + \delta, s) - y(\underline{d}, s)] \Big) ds$$

We can see that the first term compares disclosure payoff in state $\underline{d} + \delta$ to disclosure payoff in state $\underline{d}$ for a *fixed* updating rule. As such, if incentives to disclose are increasing in θ, we know this term will be positive. Hence if Assumption 1 holds and $\underline{d} + b > x_0$, the first term is positive.

Let's now focus on the second term.

$$\int f_{s|\underline{d}}(s)[y(\underline{d}+\delta,s)-y(\underline{d},s)]ds$$

We can see that if this is positive, that is sufficient for increasing conjectured $\underline{d}$ to make disclosure more appealing for the sender in state $\underline{d}$.

To sign this we can divide by δ and take the limit to obtain the partial derivative of the term with respect to $\underline{d}$:

$$\lim_{\delta\to 0}\int f_{s|\underline{d}}(s)\frac{y(\underline{d}+\delta,s)-y(\underline{d},s)}{\delta}ds=\int f_{s|\underline{d}}(s)\frac{\partial y(\underline{d},s)}{\partial \underline{d}}ds$$

This is the average change in incentives to disclose as $\underline{d}$ changes, holding fixed the distribution of $s|\underline{d}$. We can see that a sufficient condition for this to be positive is if $\frac{\partial y(\underline{d},s)}{\partial \underline{d}}>0\ \forall s$. If this average change in incentives to disclose as $\underline{d}$ changes is positive, and incentives to disclose are increasing in θ for a fixed receiver conjecture and $\theta > x_a - b$, then it follows that overall incentives to disclose for conjectured cutoff type $\underline{d}$ are increasing in $\underline{d}$, and so at most one value of $\underline{d}$ can give indifference between disclosure and non-disclosure.

We have:

$$\frac{\partial y(\underline{d},s)}{\partial \underline{d}}=\left(\frac{\partial x_{\underline{d}}}{\partial \underline{d}}-\frac{\partial x_{\underline{d}}(s)}{\partial \underline{d}}\right)(x_{\underline{d}}+x_{\underline{d}}(s)-2(\underline{d}+b))+(x_{\underline{d}}-x_{\underline{d}}(s))\left(\frac{\partial x_{\underline{d}}}{\partial \underline{d}}+\frac{\partial x_{\underline{d}}(s)}{\partial \underline{d}}-2\right)$$

Rearranged:

$$\frac{\partial y(\underline{d},s)}{\partial \underline{d}}=2\left[(x_{\underline{d}}(s)-(\underline{d}+b))\left(1-\frac{\partial x_{\underline{d}}(s)}{\partial \underline{d}}\right)+(\underline{d}+b-x_{\underline{d}})\left(1-\frac{\partial x_{\underline{d}}}{\partial \underline{d}}\right)\right]$$

We know that $\frac{\partial x_{\underline{d}}(s)}{\partial \underline{d}}$ and $\frac{\partial x_{\underline{d}}}{\partial \underline{d}}$ will be positive, given that as $\underline{d}$ increases, for each signal realisation there are fewer low states compatible with it, and for non-disclosure, there are more high states compatible with it.

We know that $\underline{d} + b > x_{\underline{d}}$ is a necessary condition for any equilibrium cutoff $\underline{d}$. When $b \leq 0$, we also have $\underline{d} + b \leq x_{\underline{d}}(s)$ for any signal realisation s. As such, we can note that if $b \leq 0$, a sufficient condition for $\frac{\partial y(\underline{d},s)}{\partial \underline{d}} > 0$ is that $\frac{\partial x_{\underline{d}}(s)}{\partial \underline{d}}, \frac{\partial x_{\underline{d}}}{\partial \underline{d}} < 1$.

Suppose instead that $b > 0$. As before if we have $\underline{d} + b \leq x_{\underline{d}}(s)$ for signal realisation s a sufficient condition for $\frac{\partial y(\underline{d},s)}{\partial \underline{d}} > 0$ is that $\frac{\partial x_{\underline{d}}(s)}{\partial \underline{d}}, \frac{\partial x_{\underline{d}}}{\partial \underline{d}} < 1$. However, note that there there may now exist signal realisations such that $\underline{d} + b > x_{\underline{d}}(s)$.

We want:

$$\frac{x_{\underline{d}}(s) - x_{\underline{d}}}{\underline{d} + b - x_{\underline{d}}} + \frac{\underline{d} + b - x_{\underline{d}}(s)}{\underline{d} + b - x_{\underline{d}}} \frac{\partial x_{\underline{d}}(s)}{\partial \underline{d}} > \frac{\partial x_{\underline{d}}}{\partial \underline{d}}$$

Our upper bound on $\frac{\partial x_{\underline{d}}}{\partial \underline{d}}$ now comprises a term that is less than 1 plus another term that is less than 1 multiplied by $\frac{\partial x_{\underline{d}}(s)}{\partial \underline{d}}$. As such, we can see that even if $\frac{\partial x_{\underline{d}}}{\partial \underline{d}} < 1$ we may have $\frac{\partial y(\underline{d},s)}{\partial \underline{d}} < 0$.

A sufficient condition for $\frac{\partial x_{\underline{d}}}{\partial \underline{d}}$ to satisfy this upper bound is $\min\{\frac{\partial x_{\underline{d}}(s)}{\partial \underline{d}}, 1\} \geq \frac{\partial x_{\underline{d}}}{\partial \underline{d}}$.

Hence we can say the following; if Assumption 1 holds (so incentives are monotone for $\theta + b > \underline{d}$ and a fixed disclosure threshold), $\frac{\partial x_{\underline{d}}(s)}{\partial \underline{d}}, \frac{\partial x_{\underline{d}}}{\partial \underline{d}} < 1$ for $\underline{d} + b < x_{\underline{d}}(s)$, *and* $\min\{\frac{\partial x_{\underline{d}}(s)}{\partial \underline{d}}, 1\} \geq \frac{\partial x_{\underline{d}}}{\partial \underline{d}}$ for $\underline{d} + b \geq x_{\underline{d}}(s)$, then as we increase conjectured $\underline{d}$, disclosure becomes more appealing for a sender observing $\underline{d}$ for any $\underline{d} + b > x_{\underline{d}}$.

It follows that there is at most one $\underline{d}^*$ such that $\underline{d}^* + b > x_{\underline{d}^*}$ and $E[U_S(x_{\underline{d}^*}(s), \underline{d}^*, b) | \theta = \underline{d}^*] = U_S(x_{\underline{d}^*}, \underline{d}^*, b)$, so equilibria in the class of lower censorship disclosure rules are unique. □

Proposition 6. *For $\theta \sim U[0,1]$ and $s = \theta + \sigma\varepsilon$ for $\varepsilon \in \mathbb{R}$ log-concave, symmetric, and independent of θ, as $\sigma \to 0$, for $b > 0$ the following can be supported as limiting equilibrium disclosure sets $\lim_{\sigma \to 0} D(b)$*

- *If $b \in (0, \frac{1}{4})$, $\lim_{\sigma \to 0} D(b) \in \{[0,1], [1-4b, 1], \emptyset\}$*
- *If $b > \frac{1}{4}$, $\lim_{\sigma \to 0} D(b) \in \{[0,1], \emptyset\}$.*

Proof. Given symmetry, we will focus on lower-censorship equilibria.

Our equilibrium indifference condition is

$$2\left(\frac{\frac{\underline{d}}{2}+E(x(s)|\theta=\underline{d})}{2}-(\underline{d}+b)\right)(\frac{\underline{d}}{2}-E(x(s)|\theta=\underline{d}))=Var[x(s)|\theta=\underline{d}]$$

As $\sigma \to 0$ we have $E[x(s)|\theta] \to \theta$ (for $s \in \hat{D}$) and $Var[x(s)|\theta] \to 0$, so our equilibrium indifference condition tends to:

$$\left(\frac{\underline{d}}{4}+b\right)\underline{d}=0$$

This tells us that if the receiver conjectures that $\underline{d}=0$ or $\underline{d}=-4b$, a sender observing $\underline{d}$ is indifferent between disclosing and not-disclosing.

We can see $\underline{d}=0$ can only be an equilibrium cutoff if $b \geq 0$, else a positive measure of types will strictly prefer non-disclosure. For $b \geq 0$ as disclosure becomes noiseless, all sender types are willing to disclose with $x_0=0$, so this can be supported in equilibrium.

Turning to $\underline{d}=-4b$, we can see that can only be supported as an equilibrium cutoff if $b<0$. If we have a lower censorship equilibrium with $b<0$ and disclosure for states above $-4b$, no sender observing a state above $-4b$ prefers non-disclosure, as this gives loss of $(\theta+3b)^2$ compared to a loss of b^2 from disclosure, and loss from non-disclosure is larger when $\theta>-4b$. As ε is unbounded a sender observing a state below $-4b$ will receives only actions greater than $-4b$ from unexpected disclosures, while they receive $-2b$ from disclosure; as their ideal action is less than $-3b$ it follows they strictly prefer non-disclosure.

As this holds for negative bias in lower censorship equilibrium, it follows by symmetry it holds for positive bias in upper censorship equilibrium, so if $b<\frac{1}{4}$, $D(b)=[0,1-4b]$ will be an equilibrium disclosure set.

Recall that for $\varepsilon \in \mathbb{R}$ we can also always support a trivial no-disclosure equilibrium with $x(s)=E(\theta)$ for all s. Our proposition follows. □

Proposition 7. *For $\theta \sim U[0,1]$ and $s=\theta+\varepsilon$ for $\varepsilon \in \mathbb{R}$ log-concave, symmetric,*

and independent of θ, as $\sigma \to \infty$, for $b \geq 0$ equilibrium disclosure sets converge to:

- *If* $b \in [0, \frac{1}{4})$, $\lim_{\sigma \to \infty} D(b) \in \{[0, \frac{1}{2} - 2b], [\frac{1}{2} - 2b, 1], \emptyset\}$
- *If* $b \geq \frac{1}{4}$, $\lim_{\sigma \to \infty} D(b) \in \{[0, 1], \emptyset\}$

Proof. Given our distributional assumptions, a solution to the limit of our equilibrium indifference condition will characterise an equilibrium disclosure rule.

As $\sigma \to \infty$ our equilibrium indifference condition for a lower censorship equilibrium tends to:

$$2\left(\frac{\frac{\underline{d}}{2} + \frac{1+\underline{d}}{2}}{2} - (\tilde{\underline{d}} + b)\right)(\frac{\underline{d}}{2} - \frac{1+\underline{d}}{2}) = 0$$

which we can rearrange to obtain:

$$\underline{d} = \frac{1}{2} - 2b$$

It follows if $b \in [0, \frac{1}{4})$, we can support $D(b) = [\frac{1}{2} - 2b, 1]$ and if $b \geq \frac{1}{4}$ we can support $D(b) = [0, 1]$.

By symmetry we can conclude that if we can support $\underline{d} = \frac{1}{2} - 2b$ as a lower censorship cutoff, we can also support $\bar{d} = \frac{1}{2} - 2b$ as an upper censorship cutoff, because the limits of the equilibrium indifference conditions are identical. It follows if $b \in [0, \frac{1}{4})$, we can also support $D(b) = [0, \frac{1}{2} - 2b]$.

As before, non-disclosure equilibria also continue to exist. The proposition follows. □

B Proofs for Section 4 (Efficiency)

Lemma 1. *The messaging rule which maximises the sender's ex-ante expected payoff also maximises the receiver's ex-ante expected payoff provided the receiver is best responding.*

Proof. The sender's ex-ante expected payoff is:

$$-E((x-\theta-b)^2) = -[E((x-\theta)^2) - 2bE(x-\theta) + b^2]$$

If $E(x) = E(\theta)$, which is the case whenever the receiver is Bayesian and best responds to the sender's messaging rule, then this simplifies to:

$$-E((x-\theta-b)^2) = -[E((x-\theta)^2) + b^2]$$

The receiver's ex-ante expected payoff is given by:

$$-E((x-\theta)^2)$$

As b^2 is a constant, it follows that the problem of maximising the sender's ex-ante expected payoff is equivalent to the problem of maximising the receiver's ex-ante expected payoff. □

Proposition 8. *If the sender's disclosure set is given by $D = [\underline{d}, 1]$, ex-ante expected payoffs are maximised when $D^* = [\underline{d}^*, 1]$, where:*

- *If $e \leq \frac{3}{6+2\sqrt{3}}$ then $\underline{d}^* = \frac{2\sqrt{3}}{3}e$*
- *If $e > \frac{3}{6+2\sqrt{3}}$ then $\underline{d}^*$ solves $3(2\underline{d}^* - 1) + \frac{(1-\underline{d}^*)^3}{e} = 0$*

Proof. The receiver's ex-ante expected payoffs are:

$$-\underline{d}\left(\frac{\underline{d}^2}{12}\right) - \left[2e(Var(\theta|s)|s \in [\underline{d}-e, \underline{d}+e]) + (1-2e-\underline{d})\frac{(2e)^2}{12}\right]$$

where we use symmetry to note that we must have $E(Var(\theta|s)|s \in [\underline{d}-e, \underline{d}+e]) = E(Var(\theta|s)|s \in [1-e, 1+e])$.

Proceeding from here using the conditional distributions derived under the as-

sumption that $\underline{d} + e \leq 1 - e$, expected payoffs are:

$$-\left(\frac{\underline{d}^3}{12}\right) - \frac{(1-\underline{d})e^2 + e^3}{3}$$

We can see that this is a strictly concave function for $\underline{d} > 0$ and is maximised by choosing $\underline{d}^* = \frac{2\sqrt{3}}{3}e$. Note we assumed that $\underline{d} + e \leq 1 - e$. This will be satisfied if $e \leq \frac{3}{6+2\sqrt{3}}$.

Repeating the above steps under the assumption that $\underline{d} + e > 1 - e$ we find that $\underline{d}^*$ solves $3(2\underline{d} - 1) + \frac{(1-\underline{d})^3}{e} = 0$ The assumption that $\underline{d} + e > 1 - e$ will be satisfied if $e > \frac{3}{6+2\sqrt{3}}$. □

Proposition 9. *Suppose the sender's disclosure set takes the form $D(b) = [\underline{d}(b), 1]$ or $D(b) = \emptyset$. For each possible value of $-\frac{1}{4} < b$, there is one value of $D(b)$ that can be supported in equilibrium, where:*

- *If $b \geq \frac{e}{3}$ and $e \leq \frac{1}{2}$, $\underline{d}(b) = 0$*
- *If $b \geq \frac{3e-1}{12e-3}$ and $e > \frac{1}{2}$, $\underline{d}(b) = 0$*
- *If $b < \frac{e}{3}$ and $\underline{d}(b) \leq 1 - 2e$, $\underline{d}(b) = 2(\sqrt{b^2 + \frac{e^2}{3} - be} - b)$*
- *If $-\frac{1}{4} < b < \frac{3e-1}{12e-3}$ and $\underline{d}(b) > 1 - 2e$, $\underline{d}(b)$ solves:*

$$1 + 3\underline{d}^2 + 6\underline{d}e + 3b(4e + 2\underline{d} - \underline{d}^2 - 1) = 3e + 3\underline{d} + \underline{d}^3$$

- *If $b \leq -\frac{1}{4}$, $D(b) = \emptyset$ or $D(b) = [0, 1]$*

Proof. Our equilibrium indifference condition (for an interior solution for $\underline{d}$) is:

$$2\left(\frac{\frac{d}{2} + E[x(s)|\theta = \underline{d}]}{2} - (\underline{d} + b)\right)(\frac{\underline{d}}{2} - E[x(s)|\theta = \underline{d}]) = Var[x(s)|\theta = \underline{d}] \quad (6)$$

We can see that, for $\theta = \underline{d}$, $s \in [\underline{d} - e, \underline{d} + e]$, and so:

$$E(\theta|s) = \int_{\underline{d}}^{s+e} \frac{\theta}{s + e - \underline{d}} d\theta = \frac{s + e + \underline{d}}{2}$$

for $s \in [\underline{d} - e, \underline{d} + e]$ and so, noting that $s|\underline{d} \sim U[\underline{d} - e, \underline{d} + e]$, we have:

$$E(E(\theta|s)|\theta = \underline{d}) = \frac{2\underline{d} + e}{2}$$

We can also derive $Var[E(\theta|s)|\theta = \underline{d}]$:

$$\begin{aligned} Var[E(\theta|s)|\theta = \underline{d}] = &\int_{\underline{d}-e}^{\underline{d}+e} \left(\frac{s + e + \underline{d}}{2}\right)^2 \frac{1}{2e} ds - \left(\frac{2\underline{d} + e}{2}\right)^2 \\ = &\frac{8(\underline{d} + e)^3}{24e} - \frac{8\underline{d}^3}{24e} - \frac{(2\underline{d} + e)^2}{4} \\ = &\frac{e^2}{12} \end{aligned}$$

Our condition is thus:

$$2\left(\frac{\frac{\underline{d}}{2} + \frac{2\underline{d}+e}{2}}{2} - (\underline{d} + b)\right)\left(\frac{\underline{d}}{2} - \frac{2\underline{d} + e}{2}\right) = \frac{e^2}{12}$$

which yields:

$$3\underline{d}^2 + 12b(\underline{d} + e) = 4e^2$$

Solving for $\underline{d}$ we have:

$$\underline{d}(b) = 2\left(\sqrt{b^2 + \frac{e^2}{3} - be} - b\right)$$

We can see that $\underline{d}(b) \geq 0$ for $b \leq \frac{e}{3}$ (else we cannot make the sender indifferent, and they always disclose).

Repeating this derivation for $\underline{d} > 1 - 2e$, we find that $\underline{d}(b)$ solves:

$$1 + 3\underline{d}^2 + 6\underline{d}e + 3b(4e + 2\underline{d} - \underline{d}^2 - 1) = 3e + 3\underline{d} + \underline{d}^3$$

which has a positive solution if $b \leq \frac{e-\frac{1}{3}}{4e-1}$ (else the sender always discloses) and is less than 1 if $b > -\frac{1}{4}$ (else the sender never discloses, if the receiver is maximally optimistic on seeing no message).

It remains to verify that no type above cutoff type $\underline{d}(b)$ prefers non-disclosure and no type below prefers disclosure. Suppose we set receiver's off-path updating rule to be $E(\theta||s) = s + e$ for $s \in [-e, \underline{d}(b) - e]$. We can note that under the distributions derived, $\frac{\partial E(\theta|s,\hat{D})}{\partial s} \in [0, 1]$ for all s, and $\frac{\partial f_{\epsilon|\theta}}{\partial s} = 0$. As such, the proof of Proposition 1 tells us that for all $\theta + b > E(\theta|\theta \leq \underline{d})$, incentives to disclose are increasing in θ. For $\theta + b < x_0$, we verify computationally that a maximally optimistic off-updating rule will deter deviation by $\theta \in [0, x_0 - b]$ for the disclosure thresholds $\underline{d}(b)$ given in our proposition.

Uniqueness follows by our uniqueness proposition, as $\frac{\partial E(\theta|\theta<\underline{d}}{\partial \underline{d}} = \frac{1}{2}$ and $\frac{\partial E(\theta|s,\hat{D}}{\partial \underline{d}} \in \{\frac{1}{2}, 1\}$ for the distributions derived (given off-path updating rule $E(\theta|s, s+e < \underline{d}) = s + e$). □

Proposition 10. *For θ and s with continuous densities, if the receiver expects the sender to use the optimal lower censorship disclosure rule $D^* = [d^*, 1]$, a sender observing d^* is indifferent between disclosure and non-disclosure iff $b = 0$.*

Proof. We will show that the indifference condition which is necessary for an equilibrium disclosure rule matches the first order condition that is necessary for an ex-ante optimal disclosure rule iff $b = 0$.

Assume the receiver conjectures the sender uses cutoff $\underline{d}$. Setting sender loss from non-disclosure equal to loss from disclosure in state $\underline{d}$ with $b = 0$ gives:

$$(\underline{d} - x_0)^2 = E[(\underline{d} - x(s, \underline{d}))^2 | \theta = \underline{d}]$$

The receiver optimal $\underline{d}^*$ will solve:

$$\max_{\underline{d}} - \int_0^{\underline{d}} (\theta - x_0(\underline{d}))^2 f_\theta d\theta - \int_{\underline{d}}^1 \int (\theta - x(s, \underline{d}))^2 f_{s|\theta} ds f_\theta d\theta$$

where for clarity we make it explicit that the receiver's updating rule depends upon $\underline{d}$.

This is a continuously differentiable function of $\underline{d}$ if θ and s have continuous densities, and we can easily verify that it is not maximised by either $\underline{d} = 0$ or $\underline{d} = 1$, so an optimal $\underline{d}^*$ must satisfy a first order condition.

Differentiating the loss relating to non-disclosure:

$$-(\underline{d} - x_0(\underline{d}))^2 f_\theta(\underline{d}) - 2\frac{\partial x_0(\underline{d})}{\partial \underline{d}} \int_0^{\underline{d}} (\theta - x_0(d)) f_\theta d\theta$$

We can note $x_0(\underline{d}) = \int_0^{\underline{d}} \theta \frac{f(\theta)}{F(\theta} d\theta$. Hence:

$$\int_0^{\underline{d}} f_\theta \theta d\theta - F_\theta(\underline{d}) x_0(\underline{d}) = 0$$

and so our second term must be 0. In other words, as $\underline{d}$ rises, the rise in $x_0(d)$ balances it out such that additional loss is only experienced at the margins.

Turning to differentiating loss related to disclosure:

$$\int (\underline{d} - x(s, \underline{d}))^2 f_{s|\theta=\underline{d}} ds f_\theta(\underline{d}) - 2 \int_{\underline{d}}^1 \int (\theta - x(s, \underline{d})) f_{s|\theta} \frac{\partial x(s, \underline{d})}{\partial \underline{d}} ds f_\theta d\theta$$

Again, let's focus on our second term. Changing the order of integration:

$$-2 \int \int_{\underline{d}}^1 (\theta - x(s, \underline{d})) f_{s|\theta} f_\theta d\theta \frac{\partial x(s, \underline{d})}{\partial \underline{d}} ds$$

We can note:

$$x(s, \underline{d}) = \int_{\underline{d}}^1 \frac{f_{s|\theta} f_\theta}{\int_{\underline{d}}^1 f_{s|\theta} f_\theta d\theta} \theta d\theta$$

Hence:

$$\int_{\underline{d}}^1 f_{s|\theta} f_\theta \theta d\theta - x(s, \underline{d})) \int_{\underline{d}}^1 f_{s|\theta} f_\theta d\theta = 0$$

and so:

$$-2\int\int_{\underline{d}}^{1}(\theta - x(s,\underline{d}))f_{s|\theta}f_{\theta}d\theta\frac{\partial x(s,\underline{d})}{\partial \underline{d}}ds = 0$$

Again, as $\underline{d}$ rises, the average guess by the receiver following disclosure also rises, such that additional loss is only experienced at the margins.

Combining, we have the following FOC:

$$\int(\underline{d} - x(s,\underline{d}))^2 f_{s|\theta=\underline{d}}ds f_{\theta}(\underline{d}) = (\underline{d} - x_0(\underline{d}))^2 f_{\theta}(\underline{d})$$

Cancelling, we have:

$$E[(\underline{d} - x(s,\underline{d}))^2|\theta = \underline{d}] \;=\; \int(\underline{d} - x(s,\underline{d}))^2 f_{s|\theta=\underline{d}}ds \;=\; (\underline{d} - x_0(\underline{d}))^2$$

which is exactly our indifference condition. As such, any solution to our FOC, and in particular the ex-ante optimal cutoff, must make the cutoff type indifferent between disclosure and non-disclosure. If Assumption 1 holds and ϵ is unbounded, a cutoff type being indifferent would also be sufficient for that cutoff to be an equilibrium cutoff.

To see that we can only make a type with $b = 0$ indifferent at the efficient cutoff, note the ex-ante optimal cutoff is independent of bias, but incentives to disclose are not. With $b \neq 0$ our indifference condition can be written:

$$E[(\underline{d} - x(s))^2|\theta = \underline{d}] + 2b(x_0 - E[x(s)|\theta = \underline{d}]) \;=\; (\underline{d} - x_0))^2$$

If we take a cutoff that made a sender with $b = 0$ indifferent, the first and third terms must be equal, while the second term will be negative if $b > 0$ (and positive with $b < 0$. It follows that if the receiver conjectures the sender uses a cutoff $\underline{d}^*$ that makes a sender with bias $b = 0$ indifferent, a sender with $b > 0$ must strictly prefer disclosure and a sender with $b < 0$ must strictly prefer non-disclosure The proposition follows. □

Chapter 4

Reputational Incentives with Networked Customers

Reputational Incentives with Networked Customers

Aidan Smith

August 8, 2022

Abstract

We propose a model of reputational incentives for a firm with privately known type whose customers share reviews via a social network. The firm can choose a non-contractible effort level for each customer, which stochastically improves that customer's review. When customers base their purchase decisions on their friends' reviews, firm incentives for effort are stronger if a customer's review will be read by many friends, and if those friends are not too connected (since their beliefs are easier to influence). From the perspective of ex-ante social welfare, this creates a trade-off between providing incentives and generating learning; more connected networks may allow customers to learn more, but remove firm incentives to exert effort when serving them. When effort is sufficiently productive, ex-ante expected total surplus can be higher when the social network has disjoint components. Our results imply that interventions to increase customer review visibility (such as auto-translate features) may be harmful for social and consumer welfare, if firm effort is sufficiently productive.

Keywords: Learning, Network, Reputation

JEL Classification: D82, D83, D85, L14

1 Introduction

When customers encounter a new firm for the first time, they may be uncertain about its type. For example, diners may be uncertain about whether the chef at a new restaurant is talented. Early customers will base their purchase decisions on their prior beliefs about the firm's type, but later customers can learn about the firm by observing the purchase decisions and experiences of previous customers.

When earlier customers share reviews with later customers, this can create *reputational* incentives for a firm. In a one-shot setting, if effort is chosen after customers have already decided to purchase, a firm would have no incentive to engage in costly high effort. However, if good reviews today will lead to higher revenue tomorrow, a firm may find it optimal to take costly non-contractible actions when serving customers who visit today. For example, a chef might take more care when preparing a meal, or waiting staff might pay more attention to a table.

In many settings, customers do not observe the full history of reviews, but instead learn via word-of-mouth from their friends on social networks. If firms observe customer networks, they should take into account a customer's popularity when choosing whether to exert high effort for them; if a customer has no friends and only visits a firm once, no future customers will observe their experience, so a firm has no incentives for effort when serving them, while if a customer has many friends, a firm may have strong incentives for effort. Firm incentives when serving a customer will depend not just on how connected that customer is, but on how connected their friends are; if their friends are well connected and have already seen many reviews of the firm, incentives will be weaker on average than when their friends are less well informed and so easier to influence.

This paper asks the following questions: how does customer social network structure determine firm incentives for effort? Which social networks and customer orders should a social planner design if they seek to maximise societal welfare? Our contribution lies in combining a model of reputational incentives with networked learning,

and our analysis generates both testable predictions and policy implications.

Our research questions are increasingly important at a time when social networks, platforms, and review aggregators (e.g. Instagram, Etsy, Tripadvisor) all have significant power to influence customer access to past reviews; algorithms can both hide reviews we do want to see (by determining when we see friends' posts) and show us reviews we didn't ask for (via promoted posts). Features such as 'auto-translate' can allow customers to learn from reviews that would otherwise be costly/cumbersome to process. Regulators often lack the knowledge of how platforms manipulate review visibility and have limited scope to intervene (for example, a regulator cannot observe or control Facebook's algorithms). Our results provide an important policy insight by giving conditions when increasing review visibility unambiguously improves social welfare, and when it does not.

We address our research questions by building a simple model in which a firm with a privately known type serves customers located on a social network, who visit according to some exogenous order. The firm charges a fixed price, and if a customer purchases, the firm chooses whether to exert low or high effort when serving them. The firm is either a commitment 'low-skill' type, who always provides bad service (exerting low effort), or a 'high-skill' type, who sometimes provides good service. Customers are willing to pay more if the firm is high-skill, and are willing to pay more when a high-skill firm exerts high effort. Customers observe the purchase decisions and quality of service provided to their neighbours on the social network, and update their beliefs about the firm's type via Bayes' rule.

For a given network and customer order, we explore what outcomes can be supported in equilibrium. Our main prediction is that persistent good reviews for a firm over time are more likely when customers observe a limited sample of past reviews, rather than the full history, as illustrated in Proposition 3. This arises because when customers observe the full history of reviews, firms can stop exerting high effort following early success and 'coast' on the back of their good reputation. In

contrast, when customers networks are less connected, news of early success spreads less widely, meaning firms have less incentive to shirk following early good reviews.

While we find there are equilibria with persistent high effort when customer networks are less connected, we also find for some networks there may be equilibria featuring *herding*, in which the firm exerts high effort for at most the first customer, and all remaining customers purchase if the first customer wrote a good review. In particular, Proposition 1 shows that when customers are located on a directed rooted tree, if customers further from the 'root' move later, there is an equilibrium in which only the 'root' receives high effort, and all remaining customers purchase if the 'root' writes a good review (either because they observe the 'root', or because they copy their friends). Customers who copy their friends' purchase decisions reason as follows: 'if my friend is sufficiently well informed, her *decision* to visit a restaurant reveals *she* must have learned the firm has had sufficiently many good reviews, so even if she has a bad experience herself I should ignore this and copy her purchase decision'.

Proposition 2 gives conditions under which equilibria are unique, but in general we find multiple equilibria for a given network and customer order. Multiple equilibria arise because equilibria have a feature of 'self-fulfilling prophecy': if customers expect a firm to be exerting high effort, they judge the firm more harshly if it receives bad reviews, which in turn can mean the firm is motivated to high effort. In contrast if they expect the firm to exert low effort, they are less sceptical of its skill level when it receives bad reviews, which can in turn mean the firm chooses not to exert effort. Our results suggest that customers may be better off when they are more optimistic about unknown firms, in the sense that for a given network, if customers expect high effort as often as possible in equilibrium, average customer payoffs may be much higher than if customers expect low effort as often as possible in equilibrium.

We investigate which social networks and customer orders a planner looking to

maximise expected social welfare would design. We show that a social planner can strengthen firm incentives by having reviewers visit simultaneously, because when reviewers visit sequentially, a firm may quit exerting high effort early if initial reviews are good enough. We show that if a planner wants to induce high effort for as many customers as possible, providing the strongest possible aggregate incentives to the firm does not in general show all later customers all past reviews; Proposition 5 shows incentives are strongest when later customers see relatively few reviews because the probability that each one of those reviews might determine their decision is larger. Proposition 8 shows that a consequence of this feature of incentives is that when high effort is sufficiently productive, efficient networks will not show customers the full history of reviews, and may have disjoint components. We argue this implies that when reputational incentives are strong, features such as 'auto-translating reviews' may harm rather than benefit total welfare.

The paper is structured as follows; in section 2, we set out the model. In section 3, we discuss what outcomes can be supported in equilibrium for a given social network and fixed order of customers. In section 4, we discuss which networks and customer orders maximise social welfare. Section 5 concludes and discusses how our analysis could be extended to more general settings. In the remainder of this section, we discuss how our paper relates to the literature. Proofs are in the appendix.

Related Literature Our paper's contribution lies in combining a model of reputational incentives with a model of social learning on a network. Our model has some similarities to Mailath and Samuelson (2001)'s model of reputations, although in contrast to their infinite horizon model in which a firm sets a price each period, we consider a firm with finite life span and a fixed price, so in our setting reputational incentives for effort are at the *extensive* margin and will eventually diminish with time. The main distinction of our model is we restrict customers to only observe their neighbours on a social network; in our model, the *identity* of a customer will

matter for a firm's effort choice (in other words, we allow *asymmetry* in how much of a game's history customers can observe).

The reputations literature has explored various settings in which customers are restricted to observing only a subset of the history of play. Pei (2022) provides a useful discussion of related literature, and considers a setting with observational learning where customers observe a bounded subset of the history of play between a long-run player and a sequence of short-run players. He shows that allowing a customer to sample a bounded number of previous customer's decisions *in addition* to the history of firm actions can be harmful, in the sense that when customers can also observe previous customers' decision, firms are only guaranteed min-max rather than Stackelberg payoffs in equilibrium. This has a similar flavour to our results on herding; in our model, when customers base their decisions on previous customers' decisions rather than on the outcomes of previous firm effort, this removes firm incentive for effort and can lead to significantly lower equilibrium payoffs.

A branch of the reputations literature following Ely and Välimäki (2003) explicitly studies when reputational incentives for a firm can be harmful for customers. They show that when an honest advisor (for example, an auto-mechanic) has a reputational incentive to distinguish themselves from a fully biased advisor, their incentives to signal their type via their advice can become so strong that they end up giving advice which is completely useless to their current customer, meaning in equilibrium customers avoid their services. Reputational incentives in our model differ in that the customer always prefers high effort from our firm, whereas with an advisor the customer's preferred action depends upon a random state privately observed by the advisor. However, our paper shares the feature that if customers can observe too much about the history of play, they may receive worse service on average if they purchase; in our model this is because a firm which has revealed itself to be a high type can quit high effort, while in Ely and Välimäki (ibid.) this is because 'good' firms have an incentive to take actions that are harmful to signal

their type[1].

In contrast to the broader reputations literature we relax two common assumptions. Firstly, we allow customers to ex-ante differ in information quality for reasons other than the date; using the language of networks allows us accommodate ex-ante *asymmetry* in what customers could observe even if they visited at the same date. Secondly, we do not restrict customers to visiting a firm sequentially, and we show there can be significant welfare gains from allowing customers to visit simultaneously, because this can strengthen incentives for the long-run player.

Our model of learning links to the social learning literature following Banerjee (1992) and Bikhchandani, Hirshleifer, and Welch (1992) (see Golub and Sadler (2016) for a recent survey of models of learning on social networks). Like Acemoglu et al. (2011) and Board and Meyer-ter-Vehn (2021), we focus on Bayesian learning restricted to learning from neighbours on a social network, but unlike in standard social learning models, we do not endow customers with access to a private signal; customers are not born with some inherent private information about firm type, and all 'signals' in our model are the endogenous result of firm effort choice. Board and Meyer-ter-Vehn (ibid.) also feature endogenous information, but in their setting, this is the result of customers trialling a product before purchase, while in our setting information is generated by purchases themselves. Like Board and Meyer-ter-Vehn (ibid.), we focus on learning in the short-run; we are interested in outcomes when a firm serves each member of a finite set of customers exactly once, rather than investigating whether asymptotic learning occurs.

Our model features no private signals, but allow agents to observe both one another's actions and payoffs, with the ability to observe one another's payoffs taking the role of private signals in a classic social learning model. Wolitzky (2018) also

[1]The analogy between our models would be closer if in our model high effort gave a quality of service that revealed the firm to be a high type, but was not actually preferred by customers to low effort (for example, a chef might produce a meal which required high technical skill but was not actually preferred by customers to a simple dish cooked well). We note that we can parameterise our model such that high effort is inefficient from the perspective of social welfare, in which case a *social planner* might view reputational incentives as bad in our setting.

studies a setting in which all learning is via agents observing one another's payoffs. His setting features agents choosing between a costly safe action and a costly action with random rewards, who learn about the expected reward of the risky action by observing earlier customers. He finds that learning is efficient when the random action is more likely to generate a high payoff, but that inefficiency is persistent if the costly action is cheaper than the safe action but less likely to generate a high payoff.

The key distinction between his model and ours is that his customers can learn about the expected reward of the risky action from observing friends who chose the *safe action* (other distinctions are that our customers can observe both past decisions and payoffs, and in our setting, the return to the risky action depends upon the *strategic* effort choice of the firm). This means that when customers observe a random sample of the history, as this sample becomes large they learn efficiently about costly but high return actions. In contrast, in our model, once customers switch to all taking the safe action, no new information about the firm can be generated, and so learning can stall. If customers choosing their outside option generated new information about the firm, incentives would differ from our model, as a good firm would be less scared of acquiring a bad reputation, knowing if they do it is still possible to rebuild a good reputation.

Both the reputations and social learning literatures commonly fix the *order* in which customers visit[2], and explore different observation rules for customers visiting sequentially. In contrast, we allow for both the customer order and observation structure to vary. We note that in the absence of discounting or a strategic long-run player, interventions on customer orders can be mapped to interventions on customer social networks with sequential visits. For example, Sgroi (2002) shows that in a social learning setting where customers observe the full history of play,

[2]A common assumption is that customer visit dates are drawn from some continuous distribution on the unit interval; we note here that absent discounting, this assumption is equivalent to assuming that customers visit the firm sequentially, one per period.

average payoffs are higher when some later customers are moved to visit in the first period (as when herding begins, herds are better informed). Absent discounting the effect Sgroi identifies does not depend inherently on *timing*; Acemoglu et al. (2011) develop a similar insight using the language of networks in a setting with one customer per period, where the first n customers do not observe one another, and the remainder can observe the full history. In our setting, however, interventions on timing are not in general equivalent to interventions on networks, because of their effect on firm incentives; a firm forced to choose efforts simultaneously for two customers faces a different problem to a firm who can observe the outcome with the first customer before choosing effort for the second.

Combining a model of learning on a network with a model of reputational incentives allows us to explore a key trade-off when incentives are reputational and customers observe endogenous signals; observation structures which encourage efficient learning may provide weaker incentives for effort. Our results on efficient networks highlight that policy conclusions drawn from social learning models may be inappropriate when firms can exercise discretionary effort; in a social learning model, it is always optimal ex-ante to allow the final customer to visit a firm to observe the full history; when firms have reputational incentives, this is no longer true (ex-ante, thought it may sometimes be true ex-post), because making the final customer better informed can remove firm incentives for effort for earlier customers. As such, our paper shows the importance of studying reputational incentives and social learning *together*, and suggests significant scope for future work studying more general reputation problems with networked short-run players.

2 Model

A long-lived firm F serves N customers $\mathcal{C} = \{C_1, .., C_N\}$ across $T \geq N$ periods. The firm has a privately known type $\theta \in \{0, 1\}$, and customers initially have a prior

belief that $P(\theta = 1) = \pi$. If $\theta = 0$, the firm is a 'low-skill' commitment type who will always provide customers with bad service; if $\theta = 1$ the firm is a 'high-skill' type who is capable of providing good service.

Each customer visits the firm once, takes a purchase decision, and receives a gross payoff equal to the quality of service received. Each customer is allocated a period in which to visit accordingly to a publicly observed order $\succeq_\alpha$: $\mathcal{C} \to \{1, ..., T\}$. We will write C_1^t to denote that $\succeq_\alpha$ allocates customer C_1 to visit in period t. Customers are located on a commonly known network $\mathcal{N}$, and observe the purchase decisions and experiences of their earlier neighbours on $\mathcal{N}$; we write $N(C_k)$ to denote the set of C_k's neighbours on $\mathcal{N}$.

In period t, the firm serves all $\mathcal{C}^t = \{C_i^\tau \in \mathcal{C} | \tau = t\}$. For each C_i^t, F and C_i^t play the following stage game:

1. C_i^t chooses to purchase ($a_i = 1$) or not purchase ($a_i = 0$) at fixed price w (observed by F).

2. F privately chooses effort level $e_i \in \{0, 1\}$.

3. Quality of service $s_i \in \{u_0, 0, 1\}$ is realised. C_i^t receives payoff $s_i - wa_i$, F receives payoff $a_i w - ce_i$, where $c > 0$.

4. a_i and s_i are observed by all $C_j \in N(C_i^t)$, and by F.

We refer to C_j observing s_i as C_j observing customer C_i's *review* of the firm. The firm serves simultaneous customers separately, so can choose different efforts for C_i^τ and C_j^τ, even though both customers are served simultaneously in period τ. The firm chooses each effort e_i to maximise their discounted sum of expected future payoffs, with discount factor $\delta \in [0, 1]$.

Firm Technology The quality of service customers received depends upon their purchase decision, firm effort choice and firm type. If customers do not purchase ($a_i = 0$), their quality of service is just their outside option payoff, which is $s_i = u_0$.

If customers do purchase ($a_i = 1$), they either receive good ($s_i = 1$) or bad ($s_i = 0$) service[3] with probabilities $P(s_i = 1|e_i, \theta) = p_{e_i\theta}$ and $P(s_i = 0|e_i, \theta) = 1 - p_{e_i\theta}$.

We make the following assumptions on firm technology: if $\theta = 0$, F is a 'low-skill' type who always provides bad service: so $p_{10} = p_{00} = 1$. Since effort is costly, it follows that if $\theta = 0$ the firm will always exert low effort. If $\theta = 1$, F is a 'high-skill' type with $p_{01} < p_{11} \leq 1$. As the firm's effort choice is only relevant if $\theta = 1$, going forward we will take e_i to mean the effort choice of F when serving C_i given $\theta = 1$.

We assume $0 < p_{01}$: good service is more likely from a high-skill form than low-skill firm regardless of firm effort choice. In other words, customers prefer a good chef putting in low effort to a bad chef. We make this assumption to guarantee that a reputation is valuable. If customers were indifferent between being served by a high-skill, low-effort firm and a low-skill firm, with a finitely lived firm we would be able to unravel any equilibria featuring high effort; final period customers would know they will receive low effort, and given this, would not care about the firm's type, meaning the firm would have no reputational incentives in the penultimate period, etc[4].

Under our assumptions, only a high-skill firm is capable of generating good service. This makes the customer inference problem simpler; observing one neighbour receive good service is conclusive evidence that the firm is high-skill (while observing only bad service is inconclusive since the firm could either be low-skill or an unlucky high-skill firm). We make this assumption for tractability, but our insights generalise easily to settings in which low-skill firms have a low probability of providing good service. In particular, results will be robust in the following sense; for almost

[3]Given how we define quality of service, s_i is a sufficient statistic for a_i; as such, the role of allowing customers to observe one another's purchase decisions is purely expositional. We also note that if customer order is commonly known, we do not actually need customers to *observe* non-purchases; in other words, observing $s_i|s_i \in \{0, 1\}$ is a sufficient statistic for s_i and a_i; non-purchases can be inferred from a lack of reviews.

[4]The important point here is not that the firm is finitely lived; the same argument can hold with an infinitely lived firm. The crucial point is that for reputational incentives to exist when incentives are for a strategic type to *distinguish* themselves from a bad type, rather than to *imitate* a good type, it must be that once the strategic type has successfully distinguished themselves, customers still care about their type.

all parameterisations, there exists some $\bar{\epsilon} > 0$ such that if a strategy profile can be supported in PBE for a given network-order pair, when $P(s_i = 1|\theta = 0) = 0$, it can also be supported in PBE when $P(s_i = 1|\theta = 0) = \epsilon$ for $\epsilon < \bar{\epsilon}$. Roughly, because our setting is discrete, for almost all parameterisations it is not crucial that good news is conclusive; it is sufficient that it is merely overwhelming evidence in favour of a high type. However, the model is much less tractable with general networks when $P(s_i = 1||\theta = 0)$ is bounded away from 0.

Equilibrium Concept Our solution concept is Perfect Bayesian Equilibrium (PBE). This requires that for any history of play, firm effort choices are a best response to customer strategies, customer purchase decisions are a best response to firm strategies given customer beliefs about firm type, and wherever possible customers derive their beliefs about firm type use Bayes's rule. As our setting is finite, the concept of PBE is well defined and we can apply standard textbook definitions. Below, we offer a formal definition of PBE our setting.

Define the history of the game at the beginning of period τ as $H_\tau = \{(a_i, s_i)|C_i \in \mathcal{C}^t\}_{t=1,\dots\tau-1}$, and define $\mathcal{H}_\tau$ as the set of all possible H_τs. An information set I_i^τ for customer C_i in period τ is $I_i^\tau = \{(a_j, s_j)|(a_j, s_j) \in H_\tau, C_j \in N(C_i)\}$, for some history H_τ; define $\mathcal{I}_i^t$ as the set of all possible I_i^τs. A behavioural strategy for each customer $\rho_i : \mathcal{I}_i^t \to [0, 1]$ assigns a probability of purchasing (setting $a_i = 1$) to each possible information set for customer C_i at time t. A behavioural strategy for the firm $\sigma : H_t \times \mathcal{C}^t \to [0, 1]^{|\mathcal{C}^t|}$ assigns an effort choice[5] to each customer who visits in period t, for each history of the game at the start of period t. Beliefs for customer C_i in period t at information set $I_i^t \in \mathcal{I}_i^t$ are given by $\mu_i^t : \mathcal{I}_i^t \to [0, 1]$, where $\mu_i^t(I_i^t) = P(\theta = 1|I_i^t, \bar{\sigma}_i)$, $\bar{\sigma}_i$ is C_i's conjecture about the firm's strategy σ, and μ_i^t is

[5]Without loss of generality, we assume a firm does not condition future effort choices on past effort choices. As customers do not observe effort choices, it is sufficient for the firm to condition future effort choices on past quality of service; if the firm shirks in period $t = 1$ and gives bad service to all customers, their problem in period $t = 2$ is identical to what their problem would have been if they had exerted high effort in period $t = 1$ but been unlucky and still achieved bad service.

computed via Bayes' rule at information set I_i^t wherever possible given conjecture $\bar{\sigma}_i$.

Formally, a Perfect Bayesian Equilibrium (henceforth equilibrium) is a strategy for each customer ρ_i $\forall i = 1, ..., N$ and a strategy for the firm σ, such that ρ_i maximises C_i's expected payoff at each information set I_i^t given conjecture $\bar{\sigma}$, and σ maximises the firm's expected discounted payoff at each history H_t given ρ_is, and $\forall i = 1, ..., N$, $\bar{\sigma}_i = \sigma$. We will focus on pure strategy equilibria.

We note that given our firm technology, with short-lived customers, off-path beliefs will not play a role in supporting outcomes in equilibria where all customers have a positive probability of purchasing from the firm (though they may play a role in supporting equilibria where some customers do not purchase). We are largely interested in equilibria with positive probabilities of trade between all customers and the firm; as such, we will typically omit discussion of off-path beliefs for the sake of brevity.

2.1 Modelling Assumptions

In the remainder of this section, we discuss some of our modelling assumptions.

Throughout, we interpret the network N as a social network, but its sole role in our analysis is to determine which customers observe one another. As such, the reader is free to treat the language of networks as purely a modelling tool capturing an underlying observation structure (for example, an organisational hierarchy, or a rule determining from how many previous periods a customer can observe reviews). Since we restrict attention to fixed endogenous orders, our analysis is also unchanged if the network is directed or undirected, in the sense that for C_i^t and C_j^{t+s}, because C_i^t moves first, it is unimportant whether C_i^t observes C_j^{t+s}, since C_i^t has no more actions left to take after period t.

While we assume that each customer visits the firm only once, it is possible to interpret distinct customers in our model as multiple instances of the same customer

acting myopically. For example, if $N(C_i^t) \cup C_i^t = N(C_j^{t+k}) \cup C_j^{t+k}$, then C_i^t and C_j^{t+k} share the same friends and we can interpret C_i^t and C_j^{t+k} as the same customer who visits both in period t and in period $t+k$. In particular, if $\mathcal{N}$ is a complete social network and $\succeq_\alpha$ is a total order on $\mathcal{C}$ (in other words, one customer visits in each period), interpreting the model as one myopic customer visiting the firm N times is equivalent to interpreting the model as N customers each visiting the firm once. Note the model *would* differ with repeat customers who were forward looking; relative to the myopic case, customers might choose to purchase from a firm not because they expected good service, but in order to gather information to inform future purchase decisions[6].

We consider a firm with two possible types; a strategic type which faces a meaningful effort choice problem, and a simple type who always provides bad service. Incentives arise because the strategic type would like to distinguish themselves from the simple type in the eyes of customers. More generally, we can separate reputational incentives in games into two types of incentive: incentives to distinguish oneself from a bad type, and incentives to imitate a good type (we can think of these incentives as analogues of incentives in separating and pooling equilibria in signalling games).

Many properties of incentives are general to both categories, but there are also important distinctions. For example, in infinite time horizon settings where customers observe the full history, incentives to distinguish oneself are inherently less persistent than incen- tives to imitate. Incentives to distinguish are not persistent because as time passes, firm type will become statistically identified on the equilibrium path for almost all histories, rendering firm effort choice irrelevant. In contrast, incentives to imitate can be very persistent, because if in equilibrium a strategic type takes exactly the same actions as some commitment (Stackelberg) type, customers

[6]This incentive for customers to gather information would be unlikely to be a significant concern if the number of total customers is large relative to the number of visits each customer makes. For example, with e.g. 50 customers each visiting twice, information gathering incentives on first visit are likely to be minimal if the social network is sufficiently connected.

do not learn more about the firm's type as time passes, so incentives are time invariant. Which form of incentives is more appropriate to focus on in modelling will depend upon the setting, but we note that empirically, businesses do differ in inherent quality level, many new businesses *do* fail when they acquire bad reputations, and many new businesses *do* succeed in building good reputations.

Throughout, we take the order of customers visits as exogenous from the firm's perspective. In some settings, this will be a reasonable assumption; for example, the order in which students take driving lessons is likely to be determined by the order of their birthdays. More generally, understanding equilibrium outcomes for a fixed order of customers is an important first step to understanding equilibria in settings in which customers or firms can influence the order of visits; for example, settings in which customers need to book restaurant tables in advance, or in which restaurants can specifically invite certain customers to their opening.

We are modelling the price of the firm as fixed for the whole game. An alternative approach to reputational incentives is to allow the firm to choose a price for each customer, or each period. Under this alternate specification, reputational incentives would be conveyed via a different channel. If pricing is fixed, incentives for effort are transmitted via the *extensive* margin; high effort may increase the probability that future customers are willing to buy from the firm at price w. If pricing is discretionary, incentives for effort are entirely at the *intensive* margin; the firm charges each customer their ex-ante willingness to pay, extracting all surplus, and all customers participate in equilibrium. Modelling prices as fixed is significantly more tractable and allows customers to draw inference from the purchase decisions of their neighbours. In many settings fixed pricing is also the more realistic assumption in the short run; a new restaurant who wants to build a reputation is unlikely to make minute adjustments to their advertised menu after serving each customer.

3 Strategies and Equilibria

In this section, we discuss how customer decisions and firm incentives depend on network structure. We first introduce customers' purchase problems and the firm's effort choice problem, before illustrating how network structure affects customer learning and firm incentives via three examples. We then state results for more general network settings.

3.1 Customer Strategy

Each customer will purchase at an information set if their willingness to pay at that information set minus price w is greater than their outside option u_0. Customer C_i, expecting $e_i = b$, purchases at information set I_t if:

$$u_0 \leq P(\theta = 1 | I_t, \bar{\sigma}_t) p_{b1} - w$$

Customers can learn both from the experiences of their neighbours and from the purchase decisions of their neighbours. If C_t believes that all of their earlier friends purchase with certainty, then C_t will only draw inference from the quality of service received by their neighbours, since the purchase decisions of earlier neighbours are uninformative about the neighbours' information sets. If C_t believes $P(a_{t-k} = 1) < 1$ for some $C_{t-k} \in N(C_t)$, C_t can also draw inference from the purchase decision of neighbour C_{t-k}; they can infer if $a_{t-k} = 1$ then C_{t-k} must have seen good news about the firm, while if $a_{t-k} = 0$, C_{t-k} can only have seen bad news. In our setting, because good news is conclusive, if a customer only sometimes purchases, when they do purchase they must be certain that the firm is a high type, since they can only have seen either conclusive good news or inconclusive bad news.

For equilibria featuring trade to exist at all, there are two important information sets at which a customer should be willing to purchase: a customer who has learned the firm to be a high-skill type must be willing to purchase expecting low effort (else

we would be able to unravel any candidate equilibrium featuring trade, since the final customer, knowing they would receive low effort, would never purchase, and hence the penultimate customer would never purchase, etc), and a customer who expects high effort must be willing to purchase from the firm under their prior. For incentives for effort to ever exist, it must also be the case that customers are not willing to purchase from a firm they know to be a low type for sure (else reputations do not matter). As such, for the remainder of the paper, we make the following assumption:

Assumption 1. *$w \in (-u_0, p_{01} - u_0]$ and $w \leq \pi pi_{11} - u_0$.*

We will define a *critical customer* as a customer in an equilibrium who purchases from the firm at some (but not all) of their information sets on the equilibrium path.

Definition 1. *A critical customer for a given strategy profile $(\boldsymbol{\rho}, \sigma)$ is a customer C_i such that ex-ante, $0 < P(a_i = 1|\boldsymbol{\rho}, \sigma) < 1$.*

3.2 Firm Strategy

The firm will exert high effort when serving customer C_i if doing so increases their discounted expected revenue by more than the cost of effort, c. Formally, when serving C_i^t, the firm's incentive compatibility (IC) constraint for high effort is given by:

$$c \leq \sum_{k=1}^{T-t} \delta^k w \sum_j = 1^M [P(a_j = 1|e_i = 1) - P(a_j = 1|e_i = 0)]$$

We can see a necessary condition for exerting high effort is that doing so increases the probability that future customers purchase. If the firm believes that all of C_i's neighbours will always purchase regardless, the firm has no incentive for effort when serving C_i and should set $e_i = 0$. In other words the firm will consider exerting high effort for C_i if and only if C_i is observed by a critical customer. In particular, any customers who are observed by no-one (for example, the final customer) will receive low effort in any equilibrium.

3.3 Examples

We now give three examples of networks and customer orders to illustrate the firm and customer problems in our model:

Example 1: Customers observe their predecessor In example 1, we have 4 customers who visit sequentially, each of whom observe their immediate predecessor (for example, we could think of a setting where all customers write reviews, and review websites only display a firm's most recent review).

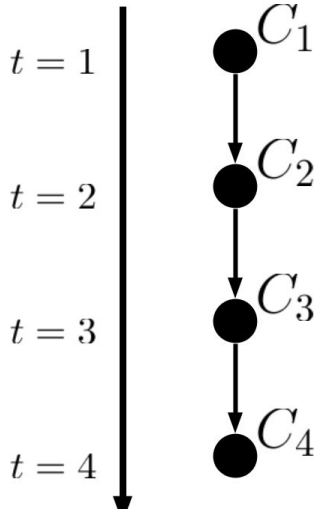

Figure 1: Example 1: Customers observe their predecessor

We will focus on the case where if the final customer observes a bad review, they are unwilling to purchase (if all customers who have seen a bad review still purchase, in the firm optimal equilibrium the firm would never exert effort and all customers would purchase regardless). Formally, assume $\frac{\pi(1-p_{01})p_{01}}{\pi(1-p_{01})+(1-\pi)} < w + u_0$.

To understand how incentives and learning operate in this setting, let's suppose that the firm exerts high effort for the first customer: $e_1 = 1$ (we know by assumption that if the firm plans high effort for C_1, C_1 is willing to purchase), and ask what this implies about equilibrium outcomes.

The first thing we can note is that if the firm exerts high effort for C_1, it must be that C_2 only purchases if they see a good review from C_1; if C_2's purchase decision

does not depend on C_1's review, the firm would be better off setting $e_1 = 0$ as effort is chosen after C_1 chooses whether or not to purchase.

The second point to note is that if C_2 only purchases when they see a good review ($s_1 = 1$), then C_3 can learn from C_2's purchase decision. If they see C_2 purchase they learn the firm is a high type for sure, whereas if they see C_2 does not purchase, they should revise their beliefs downwards. In other words, although C_3 does not observe C_1 directly, if C_2's purchase decision depends on s_1, then in equilibrium C_3 learns the value of s_1 anyway.

An important implication of C_3 learning via C_2's purchase decision is that the firm should not exert effort for C_2. To see this, note that the firm should only exert effort for C_2 if C_2's review influences C_3's decision. But we can note that C_2's review can never influence C_3's decision; C_2 only purchases if they see a good review, which implies the firm is a high type for sure. As a result, any time C_3 sees C_2 write a review, they can deduce that the firm is a high type because C_2 purchased, and so C_3 learns nothing extra from the review's content[7]. In other words, high effort for C_1 implies low effort for C_2.

Combining these insights, we can reason as follows; C_3 learns everything C_2 knows in equilibrium, so if C_3 also expects low effort from the firm, C_3's problem is identical to C_2's. As a result, if they expect low effort, they should always copy C_2's purchase decision (even if they expect high effort, they should copy C_2 if $\frac{\pi(1-p_{11})p_{11}}{\pi(1-p_{11})+(1-\pi)} < w$). Further, if C_3 copies C_2, C_4 should copy C_3 (note C_4 *knows* they will receive low effort, because they are observed by no-one).

What this tells us is that in example 1, we can find an equilibrium where if C_1 receives high effort, all remaining customers purchase only if C_1 writes a good review, and the firm exerts high effort for C_1 only. Formally, this equilibrium exists

[7]This would not be true if good reviews did not fully reveal the firm to be a high type, but an analogous point holds in more general settings; if the amount C_3 can learn from C_2's review is small compared to what they learn from C_2's purchase decision, the firm will have minimial incentives for effort for C_2.

if $c < (p_{11} - p_{01})w\sum_{i=1}^{3}\delta^i$ and $\frac{\pi(1-p_{11})p_{01}}{\pi p_{11}+(1-\pi)} < w + u_0 < \pi p_{11}$, and is unique[8] if $\frac{\pi(1-p_{11})p_{11}}{\pi(1-p_{11})+(1-\pi)} < w + u_0$.

This equilibrium is quite striking. When customers observe only their immediate predecessor, there is an equilibrium in which the firm exerts high effort at most once and all remaining customers base their purchase decisions solely on the review of the first customer. We can see this equilibrium is bad for customers in two ways; firstly, each customer only bases their decision on one review (so with probability $\pi(1 - p_{11})$ the firm is high-skill but customers fail to buy), and secondly, only one customer receives high effort when the firm is high-skill.

Example 1 illustrates an important point in general settings; herding is possible in equilibrium and is harmful for firm incentives, because if all remaining customers plan to copy their predecessor's purchase decision, the firm can no longer influence purchase decisions, and so should exert low effort for all remaining customers. We discuss how the possibility of herding generalises in Proposition 1.

Example 2: Two reviewers, one reader In example 2, we have two 'reviewers', each of whom observes no-one, and one 'reader', who observes both reviewers and visits the firm after them. Customer C_t visits in period C_τ.

Again, we will focus on the interesting case where the 'reader' bases their decision on the contents of the reviews. Formally, assume $\frac{\pi(1-p_{01})^2p_{01}}{\pi(1-p_{01})^2+(1-\pi)} < w+u_0$, to rule out an equilibrium in which the firm exerts no effort for any customer and all customers always purchase. Figure 3 illustrates how equilibrium possibilities depend upon c and π.

Because good reviews are fully revealing, if C_3 bases their decision on the contents of reviews, they should only purchase if they see at least one good review. Stepping back to the firm's effort choice for C_2, this tells us that the firm should only consider

[8]If this is not the case, then if $c < (p_{11} - p_{01})w\delta$ there is an additional equilibrium, in which C_2 and C_4 receive low effort and purchase only if they see a good review, while C_1 and C_3 receive high effort, and always purchase at any history.

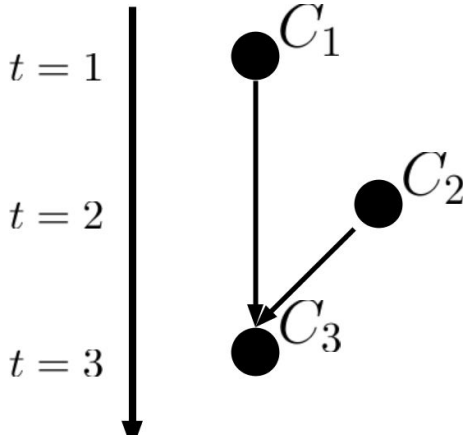

Figure 2: Example 2: Two reviewers, one reader

high effort for C_2 if $s_1 = 0$; if the firm gets a good review from C_1, C_3 will learn that the firm is a high type for sure, and so the firm has no more reputational incentives[9] for effort. If $s_1 = 0$, the firm should exert high effort for C_2 if $(p_{11} - p_{01})w\delta > c$.

Suppose the firm never exerts high effort for C_2. We can argue that in that case, they should also never exert high effort for C_1. On an intuitive level, when the firm serves C_2 after getting a bad review from C_1, the firm understands that it is their last chance to convince C_3 to purchase. In contrast, when serving C_1, their incentives for effort are weaker, as the firm knows they will have a second chance; if they get a bad review from C_1, that may not matter if they get a good review from C_2. In region C on figure 3, the firm sets $e_1 = 0$, $e_2 = 0$, and C_1 and C_2 always purchase. In region E, the firm would set $e_1 = 0$, $e_2 = 0$ if C_1 or C_2 purchased, and understanding this, C_1 and C_2 are too sceptical of firm type so do not purchase.

If the firm is willing to exert high effort for C_2, we can ask whether they should exert high effort for C_1? Their IC constraint for C_1 is now $(p_{11}-p_{01})(1-p_{11})w\delta^2 > c$. We can see that (even for $\delta = 1$), this is harder to satisfy than their IC constraint for C_2 conditional on $s_1 = 0$, because the firm has an incentive to 'wait-and-see' when

[9]With fully revealing good news, if all remaining customers have seen at least one good review, the firm no longer has incentives for effort, but the benefits to the firm of quitting high effort after early success arise more generally. If reviews were only boundedly informative about firm type, then the firm only has reputational incentives in histories where it is still possible to influence a customer's decision; if a customer needs to see four good reviews out of five and the firm has already received four good reviews, the firm no longer has reputational incentives related to that customer.

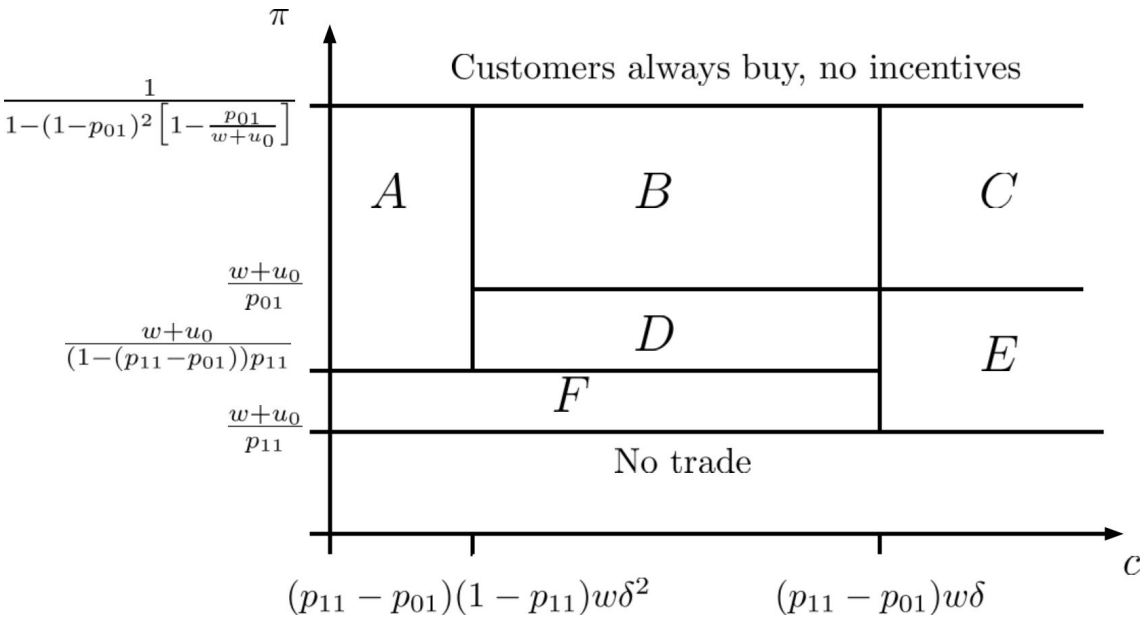

Figure 3: Example 2: Equilibrium Possibilities

serving C_1, knowing they will get a second chance at a good review when serving C_2.

In region A effort is sufficiently cheap that the firm finds the following strategy optimal; exert high effort for C_1, and exert high effort for C_2 iff $s_1 = 0$, and C_1 and C_2 have high enough priors to always purchase from the firm.

In contrast in region F, C_2's prior is too low to be willing to purchase if the firm plans high effort for C_1 (since C_2 knows they only get high effort if the firm fails to receive a good review from C_1), and so in equilibrium only one of the two customers can purchase from the firm. Effort remains cheap enough that the firm is willing to exert effort if only one customer visits, so in region F in equilibrium one of C_1 and C_2 always purchases and receives high effort and the other never purchases.

In region B the firm exerts high effort for C_2 iff $s_1 = 0$, and does not find it worthwhile to exert high effort for C_1 given they know they will have a second chance when serving C_2. C_1 and C_2 always purchase, and C_3 purchases if she sees a good review. In region D the unique equilibrium has C_2 always purchase and receive high effort, while C_1 does not purchase. C_1 understands that if they visit knowing C_2 also visits in equilibrium they will receive low effort, and their prior

is low enough that this deters purchasing, while C_2 is willing to visit regardless of whether C_1 visits or not, because even if C_1 does deviate to visit, the probability of $e_2 = 1$ remains sufficiently high.

If we consider expected payoffs, we can see that there are parameterisations where C_2's expected payoff in equilibrium is smaller when effort is cheaper. For example, in region A, with cheap effort C_1 and C_2 purchase and C_2 receive high effort with probability $\pi(1-p_{11})$, while in region B with an intermediate effort cost again C_1 and C_2 purchase, but now C_1 receives low effort and C_2 receives high effort with probability $\pi(1-p_{01})$. In other words, weakening firm incentives for C_1 by increasing c can be good for C_2, since from the firm's perspective, efforts for C_1 and C_2 are substitutes (similarly C_2 is better off with parameterisations in D than in B, so at the margins may also benefit from lower π).

If we consider relative payoffs, C_2's expected payoff is higher than C_1's for intermediate costs (regions B and D) but lower than C_1's for low costs (region A). If effort is very cheap, C_1 receives high effort whenever the firm is high-skill, whereas C_2 only receives high effort from a high-skill firm with probability $(1-p_{11})$. If effort is more expensive however, the incentives to 'wait-and-see' become too strong, and C_2 is better off than C_1; C_1 never receives high effort from a high-skill firm, while C_2 continues to receive high effort from a high-skill firm with probability $(1-p_{11})$.

We can see that if effort is sufficiently cheap ($c < p_{11} - p_{01})(1-p_{11})w\delta^2$), C_2 would be better off if they could move their visit to the firm to $t = 1$. As they do not observe anyone they derive no informational benefit from visiting the firm in $t = 2$, and when effort is cheap they are less likely to receive high effort than a $t = 1$ customer because the firm can condition $t = 2$ effort choices on $t = 1$ performance. Visiting in $t = 1$ removes the firm's ability to 'wait-and-see', and so for cheap effort, high-skill firms would exert high effort for both C_1 and C_2 if both visited at $t = 1$. This illustrates an important more general point; even for a fixed network, the timing of customer visits is important for expected payoffs; it is good

to be an early reviewer when effort is cheap, and a later reviewer when effort is expensive.

Our second example demonstrated how firm incentives for effort can vary according to the timing of reviewer visits. Our third example illustrates how firm incentives for effort depend on how many customers observe a review, and how well informed those customers are.

Example 3: Shared and Exclusive Readers In Example 3, we have 3 'reviewers' who visit at $t = 1$, and 3 'readers' who visit at $t = 2$. C_1 and C_2 are both observed by C_4 and C_5, and C_3 is observed by C_6.

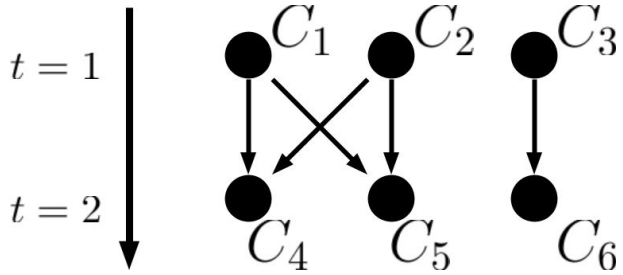

Figure 4: Example 3: Shared and Exclusive Readers

For all readers to base their decisions on the contents of reviews in equilibrium, we need $\frac{\pi(1-p_{01})p_{01}}{\pi(1-p_{01})+(1-\pi)} < w + u_0$, to rule out an equilibrium in which the firm exerts no effort for any customer and all customers always purchase. Note however that if $\frac{\pi(1-p_{01})p_{01}}{\pi(1-p_{01})+(1-\pi)} < w + u_0 < \frac{\pi(1-p_{01})^2 p_{01}}{\pi(1-p_{01})^2+(1-\pi)}$, it could be that in equilibrium C_6 always purchases and C_3 receives low effort, while C_4 and C_5 base their decisions on review contents. In other words, if customers have relatively high priors, it may be one bad review would not deter them from purchasing but two would..

We are interested in comparing customer learning from low-effort reviews to learning from high-effort reviews, so we will also assume for this example that customers are willing to purchase from the firm under their prior regardless of effort level ($w + u_0 < \pi p_{01}$).

If $\frac{\pi(1-p_{01})p_{01}}{\pi(1-p_{01})+(1-\pi)} < w + u_0$, the firm knows that 'readers' will only purchase if

they have seen at least one good review. The firm's effort choice problem for C_6 is simple: if $(p_{11} - p_{01})w\delta \geq c$, exert high effort, otherwise exert low effort.

Turning to C_1 and C_2, we can see that the firm's effort for one reviewer will affect their incentives for effort for the other. Note that this is not moral-hazard-in-teams; as the firm chooses e_1 and e_2 simultaneously to maximise a single objective, it can choose the co-operative solution and there is no hold-out problem. Suppose the firm plans high effort for C_1. Then the firm should exert high effort for C_2 if $(p_{11} - p_{01})(1 - p_{11})2w\delta \geq c$, and should set $e_2 = 0$ otherwise. The firm's effort choice problems for C_1 and C_2 are symmetric, so if $(p_{11} - p_{01})(1 - p_{11})2w\delta \geq c$, the firm should exert high effort for both.

If $(p_{11} - p_{01})(1 - p_{01})2w\delta \geq c > (p_{11} - p_{01})(1 - p_{11})2w\delta$, the firm only finds it optimal to exert high effort for one of C_1 and C_2. If $(p_{11} - p_{01})(1 - p_{01})2w\delta < c$, the firm finds high effort for neither customer optimal.

Comparing incentives for effort for C_1, C_2, and C_3, we can see that if $p_{11} < \frac{1}{2}$, then if C_3 receives high effort in equilibrium, C_1 and C_2 must also receive high effort in equilibrium. In contrast if $p_{11} \geq \frac{1}{2}$, it is possible in equilibrium that C_3 receives high effort, but either C_1 or C_2 does not. This illustrates an important point; the strength of firm incentives for a reviewer depends not only on how many later customers will read their review, but also on how many other reviews those later customers read. If a single high effort review is very informative ($p_{11} > \frac{1}{2}$), a firm may not find it worthwhile to exert high effort for multiple reviewers, since it is unlikely a second high effort review will change the mind of a customer who has already seen one high effort review. The intuition is particularly clear in the extreme case as $p_{11} \to 1$; if high-effort guarantees a good review, then the firm should never exert high effort for both C_1 and C_2, since exerting high-effort once is sufficient to guarantee C_4 and C_5 observe a good review.

From the perspective of customer expected payoffs, we can say the following: if effort is very cheap then all $t = 1$ customers have the same expected payoff, and C_4

and C_5 have higher expected payoffs than C_6 as they observe more reviews. If the cost of effort is very large, then again all $t = 1$ customers have the same expected payoff as none receive high effort, and C_4 and C_5 have higher expected payoffs than C_6 as they observe more reviews. For intermediate costs of effort, however, if p_{11} is sufficiently large (high effort reviews are sufficiently informative), it may be that C_3 has a greater expected payoff than C_1 and C_2 and C_6 has a greater expected payoff than C_4 and C_5, if the firm chooses high effort for C_3 but not for C_1 or C_2. Note in particular that if one high effort review is more informative than two low effort reviews, for intermediate costs of effort and p_{11} sufficiently large, C_4 would be better off ex-ante if they could commit to not reading C_2's review, since this would strengthen firm incentives when serving C_1 and improve C_4's information quality.

Example 3 illustrates an important general point; which customers are best off for a given order and given network depends in a non-trivial way on the parameters of the firm's moral hazard problem. When the firm finds effort cheap and reviews are relatively uninformative, in general customers benefit from being more connected. In contrast, if the firm finds effort more expensive and reviews are relatively informative, uninformed customers (reviewers) may be better off being observed by a few 'captive' readers than many shared readers, and readers may be better off committing to reading only a few reviews from reviewers who receive high effort. We return to this further in section 4 when we discuss efficient networks.

3.4 General Networks

We now turn to state three results for general networks. Firstly, we show that the 'herding' behaviour in Example 1 is possible in much more general network settings. Secondly, we give conditions for equilibria to be unique, but argue in general we should expect multiple equilibria for a given network and customer order. Finally, we compare networks in which customers observe only the previous customer to networks in which customers observe the full history, and show that if we select

the equilibria which feature the most high effort, total effort can be higher when customers only observe the previous customer than when customers observe the full history.

Herding: Example 1 (where customers observe only their predecessor and visit sequentially) showed that *herding* is possible in equilibrium, with all customers after some date copying the purchase decision of their predecessor regardless of reviews. We can extend this insight to *any* network where customers visit sequentially and all customers are connected by a path on the network to the first customer:

Proposition 1. *If one customer C_t visits in each period t, $c \leq (p_{11}-p_{01})w\sum_{i=1}^{N-1}\delta^i$, $\frac{\pi p_{11}p_{01}}{\pi p_{11}+(1-\pi)} < w + u_0 < \pi p_{11}$, and all C_τ for $\tau \geq 2$ are connected to C_1 via a path $\langle C_1, C_{j_1}\rangle, \langle C_{j_1}, C_{j_2}\rangle ..., \langle C_{j_n}, C_\tau\rangle$ (for $j_n < \tau$ increasing in n), there exists an equilibrium in which $e_1 = 1$, C_1 always purchases, and all remaining customers C_τ purchase iff $s_1 = 1$, with $e_\tau = 0$ for $\tau \geq 0$.*

Note that any network where all customers observe the full history of reviews satisfies the requirement that C_τ is connected to C_1 via a path through customers who visit between $t = 1$ and $t = \tau$. This requirement is also satisfied by networks which are directed rooted trees (formally, networks satisfying, for $t > 1$, $|\{C_{t-k}|C_{t-k} \in N(C_t), k > 0\}| = 1$).

This result gives a sufficient condition for there to exist an equilibrium in which all customers base their purchase decisions on the review of the first customer. We can note that if C_2 bases their purchase decision on C_1's review, and all remaining customers are connected via a directed path to C_2, we can also restate the result as giving a sufficient condition for all remaining customers to copy C_2's purchase decision. This gives us the following corollary:

Corollary 1. *Suppose one customer C_t visits in each period t, and all customers C_k who visit after τ are connected to C_τ by a path $\langle C_\tau, C_{j_1}\rangle, \langle C_{j_1}, C_{j_2}\rangle ..., \langle C_{j_n}, C_k\rangle$ (for $j_n < k$ increasing in n), and are not connected to any C_j for $j < \tau$. Then if C_τ*

is a critical customer, there exists an equilibrium in which all remaining customers copy C_τ's purchase decision.

Our corollary highlights that even on networks where not all customers are connected via a path to the first customer, we can identify subgraphs of a social network on which herding *is* possible.

Our corollary highlights that even on networks where not all customers are connected via a path to the first customer, we can identify subgraphs of a social network on which herding *is* possible. We can also note that it may provide some insight into which links on a network are relevant in equilibrium; in particular, we can see that if C_k copies C_τ's purchase decision in equilibrium, then C_k learns no payoff-relevant information from any links to earlier customers beyond links on one path between them and C_τ, and so deleting those links should not affect equilibrium strategies.

Our results on the possibility of herding highlight how much the strength of reputational incentives can differ when incentives are at the *extensive* rather than *intensive* margin. In our model, if each customer only observes the previous customer, two consecutive customers cannot both receive high effort, because when purchase decisions are informative about firm type, this removes incentives for consecutive effort. In contrast, if a firm could set prices, they would always charge customers their willingness to pay, and customers would always purchase. As a result, while with a fixed price firm cannot exert high effort for two consecutive customers when customers only observe the previous period, it is possible that a *price setting* firm could exert high effort for all but the final customer, because for a price setting firm, customer purchase decisions are not informative about firm type, and all learning is from *reviews*.

Uniqueness: Multiple equilibria arise in general in our setting because customer willingness to pay depends upon whether customers believe a firm has been exerting high effort (and whether they expect the firm to continue to exert high effort).

Customers judge a firm which receives bad reviews more harshly if they believe it has been trying, but are willing to pay more if they think they will receive high effort if the firm is high skill. For equilibria to be unique, we must remove the possibility of self-fulfilling prophecies, where the firm's equilibrium effort choices depend upon customers' conjectures about firm effort.

We can define the following two thresholds for price w:

$$\bar{t} = \max\{t | t \in \mathcal{N}, \frac{(1-p_{01})^t \pi}{(1-p_{01})^t \pi + (1-\pi)} p_{11} \geq w + u_0\}$$
$$\underline{t} = \max\{t | t \in \mathcal{N}, \frac{(1-p_{11})^t \pi}{(1-p_{11})^t \pi + (1-\pi)} p_{01} \geq w + u_0\}$$

$\underline{t}$ is the largest degree a customer can have for which we can guarantee that they always purchase in any equilibrium for any order of customers; $\bar{t}$ is the largest degree a customer can have such that for any order of customers it is possible they always purchase in *some* equilibrium[10]. We can use these thresholds to give the following result:

Proposition 2. *If $\bar{t} = \underline{t}$, and each customer C_t visits in period t, actions on the equilibrium path in a pure strategy PBE are unique for almost all parameterisations.*

If $\bar{t} = \underline{t}$ the initial set of customers who purchase only if they have observed a good signal does not depend on customer conjectures about firm effort choice. Hence we can eliminate self-fulfilling prophecies in equilibria, since if customer strategies are independent from conjectures about firm type, firm strategies should also be independent from customer conjectures.

Further, if we only have one customer per period, given our discrete setting, *firm* optimal strategies will generically be unique for fixed customer strategies (we can reason that with all effort choices taken sequentially, for almost all parameterisa-

[10]Note if the highest degree d_{max} on network $\mathcal{N}$ satisfies $d_{max} \leq \underline{t}$, no equilibrium features high effort; if there is no customer on the network who can ever observe enough signals to be unwilling to purchase (for example, if π is very large or w is very small), then all customers always purchase and the firm will never exert high effort.

tions the firm will have a unique optimal action when serving the final customer at each history, and when serving the penultimate customer, etc.). Note we cannot guarantee equilibrium uniqueness in a general setting when customers visit the firm simultaneously, because for many parameterisations a firm may be indifferent between several strategies. For example, we can see that in Example 3 C_1 and C_2 are symmetric, so if there is an equilibrium in which the firm only exerts high effort for C_1 there is also an equilibrium in which the firm only exerts high effort for C_2.

Note that $\bar{t} = \underline{t}$ may not in general be an appealing requirement for a parameterisation; one interpretation of the proposition would be that equilibria are generically unique when firm effort choice makes a negligible difference ($p_{01} \approx p_{11}$), which is not the most profound insight. For most interesting parameterisations of our model, multiple equilibria are likely to be possible. However, it worth noting that $\bar{t} = \underline{t} = 0$ is a potentially interesting special case in which firm effort choice may remain important; this is the case where no matter what effort choice they expect, customers are turned off the firm by one bad review. Here it may be firm effort choice is still very productive, but customers have sufficiently low priors they avoid firms with even one bad review, and so studying incentives for effort remains both important *and* tractable.

Persistent High Effort: Our results so far have demonstrated when it is possible in equilibrium for all customers beyond a certain date to copy their predecessors, and when we should expect to find unique equilibria. Given the pervasiveness of multiple equilibria in our model, without a strong equilibrium selection argument it is hard to make definitive predictions in general settings. However, we can provide insights into inference problems. In particular, we can suggest an answer to the following question: if we observe a firm receive persistent good reviews over time, what does this say about customer networks?

We will do so by comparing the maximum high effort that can be incentivised in

equilibrium for two commonly studied network specifications: the complete network (where all customers observe all their predecessors) and the line network (where each customer only observes their immediate predecessor).

When the social network is complete and customers visit sequentially, we will say that C_t visits in period t, and observes all the customers C_{t-1} observed[11]. As all customers observe the full history of reviews, it follows the firm will only exert high effort if it is yet to achieve a good review. We can reason that if the firm initially exerts high effort for customers in equilibrium, there are two reasons they might stop; either they have already received a good review (meaning all remaining customers know they are a high type) or customers have observed too many bad reviews and are unwilling to purchase.

It follows that we can support high effort (conditional on being yet to receive a good review) for all customers up to some cutoff date t^* as an equilibrium strategy for the firm if the following conditions hold:

$$\begin{aligned} w \geq & \frac{c(1-\delta)}{(1-p_{11})^{t^*-1}(p_{11}-p_{01})\delta^{t^*-1}(1-\delta^{T-t^*+1})} \\ w+u_0 \geq & \frac{(1-p_{11})^{N-1}\pi}{(1-p_{11})^{N-1}\pi+(1-\pi)}p_{01} \\ t^* = & \max_t\{t|t\in\mathcal{N}^+, t<N, \frac{(1-p_{11})^t\pi}{(1-p_{11})^t\pi+(1-\pi)}p_{11}\geq w+u_0\} \end{aligned}$$

As $c \to 0$, it follows that the maximum number of customers who can receive high effort on a complete network is pinned down by:

$$t^* = \max_t\{t|t\in\mathcal{N}^+, t<N, \frac{(1-p_{11})^t\pi}{(1-p_{11})^t\pi+(1-\pi)}p_{11}\geq w+u_0\}$$

When the social network is a line network and C_t visits in period t, such that

[11] A common approach to restricting customer observations in the reputations literature is to limit customers to visiting sequentially but observing only the outcomes of the past K periods for finite K. In our setting, we note that either the finite memory setting has an equivalent equilibrium to a complete network with sequential visits, if for a complete network there is a critical customer among the first K, or in equilibrium features no high effort.

$N(C_t) = \{C_{t-1}, C_{t+1}\}$ for $1 < t < N$, we know that there exist equilibria featuring herding, where the firm exerts high effort only for C_1 (as in Example 1). However, if $\frac{\pi(1-p_{11})p_{01}}{\pi(1-p_{11}+(1-\pi)} < w+u_0 < \frac{\pi(1-p_{11})p_{11}}{\pi(1-p_{11}+(1-\pi)}$, there also exist other equilibria. In particular, suppose N is even. Then for c sufficiently small, there exist equilibria in which the firm sets $e_i = 1$ and customer always sets $a_i = 1$ if i is odd, while if i is even the firm sets $e_i = 0$ and C_i sets $a_i = 1$ iff $s_{i-1} = 1$. In other words, odd customers always purchase and receive high effort, and even customers purchase if their immediate predecessor wrote a good review and receive low effort. This constitutes an equilibrium because if odd customers expect high effort they are happy to purchase at any information set and as such will ignore whether their predecessor purchases, while if even customers know they will receive low effort, they are only willing to purchase if they see a good review. As such, as $c \to 0$, if the total number of customers is even, the maximum number of customers who can receive high effort when customers only observe the previous period is $\frac{N}{2}$.

The following result is immediate:

Proposition 3. *If the number of customers $N > \bar{N}$ and c is sufficiently small, the maximum number of customers who can receive high effort in some equilibrium history is higher with a line network than a complete network, for:*

$$2\left\lfloor ln\left(\frac{(1-\pi)(w+u_0)}{\pi(p_{11}-w-u_0)}\right)\right\rfloor = \bar{N}$$

This result tells us that if the number of customers is sufficiently large, the maximum number of customers who can receive high effort in equilibrium is larger when customers observe only the previous period than when they observe the full history.

Note that this is the *maximum* number who receive high effort, not the expected number who receive good service. For the line network the expected number of customers who receive good service in equilibrium grows linearly with the maximum

number who receive high effort, but for the complete network, it does not. As a result, if we consider the expected number of customers receiving good service in equilibrium, the complete network compares even less favourably to the line network.

Note while we make a simple comparison between a line network and a complete network when effort is cheap, analogous results hold for higher effort costs[12] and observation structures where customers observe something between the full history and their immediate predecessor. The driving force behind this result is that when all customers can observe sufficiently many previous periods, then beyond some date (which is pinned down by the firm technology and *not* the total number of customers), customers will only be willing to purchase if there is a good review in the history (regardless of what effort the customers expect from the firm), and after that date the firm will have no more incentives for effort. In contrast, if customers only observe a limited subset of the firm's past reviews, it is possible in equilibrium for the firm to restart high effort even if it has already achieved one good review, if not all later customers have learned about that good review. As a result, expected firm effort can grow in proportion to the total number of customers.

Our result thus suggests we should draw the following inference:

Observation 1. *Firms who receive persistent good reviews over time serve networks with intermediate customer connectivity.*

4 Efficient Networks

In many settings we would like to understand which customer social networks are optimal from the perspective of various welfare measures. Social planners may sometimes have the ability to influence which social networks form: for example, a

[12]In particular, suppose for the line network, the firm will only exert high effort if it determines the next M customers' decisions. Then we can find equilibria in which C_1 receives high effort, C_2 to C_{M+1} herd (only purchasing if $s_1 = 1$), C_{M+2} always purchases and receives high effort, C_{M+3} to C_{2M+3} herd, and so on. The same insight obtains; for N sufficiently large, the highest effort possible in equilibrium will be higher for the line network than complete network.

university building a new campus may have discretion over the layout of halls of residence and the number of communal spaces available. Policy makers and platforms may also have access to interventions which change the visibility of online customer reviews; for example, a reviews platform could add an auto-translate feature for foreign language reviews. In this section, we begin by presenting some simplifying results about what outcomes a planner designing a customer network and order can implement in equilibrium. We then explore in greater depth the problem of a planner whose specific aim is to maximise total expected surplus. We also briefly discuss how consumer optimal networks relate to socially efficient networks.

We will assume that effort is ex-ante desirable from the perspective of total surplus (equivalently, we assume if effort were contractible high-skill firms would want to write contracts which induced high effort). Formally:

Assumption 2. $p_{11} - p_{01} > c$

This assumption best fits our motivations; exploring settings where reputational incentives can overcome moral hazard problems[13] in the absence of formal contracting.

We will continue to refer to a network $\mathcal{N}$ as a social network, but we reiterate that formally, $\mathcal{N}$ only represents an observation structure, and the reader is free to assign it different interpretations. When discussing a planner designing $\mathcal{N}$, for example, it may be appealing to interpret $\mathcal{N}$ as determining review visibility on an online platform.

We will consider the problem of a planner who before the first period publicly chooses a network-customer order pair, and suggests an equilibrium strategy profile to all players. We begin by providing some simplifying results which are independent of the planner's objective, before exploring the problem of a planner who seeks to

[13]Parametrisations with inefficient effort, where high effort plays the role of noisy inefficient Spencian signalling, are also of theoretical interest, but are arguably less compelling as applications. As such, we assume effort is ex-ante efficient outright, for the sake of minimising caveats in stating our results.

maximise total expected surplus.

4.1 Feasible Outcomes

In order to understand optimal networks from the perspective of a specific objective, it is useful to build some insight into what outcomes it is feasible for a planner to implement by choosing a network-order pair.

Bipartite Networks We begin with a significant simplifying result; with fully revealing good news, any equilibrium outcome that can be supported on a network $\mathcal{N}$ can also be supported on a bipartite network $\mathcal{N}_B$. Formally, a network $\mathcal{N}$ with node set $\mathcal{C}$ and edge set $\mathcal{E}$ is a bipartite network if there exists disjoint $\mathcal{C}_1$ and $\mathcal{C}_2$ such that every edge in $\mathcal{E}$ connects a node in $\mathcal{C}_1$ to a node in $\mathcal{C}_2$.

Proposition 4. *If $\mathcal{N}, \succeq_\alpha$ is a network-order pair such that equilibrium strategy profile $(\sigma, \boldsymbol{\rho})$ generates ex-ante distribution over histories $\Delta(H_t)$, there exists a bipartite network $\mathcal{N}_\mathcal{B}$ with node sets ${\mathcal{C}_\mathcal{B}}^1$ and ${\mathcal{C}_\mathcal{B}}^2$, such that for network-order pair $\mathcal{N}_\mathcal{B}, \succeq_\alpha$, we can find an equivalent equilibrium strategy profile $(\sigma_B, \boldsymbol{\rho}_B)$ which also generates ex-ante distribution over histories $\Delta(H_t)$, and in equilibrium all $C_i \in {\mathcal{C}_\mathcal{B}}^1$ always purchase on the equilibrium path.*

This proposition states that for any outcomes that can be supported in equilibrium with a non-bipartite network, we can find a bipartite network in which customers visit in the same order and the same outcomes can be supported in equilibrium. The implication for a social planner is that (independent of planner objective function) there is no loss to the planner from restricting their search to only considering bipartite networks.

The reason is as follows: firstly, we can note there is no benefit to allowing customers who 'always buy' to observe one another, as they purchase at every information set on the equilibrium path. Secondly, because good news is fully revealing, anything one critical customer could learn from observing another critical customer,

they could also learn by observing that critical customer's friends instead. In other words, if Bob is a critical customer and Alice observes only Bob, Alice takes exactly the same decisions in each on-path history as if she only observed Bob's friends but did not observe Bob, because any time she sees Bob purchase, she knows one of his friends wrote a good review, so whether Bob writes a good review himself is irrelevant. As a result, for any non-bipartite network in which critical customers learn from one another, we can design a new network, cutting out the middle man, in which they take exactly the same decisions at every on-path history but each critical customer observes only 'always buys' (note that this may involve *adding* links to the non-bipartite network as well as removing them).

It is worth emphasising that our assumption that $p_{10} = p_{00} = 0$ is crucial for this result; if good news is not fully revealing, a planner would not be able to restrict attention to only bipartite networks, because it would still be possible for customers to learn from one another after the first piece of good news was observed. In other words, 'sometimes buys' could learn from other 'sometimes buys' if good news were not fully revealing.

When choosing network-order pairs, planners face a trade-off between inducing high effort from the firm and providing information to later customers. How the planner values the trade-off will depend upon their specific objective function, but regardless of their objective it is useful to understand what the planner can achieve at either extreme. In other words, if we think of the planner as implicitly choosing a payoff vector from a set of feasible equilibrium payoffs, it is useful to characterise the extreme points of that set. As such, it is illuminating to ask; what should a planner who wants to induce as much high effort as possible do? What should a planner who wants to minimise the probability of customer 'mistakes' (buying from bad firms or failing to buy from good firms) do?

Maximal High Effort Example 3 illustrated that if p_{11} is sufficiently large (high effort reviews are sufficiently informative) then it may be easier to incentivise high effort for always buys when later customers observe fewer reviews. This suggests the following question: if customers visit at $t = 1$ and $t = 2$, which networks maximise the number of customers who can visit at $t = 1$ and receive high effort in equilibrium?

To answer this question, it is useful to state the following lemma on when customer incentives are strongest:

Lemma 1. *If all customers visit in $t = 1$ and $t = 2$, and all $t = 1$ customers receive high effort, aggregate expected benefit from effort for firms is maximised if $t = 2$ customers are critical and have degree m^*, where*

$$m^* =\in \left\{ \lfloor \frac{-1}{\ln(1-p_{11})} \rfloor, \lceil \frac{-1}{\ln(1-p_{11})} \rceil \right\}$$

We can reason that roughly, the planner will be able to induce the most high effort by designing the network so that each $t = 2$ customer makes the largest possible contribution to aggregate incentives (since the more each $t = 2$ customer contributes to incentives, the fewer $t = 2$ customers are necessary to satisfy $t = 1$ IC constraints). The planner faces the following trade off; $t = 2$ customers with low degrees provide strong incentives to a few $t = 1$ customers, while $t = 2$ customers with high degrees provide weak incentives for many $t = 1$ customers. If a $t = 2$ customer has degree 1, they appear in 1 IC constraint at $t = 1$, while if they have degree 2 they appear in 2 IC constraints at $t = 2$, etc. Fixing the number of $t = 1$ and $t = 2$ customers and maximising aggregate incentives with respect to degree of $t = 2$ customers gives our lemma.

Note that in general may not be feasible to give $t = 2$ customers degree m^* in an efficient network-order pair. For example, it may be that $m^* > n_1$, where n_1 is the number of $t = 1$ customers, or it may be that $t = 2$ customers with degree

m^* cannot be symmetrically and evenly allocated between $t = 1$ customers without divisibility issues. In settings in which $\pi p_{01} - w \geq u_0$, it might also be the case that $t = 2$ customers with degree m^* are not critical customers.

We can use our lemma to give an upper bound on the number of customers who can receive high effort in equilibrium. We do so using the following argument: if, using the strongest possible incentives (which may not be feasible due to divisibility issues), we would need more than a $t = 2$ customers for each $t = 1$ customer, then on any *feasible* network $\frac{n_1}{n_2} < \frac{1}{a}$:

Proposition 5. *In a bipartite network featuring n_1 customers visiting at $t = 1$ and receiving high effort, and n_2 customers visiting at $t = 2$, we must have $\frac{n_1}{n_2} \leq \frac{1}{a}$, where a is given by:*

$$a = -\frac{[ln(1 - p_{11})]c}{(p_{11} - p_{01})\delta w}(1 - p_{11})^{\frac{1+ln(1-p_{11})}{ln(1-p_{11})}}$$

We bound by looking at the number of units of strongest possible incentives we would need to incentivise high effort and saying we need at least this many units of weaker incentives. The logic here is the following; suppose that using the strongest incentives conceivable (which may not be feasible), we need 3.5 $t = 2$ customers for each $t = 1$ customer to satisfy aggregate IC. Then for any feasible network (featuring weakly weaker incentives), we must have at least 3.5 $t = 2$ customers for each $t = 1$ customer[14].

Minimal Customer Mistakes A planner who wants to minimise customer 'mistakes' can achieve arbitrarily few mistakes with a large enough number of customers if customers are always willing to purchase under their priors. In particular, the planner can do the following; allocate n_1 customers to $t = 1$ and n_2 customers to $t = 2$, and allow all n_2 customers to observe all n_1 customers. As $N = n_1 + n_2 \to \infty$,

[14]Note it does not follow that we should have 4 $t = 2$ customers for each $t = 1$ customers; for example, we may be able to do better by starting with 3 $t = 2$ customers for each $t = 1$ and adding some additional $t = 2$ customers who observes all $t = 1$ customers to 'top up' incentives until we satisfy aggregate IC.

if the planner lets $n_1 \to \infty$ and $n_2 \to \infty$ but $\frac{n_1}{n_2} \to 0$, all $t = 2$ customers will have arbitrarily accurate posteriors and the proportion of customers who only purchase when $\theta = 1$ will go to 1 (this insight is from Acemoglu et al. (2011), who state it in a social learning setting). Note that since $t = 2$ customers have arbitrarily accurate posteriors, the firm would never find high effort optimal at $t = 1$, since the expected change in revenue as a result of high effort becomes vanishingly small as $n_1 \to \infty$.

The problem of minimising customer mistakes is more interesting with a small number of customers when the planner considers choosing network-order pairs which *do* induce some high effort. Then the planner must contend with an additional effect; reviews from customers who receive high effort are more informative than reviews from customers who receive low effort. As a result, when the total number of customers is small, purely from the perspective of learning the planner may prefer to design a network in which $t = 2$ customers do not observe all $t = 1$ customers (as illustrated in example 3). Given the discrete nature of our model, it is hard to give a clean characterisation of what a planner who wants to minimise customer 'mistakes' should choose for a given number of customers, but if $p_{11} - p_{01}$ is sufficiently large (and c sufficiently small) we would expect that with a small number of customers, the planner would do best choosing a network such that $t = 2$ customers read as many reviews as they can, subject to the constraint that the firm is 'just' incentivised for high effort for some $t = 1$ customers.

4.2 Socially Optimal Networks

We now explore in more depth the problem of a planner who designs a network-order pair to maximise expected social surplus. Formally, we define total ex-ante expected surplus S for a given network-order pair $(\mathcal{N}, \succeq_\alpha)$ and strategy profile $(\boldsymbol{\rho}, \sigma)$ as:

$$S = \sum_{t=1}^{T} \delta^{t-1} \Bigg[\sum_{C_i^t \in \mathcal{C}^t} E(a_i(s_i - ce_i) + (1 - a_i)u_0 | \boldsymbol{\rho}, \sigma, \mathcal{N}, \succeq_\alpha) \Bigg]$$

In other words, it is the discounted sum of consumer and producer surplus. Note that total expected welfare S is *not* a function of the price w, as purchases are transfers between customers and firms, and so cancel out.

We say that a network-order pair $(\mathcal{N}^*, \succeq_\alpha^*)$ is *efficient* if there exists a strategy profile $(\boldsymbol{\rho}^*, \sigma^*)$ such that $(\boldsymbol{\rho}, \sigma)$ constitutes a PBE given $(\mathcal{N}^*, \succeq_\alpha^*)$, and $(\boldsymbol{\rho}^*, \sigma^*, \mathcal{N}^*, \succeq_\alpha^*) \in \text{argmax}_{\boldsymbol{\rho},\sigma,\mathcal{N},\succeq_\alpha} S$. Note that efficient network-order pairs will not be unique in our setting; networks may have links that are irrelevant for equilibrium outcomes for a given order (say, if they link two customers who visit simultaneously), and as all customers in our model are ex-ante identical we can always swap C_i and C_j in $\mathcal{N}^*, \succeq_\alpha^*$ and the new network-order pair will also be efficient. Note also that as our setting can feature multiple equilibria (because customer willingness to pay depends on conjectures about firm effort), an efficient network-order pair $(\mathcal{N}^*, \succeq_\alpha^*)$ may also have other PBE strategy profiles $(\boldsymbol{\rho}', \sigma')$ such that $(\boldsymbol{\rho}', \sigma', \mathcal{N}^*, \succeq_\alpha^*) \notin \text{argmax}_{\boldsymbol{\rho},\sigma,\mathcal{N},\succeq_\alpha} S$. If a network-order pair $(\mathcal{N}^*, \succeq_\alpha^*)$ is *efficient*, we will refer to it as solving the social planner's problem.

What outcomes a social planner would like to induce will depend upon customer outside options. If $u_0 \leq 0$ trade between a low-skill firm and a customer is ex-ante desirable for total surplus (eating at a bad restaurant is still better than cooking at home), even if customers would prefer to avoid low-skill firms (because from their perspective, low skill firms are overpriced). If $u_0 > 0$ trade between a low-skill firm and a customer is ex-ante undesirable, and so a planner would like customers to avoid purchasing from low-skill firms.

The decision that customers would take under their prior also plays an important role in shaping the planner's problem. If $\pi p_{01} - w \geq u_0 > -w$, uninformed customers (inheriting prior) are willing to purchase regardless of their conjectures about firm effort. Hence if no customers learn anything about the firm (for example, if the social network is empty), all customers will purchase. Customers will only stop purchasing from the firm if they have seen a sufficient amount of bad news and revised their

beliefs down relative to their prior. The planner can thus guarantee surplus of $S_0 = N\pi p_{01}$ by allocating all customers to $t = 1$ and choosing any network.

In contrast if $\pi p_{01} - w < u_0 \leq \pi p_{11} - w$, uninformed customers will purchase only if they expect high effort. Hence for any trade to occur a planner will need to incentivise high effort for a first wave of customers, from whom later customers can learn about firm type. The planner can guarantee surplus of $S_\emptyset = Nu_0$ by allocating all customers to $t = 1$ and choosing any network.

We can simplify the customer orders a planner needs to consider for optimal network-order pairs.

Proposition 6. *If $\delta < 1$ and the optimal network $\mathcal{N}^*$ is a symmetric bipartite network, then any optimal order $\succeq^*$ allocates all customers to periods $t = 1$, $t = 2$, or $t = 3$. If $\pi p_{01} - w < u_0$, any optimal order $\succeq^*$ allocates all customers to periods $t = 1$ or $t = 2$.*

This proposition states that if the planner discounts the future, and links between 'sometimes buys' and 'always buys' are symmetrically allocated, the planner will allocate customers to at most the first three periods. If customers are only willing to purchase expecting high effort under their priors, the planner will allocate all customers to at most the first two periods[15]. The role of assuming $\delta < 1$ here is simply to break ties; if $\delta = 1$ then we will have many optimal orders since the planner is happy to leave arbitrary gaps between customers (provided during the gaps customers learn nothing new).

On an intuitive level, we can think of this result as saying we can divide the optimal equilibrium path into up to three phases: reputation decay for the firm, high effort for the firm, and informed decisions for customers. This result claims that each of these phases can be achieved in one period if the network is a symmetric bipartite network.

[15]Note there is no obligation on the planner to use more than *one* period: if the planner wants to induce full trade or no trade equilibria, with $\delta < 1$ they will do best allocating all customers to $t = 1$.

To sketch the argument, firstly note that for an impatient planner, it is optimal for all critical customers visit at once. Proposition 4 tells us there is no benefit to critical customers observing one another, so with $\delta < 1$ the planner will do best when they visit simultaneously.

Secondly, note that expected total effort is higher when 'always buys' who receive high effort at some histories visit simultaneously for a symmetric network. If 'always buys' visit sequentially the firm can quit high effort for later 'always buys' once all their neighbours have already observed one good review [16]. Moving later 'always buys' earlier to visit at the same time as the first customer who sometimes receives high effort increases expected effort for the moved customers while leaving expected effort for other customers unchanged. In a symmetric setting[17], removing the firm's ability to wait-and-see with effort choices can only strengthen aggregate incentives.

Finally, we know from example 2 that it may sometimes not be possible to incentivise high effort unless the firm has already received sufficiently many bad reviews, if customers have high priors. As a result, for some parameterisations the optimal order may involve an initial phase of reputation decay, because this is the only way to incentivise effort at a later date. If the planner has $\delta < 1$ the planner will do best allocating all of the 'always buys' who receive low effort to a single period.

Our results so far tell us a planner can restrict their search to bipartite networks, and which orders a planner should consider for a *symmetric* bipartite network. We now give conditions for when a *complete* bipartite network is optimal, and argue

[16]Note if the planner maximised firm profits (or if effort were ex-ante inefficient and had purely signalling value) this would be a beneficial effect; with inefficient high effort the planner would like to let the firm wait-and-see specifically to allow them to quit effort after the first success.

[17]Asymmetric settings are more complicated; it may be that if C_j has a lower degree than all the other 'always buys', they cannot receive high effort if they visit at the same time as other 'always buys', but can receive high effort in some histories if they visit later. In practice, optimal networks are likely to be approximately symmetric (the planner has no intrinsic reason to provide a customer with stronger incentives than necessary, and for critical customers there is diminishing marginal value to observing an extra review, both of which point the planner towards symmetry), but given the discreteness of our setting, we cannot rule out asymmetry of efficient networks due to divisibility issues.

that if the number of customers is sufficiently large and high effort is sufficiently productive, the planner will not want to choose a complete bipartite network.

Optimality of Complete Bipartite Network To help understand when a planner would like to use a complete bipartite network, we can ask: when are learning and incentives are complements for a planner? In other words, when does a more connected network improve both customer decision making and firm incentives for effort?

Recall in lemma 1 we found:

$$m^* = \left\{ \lfloor \frac{-1}{\ln(1-p_{11})} \rfloor, \lceil \frac{-1}{\ln(1-p_{11})} \rceil \right\}$$

where m^* is the degree for $t = 2$ customers which maximises their contribution to aggregate incentives for $t = 1$ customers. Understanding when incentives are strongest, we can note the following; if we fix the node sets of a bipartite network, and all critical customers currently have a degree below m^*, increasing all of their degrees by one should strengthen aggregate incentives *and* improve customer learning. Hence the following result holds:

Proposition 7. *In the class of bipartite networks with node sets* $\mathcal{C_B}^1$ *and* $\mathcal{C_B}^2$ *and* $|\mathcal{C_B}^1| = n_1$, $|\mathcal{C_B}^2| = n_2$, *where all members of* $\mathcal{C_B}^1$ *visit the firm at* $t = 1$, *the complete bipartite network is optimal if* $n_1 \leq m^*$, $u_0 > 0$, *and* c *is sufficiently small.*

In other words, if the highest possible number of 'always buys' it is possible for a critical customer to observe is below m^*, and a social planner only wants customers to purchase from high-skill firms, it is socially optimal to let all critical customers observe all 'always buys'.

In contrast if $n_1 > m^*$, the planner may face a trade off between learning and incentives; other things equal, a (uniformly) more connected network provides weaker incentives even if it may provide better learning. The driving force behind this trade

off is simple; as customers have access to more information, the chance that any one review changes their decision goes to zero. Hence, if customers read too many reviews each, a firm may not find high effort worthwhile, reasoning that customers are likely to learn the firm's type regardless of firm effort.

Note that the condition $n_1 \leq m^*$ may be very hard to satisfy; for example, if $p_{11} \approx 0.6$, we have $m^* = 2$, so with more than two customers who always buy, increasing the number of reviews later customers can read harms firm incentives. We can also note $n_1 \leq m^*$ is not a *necessary* condition for a complete bipartite network to be optimal; if we fix the two node sets of a bipartite network, the planner may prefer to use the weakest incentives that still motivate a firm to effort. If we have customers visiting at $t = 1$ and $t = 2$ and the planner wants customers to learn the firm's type, the planner can improve social welfare by adding links between $t = 1$ and $t = 2$ customers provided this does not affect which customers a firm wants to exert high effort for.

Observing that if we increase the number of customers beyond a certain point, increasing connectivity further will weaken firm incentives, we can say:

Proposition 8. *There exists some n' and some $\beta \geq 0$ such that if $N > n'$ and $p_{11} - c \geq \beta$, if high effort is possible in equilibrium for some network-order pair, no complete bipartite network is an efficient network.*

This result constitutes two claims: firstly, if high effort is sufficiently productive, as N becomes large a social planner will want to incentivise high effort for more than one customer; secondly, as N becomes large, it will become impossible to incentivise high effort for more than one customer in equilibrium.

This is true because incentives for firm effort are strongest when the watching customers have finite degree. For example, consider the case where $p_{11} = 1$; then it is impossible for more than one customer to receive high effort on a complete bipartite network in which customers visit at $t = 1$ or $t = 2$. However, for n sufficiently large, it will be possible to incentivise high effort for more than one customer on a disjoint

network, if each 'always buy' customer has enough exclusive followers. Where a planner places significantly more weight on incentives than customer learning, non-complete social networks may prove optimal.

A corollary with important policy implications is:

Corollary 2. *Efficient networks may feature disjoint components.*

This tells us that interventions such as auto-translating foreign language reviews may be harmful for social welfare, if the result is that they connect two previously disjoint cliques on which a firm previously had to build a reputation separately[18].

Customer Optimal Networks: We have focused on socially optimal networks, but can argue that, for many parameterisations, consumer optimal networks will be qualitatively very similar to socially optimal networks. In particular, when $(p_{11} - p_{01})w > c$ and $u_0 > 0$, both total surplus and consumer surplus satisfy the following properties: they are linear and increasing in e_i and $P(a_i|\theta = 1)$, and linear and decreasing in $P(a_i|\theta = 0)$, for all C_i. In other words, when effort is ex-ante efficient and customers' outside options generate more surplus than purchasing from a bad firm, both total and consumer surplus agree high effort is good, purchasing from high-skill firms is good, and purchasing from low-skill firms is bad.

The exact rate at which a planner values the trade-off between firm effort and customer learning will depend on their specific objective; for example, we know the objective of a planner who cares about total surplus does not depend on prices, while a planner who cares about consumer surplus values customers avoiding bad firms more when prices are higher. As such, for a given parameterisation, the socially optimal network is unlikely to be exactly the customer optimal network.

[18]Note however that because the planner is risk neutral, there is no inherent power in disjoint networks over connected networks with similar properties. For example, if we divide customers into equal numbers of 'always buys' and critical customers, any 2-regular bipartite network will give the same expected surplus, whether it is a circle network or divides the network into disjoint cliques of 4 customers each. As such, our insight is that disjoint networks may be socially preferable to complete networks, *not* that disjoint networks are *uniquely* socially optimal.

Nonetheless, we can note the following; when $(p_{11} - p_{01})w > c$ and $u_0 > 0$, consumer optimal networks will be *qualitatively* similar to socially optimal networks (if not quantitatively identical), because consumer and social surplus make qualitatively similar trade-offs between firm effort and learning. As such, we can develop an understanding of the properties of consumer optimal networks by studying socially optimal networks; there is unlikely to be significant benefit to a formal analysis of customer optimal network-order pairs on top of an analysis of socially optimal network-order pairs[19].

Firm Optimal Networks: When $w < \pi p_{01}$, the question of firm optimal networks is simple; if customers are always willing to purchase from an unknown firm under their prior, regardless of firm type the firm is best off with a network order-pair which induces no learning and no incentives for effort (for example, any network-order pair where all customers visit at $t = 1$). In contrast, if $w > \pi p_{01}$, uninformed customers must expect to receive high effort to be willing to purchase, so (regardless of firm type) the optimal network-order pair must be one which would induce some high effort from a high-skill firm.

We can see that when $w > \pi p_{01}$ the firm optimal network will depend upon firm type. If the firm is a high-skill type, they value both incentivising their own effort (so that customers are willing to purchase under their prior) and providing later customer with the best information possible (as if later customers learn the firm is high-skill they will visit). In contrast, if the firm is a low-skill type, while they want the network to be one which would incentivise effort from a high-skill firm (so that customers are willing to purchase under their prior), they do not care about the information quality of later critical customers, since with fully revealing good reviews, no critical customer will ever purchase from a low-skill firm, as a low-skill

[19]Phrased another way, we can conjecture the following: if for some parameterisation $\mathcal{N}^*, \succeq^*$ is a consumer optimal network, there exists a reparameterization of the model for which $\mathcal{N}^* \succeq^*$ is socially optimal. In other words, we do not expect consumer optimal networks to explore parts of the network-order space not explored by socially optimal networks.

firm cannot receive good reviews.

We can reason that a low-skill firm should prefer a network-order pair which maximises aggregate incentives, as discussed in proposition 5. In contrast, it is possible a high-skill firm may not want to maximise aggregate incentives; they may prefer ex-ante to have fewer $t = 1$ customers and more $t = 2$ customers (even though $t = 2$ customers do not always purchase), if they need to exert high effort for all $t = 1$ customers to motivate them to visit. This is particularly clear if p_{11} is very close to 1; if $(p+11-p_{01})\delta w > c$, a high-skill firm prefers to have one t = 1 customer, observed by $N - 1$ $t = 2$ customers, while with $N = 2n$, the low-skill firm prefers a one-to-one matching between n $t = 1$ and n $t = 2$ customers (each reviewer is viewed by one reader, each reader reads one review). As such, there is a sense in which a low-skill firm's ex-ante preferences over networks may appear closer to consumer or a social planner's preferences than a high-skill firm's[20].

5 Conclusions

We have shown that when incentives are reputational, customer social networks matter for incentives; a firm has incentives for effort only if the quality of service they provide will be observed by a customer whose purchase decision depends upon reviews. The strength of these incentives depends on the network structure; incentives are stronger when many critical customers (and associated herds) are watching, but weaker when those critical customers have access to many other reviews. We have argued that as service generated by high effort becomes fully informative about firm type, efficient social networks can feature disjoint components if high effort is sufficiently desirable.

For tractability, we have focused on a simple model in which all players have

[20]Of course, conditional on the firm being a low type, customers and social planners would prefer a network which induced no trade; the point here is that when firms have discretion over observation structures, customers should be more sceptical of firms who choose to implement observation structures which appear ex-ante customer optimal.

perfect information about the network and customer order, and in which good service fully reveals the firm's type. Many of our insights generalise naturally to broader settings. For the remainder of this section, we discuss possible generalisations, and relate our results to alternative approaches to modelling reputation.

We would expect herding to be a feature of equilibrium which carried over to more general settings. In particular, wherever customers make a binary decision between purchasing and not purchasing at a fixed price, observing a customer's purchase decision will reveal information about their beliefs, and so herding is always possible when a poorly informed customer observes a well informed customer. If customer C_i observes C_j, and believes that customer C_j is sufficiently better informed than them (for example, if customer C_i's neighbours are a subset of customer C_j's neighbours), then customer C_i should purchase if C_j does. This would be true even if we allowed low skill firms to sometimes generate good service, or made quality of service a continuous random variable.

We might be interested in settings featuring weaker assumptions on the firm's knowledge of the network. If a firm is uncertain about who observes whom (either because they do not know the network structure or because they do not know where on the network a customer is located), then for each customer the firm will assign probabilities to that customer being observed by each other customer. Each customer will have a cut-off strategy in beliefs, which will depend on their location on the network[21]. We would expect analogues of many of our insights to hold in this more general setting; the firm will exert high effort for a customer if they believe it is *sufficiently likely* that customer is observed by a critical customer. Similarly, if a customer believes it is sufficiently likely that their friend is well informed, they will interpret that friend purchasing from the firm as a sign the firm is high quality and

[21]If we assumed the firm knew topological properties of the network but not the sequence of customers, an infinite network would simplify the firm's problem significantly; with an infinite network known to possess some kind of regular topology, the firm will not need to worry about the correlation in customer 'types' that occurs when drawing nodes from a finite network without replacement.

choose to purchase themselves.

While we can speculate about how a firm would behave if they did not know the network or order of customers, a full analysis is not easy. Note that the firm's problem is potentially very complicated; their strategy in a given period will depend on in how many previous periods they achieved high effort and in how many previous periods they were rejected. Historic rejections may suggest that herding has already begun to the firm's detriment, meaning the firm may be deterred from future high effort, believing there are few customers remaining whose purchase decisions they can influence.

We might also want to explore endogenous customer orders. In many settings it is unrealistic to assume that the orders in which customers visit is exogenously fixed; for example, customers generally choose when to book to visit a new restaurant. We would expect the order in which customers visit to depend upon their location on the social network; for example, central customers might visit early, uncertain about the firm's type but knowing the firm has an incentive for high effort if skilled, while less central customers might choose to visit later, letting their better connected friends act as reviewers first. This is work in progress.

It is worth noting that in the presence of reputational incentives, the predictions of a model will depend importantly upon the possible types of a firm. We have considered a model in which a high-skill strategic firm seeks to distinguish itself from a low-skill commitment type firm which always provides bad service. As a result, incentives are driven by a desire by the firm to distinguish themselves; the firm has an incentive to exert high effort if doing so can statistically identify them as high-skill rather than low skill to later customers. A consequence of this is that incentives strengthen over time in 'bad histories'; a firm who is yet to distinguish themselves as high-skill has a finite number of opportunities to do so in equilibrium, and finds incentives strengthen as the number of remaining opportunities reduces (other things equal).

In contrast, suppose we considered a model in which a low-skill strategic firm tried to imitate a high-skill commitment type (alternatively, a strategic firm seeking to imitate a Stackelberg type). Here, the firm has incentives for effort if doing so makes it harder for customers to identify them as low-skill. The consequence of this is that while incentives may strengthen in 'good histories' (as the chance of successfully going undetected given high effort goes up over time), they weaken in 'bad histories', as a firm who has already been 'caught' has nothing to gain from further high effort. In other words, if good firms seeks to distinguish themselves from bad, firms with good reviews are less likely to be exerting high effort; if bad firms seek to imitate good, firms with good reviews are more likely to be exerting high effort. We have shown that when incentives are reputational, customer social networks matter for incentives, because a firm has incentives for effort only if the quality of service they provide can affect the probability that friends of their current customer purchase in future. The strength of these incentives depends on the network structure; incentives are stronger when many critical customers (and associated herds) are watching, but weaker when those critical customers have access to many other reviews. We have argued that as service generated by high effort becomes fully informative about firm type, efficient social networks can feature disjoint components if high effort is sufficiently desirable.

For tractability, we have focused on a simple model in which all players have perfect information about the network and customer order, and in which good service fully reveals the firm's type. Many of our insights generalise naturally to broader settings. For the remainder of this section, we discuss possible generalisations, and relate our results to alternative approaches to modelling reputation.

We would expect herding to be a feature of equilibrium which carried over to more general settings. In particular, wherever customers make a binary decision between purchasing and not purchasing at a fixed price, observing a customer's purchase decision will reveal information about their beliefs, and so herding is always possible

when a poorly informed customer observes a well informed customer. If customer C_i observes C_j, and believes that customer C_j is sufficiently better informed than them (for example, if customer C_i's neighbours are a subset of customer C_j's neighbours), then customer C_i should purchase if C_j does. This would be true even if we allowed low skill firms to sometimes generate good service, or made quality of service a continuous random variable.

We might be interested in settings featuring weaker assumptions on the firm's knowledge of the network. If a firm is uncertain about who observes whom (either because they do not know the network structure or because they do not know where on the network a customer is located), then for each customer the firm will assign probabilities to that customer being observed by each other customer. Each customer will have a cut-off strategy in beliefs, which will depend on their location on the network[22]. We would expect analogues of many of our insights to hold in this more general setting; the firm will exert high effort for a customer if they believe it is *sufficiently likely* that customer is observed by a critical customer. Similarly, if a customer believes it is sufficiently likely that their friend is well informed, they will interpret that friend purchasing from the firm as a sign the firm is high quality and choose to purchase themselves.

While we can reason that many of our insights would carry over if a firm did not know the network or order of customers, a full analysis is not easy. Note that the firm's problem is potentially very complicated; their strategy in a given period will depend on in how many previous periods they achieved good service and in how many previous periods they were rejected by customers. Historic rejections may suggest that herding has already begun to the firm's detriment, meaning the firm may be deterred from future high effort, believing there are few customers remaining

[22]If we assumed the firm knew topological properties of the network but not the sequence of customers, an infinite network would simplify the firm's problem significantly; with an infinite network known to possess some kind of regular topology, the firm will not need to worry about the correlation in customer 'types' that occurs when drawing nodes from a finite network without replacement.

whose purchase decisions they can influence.

We might also want to explore endogenous customer orders. In many settings it is unrealistic to assume that the orders in which customers visit is exogenously fixed; for example, customers generally choose when to book to visit a new restaurant. We would expect the order in which customers visit to depend upon their location on the social network; for example, central customers might visit early, uncertain about the firm's type but knowing the firm has an incentive for high effort if skilled, while less central customers might choose to visit later, letting their better connected friends act as reviewers first.

It is worth noting that in general in settings of reputational incentives, the predictions of a model will depend importantly upon the possible types of a firm. We have considered a model in which a high-skill strategic firm seeks to distinguish itself from a low-skill commitment type firm which always provides bad service. As a result, incentives are driven by a desire by the firm to distinguish themselves; the firm has an incentive to exert high effort if doing so can statistically identify them as high-skill rather than low skill to later customers. A consequence of this is that incentives strengthen over time in 'bad histories'; a firm who is yet to distinguish themselves as high-skill has a finite number of opportunities to do so in equilibrium, and finds incentives strengthen as the number of remaining opportunities reduces (other things equal).

In contrast, suppose we considered a model in which a low-skill strategic firm tried to imitate a high-skill commitment type (alternatively, a strategic firm seeking to imitate a Stackelberg type). Here, the firm has incentives for effort if doing so makes it harder for customers to identify them as low-skill. The consequence of this is that while incentives may strengthen in 'good histories' (as the chance of successfully going undetected given high effort goes up over time), they weaken in 'bad histories', as a firm who has already been 'caught' has nothing to gain from further high effort. In other words, if good firms seeks to distinguish themselves

from bad, firms with good reviews are less likely to be exerting high effort; if bad firms seek to imitate good, firms with good reviews are *more* likely to be exerting high effort.

A final important distinction between our model and the general literature on reputations is that much of the literature focuses on settings in which the long-run player actions are statistically identified by observing an unbounded history of play. A key component to this assumption is that regardless of short-run player actions, outcomes are informative about long-run player actions. In our setting, this would be equivalent to requiring that observing the quality of service of an earlier friend is informative about firm effort even if that earlier friend did not purchase (for example, if firm effort generated externalities and so affected customer payoffs from outside option). If long-run player actions are statistically identified even in 'bad histories' (in which almost all customers do not purchase), later customers can still learn about firm type if they observe the full history. In contrast, in our model, if almost all of a customer's earlier friends have not purchased, with positive probability that customer will hold incorrect beliefs about the firm's type regardless of their visit date. Hence, compared to the standard reputations literature, our results place much more importance on ensuring initial customers purchase; once customers stop purchasing in our setting a firm can earn no more revenue, whereas if customers can learn about a firm's type even if their friends do not purchase, firms understand they can recover from initial non-purchases and a bad reputation is always salvageable for a good firm.

References

Acemoglu, Daron et al. (2011). "Bayesian Learning in Social Networks". In: *The Review of Economic Studies* 78.4, pp. 1201–1236. ISSN: 00346527, 1467937X. URL: http://www.jstor.org/stable/41407059.

Banerjee, Abhijit V. (1992). "A Simple Model of Herd Behavior*". In: *The Quarterly Journal of Economics* 107.3, pp. 797–817. ISSN: 0033-5533. DOI: 10.2307/2118364. eprint: https://academic.oup.com/qje/article-pdf/107/3/797/5298496/107-3-797.pdf. URL: https://doi.org/10.2307/2118364.

Bikhchandani, Sushil, David Hirshleifer, and Ivo Welch (1992). "A Theory of Fads, Fashion, Custom, and Cultural Change as Informational Cascades". In: *Journal of Political Economy* 100.5, pp. 992–1026. ISSN: 00223808, 1537534X. URL: http://www.jstor.org/stable/2138632.

Board, Simon and Moritz Meyer-ter-Vehn (2021). "Learning Dynamics in Social Networks". In: *Econometrica* 89.6, pp. 2601–2635. DOI: https://doi.org/10.3982/ECTA18659. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA18659. URL: https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA18659.

Ely, Jeffrey C. and Juuso Välimäki (2003). "Bad Reputation". In: *The Quarterly Journal of Economics* 118.3, pp. 785–814. ISSN: 0033-5533. DOI: 10.1162/00335530360698423. eprint: https://academic.oup.com/qje/article-pdf/118/3/785/5472938/118-3-785.pdf. URL: https://doi.org/10.1162/00335530360698423.

Golub, Ben and Evan Sadler (2016). "Learning in Social Networks". In: *Oxford Handbook of Economic Networks*. URL: https://dx.doi.org/10.2139/ssrn.2919146.

Mailath, George J. and Larry Samuelson (2001). "Who Wants a Good Reputation?" In: *The Review of Economic Studies* 68.2, pp. 415–441. ISSN: 0034-6527. DOI: 10.1111/1467-937X.00175. eprint: https://academic.oup.com/restud/

article-pdf/68/2/415/4369473/68-2-415.pdf. URL: https://doi.org/10.1111/1467-937X.00175.

Pei, Harry (2022). "Reputation Building under Observational Learning". In: *The Review of Economic Studies*. rdac052. ISSN: 0034-6527. DOI: 10.1093/restud/rdac052. eprint: https://academic.oup.com/restud/advance-article-pdf/doi/10.1093/restud/rdac052/45068574/rdac052.pdf. URL: https://doi.org/10.1093/restud/rdac052.

Sgroi, Daniel (2002). "Optimizing Information in the Herd: Guinea Pigs, Profits, and Welfare". In: *Games and Economic Behavior* 39.1, pp. 137–166. URL: https://ideas.repec.org/a/eee/gamebe/v39y2002i1p137-166.html.

Wolitzky, Alexander (2018). "Learning from Others' Outcomes". In: *American Economic Review* 108.10, pp. 2763–2801. DOI: 10.1257/aer.20170914. URL: https://www.aeaweb.org/articles?id=10.1257/aer.20170914.

A Proofs for Section 3 (Strategies and Equilibria)

Proposition 1. *If one customer C_t visits in each period t, $c \leq (p_{11}-p_{01})w\sum_{i=1}^{N-1}\delta^i$, $\frac{\pi p_{11}p_{01}}{\pi p_{11}+(1-\pi)} < w+u_0 < \pi p_{11}$, and all C_τ for $\tau \geq 2$ are connected to C_1 via a path $\langle C_1, C_{j_1}\rangle, \langle C_{j_1}, C_{j_2}\rangle ..., \langle C_{j_n}, C_\tau\rangle$ (for $j_n < \tau$ increasing in n), there exists an equilibrium in which $e_1 = 1$, C_1 always purchases, and all remaining customers C_τ purchase iff $s_1 = 1$, with $e_\tau = 0$ for $\tau \geq 0$.*

Proof. Suppose that customer C_t visits in period t. If $w+u_0 < \pi p_{11}$, then a customer C_t is willing to purchase from the firm under their prior if they conjecture $\bar{e}_t = 1$. If $\frac{\pi p_{11}p_{01}}{\pi p_{11}+(1-\pi)}$ then a customer C_t who conjectures $\bar{e}_1 = 1$, observes $s_1 = 0$, and anticipates $\bar{e}_t = 0$ is unwilling to purchase from the firm, because they believe it is too unlikely that the firm is a high type given they have observed one bad review.

It follows that if C_1 conjectures that $\bar{e}_1 = 1$ they are willing to purchase. Now suppose all C_τ for $\tau \geq 2$ are connected to C_1 via a path (routed only via customers who visit between $t = 1$ and $t = \tau$). Hence C_2 observes C_1. If C_2 conjectures $\bar{e}_1 = 1$ and $\bar{e}_2 = 0$ (in any history), it follows that $a_2 = 1$ iff $s_1 = 1$ is an optimal strategy for C_2. Now consider C_3. If C_3 observes s_1, then (conjecturing $\bar{e}_1 = 1$ and $\bar{e}_3 = 0$ (in any history) and understanding C_2s strategy), it also follows that $a_3 = 1$ iff $s_1 = 1$ is an optimal strategy for C_3; note if $s_1 = 0$ then we know $a_2 = 0$, so C_3 learns nothing from C_2, while if $s_1 = 1$ C_3 believes $\theta = 1$ regardless of s_2. If C_3 does not observe s_1 then (as they are connected to C_1 via a path) they must observe a_2. But as $a_2 = 1$ iff $s_2 = 1$ then this is identical to the case where they observe s_1, and an optimal strategy is to set $a_3 = 1$ iff $a_2 = 1$ iff $s_1 = 1$.

We can repeat this argument for $C_4, ..., C_N$; if C_τ conjectures $\bar{e}_1 = 1$ and $\bar{e}_\tau = 0$, then it is optimal for them to purchase if $s_1 = 1$, which they can infer either directly from observing s_1, or from observing the purchase decision of an earlier customer who themselves only purchases if they have inferred or observed that $s_1 = 1$. Hence

it is optimal for all customers to follow the suggested strategies.

Given customer strategies, a high-type firm finds it worthwhile to exert effort for C_1 if $c \leq (p_{11} - p_{01})w \sum_{i=1}^{N-1} \delta^i$ (since if $s_1 = 1$ then $a_\tau = 1$ for all $\tau \geq 2$, while if $s_1 = 0$ then $a_\tau = 0$ for all $\tau \geq 2$). Since all remaining customers purchase iff $s_1 = 1$, *conditional* on C_τ purchasing the firm does not find it worthwhile to exert effort for C_τ, since in any on-path history in which C_τ purchases, all remaining customers will purchase regardless of s_τ (note that off-path they might still want to exert effort, for example if C_2 unexpectedly purchased following $s_1 = 0$). Hence it is also optimal for the firm to follow the suggested strategy. □

Proposition 2. *If $\bar{t} = \underline{t}$, and each customer C_t visits in period t, actions on the equilibrium path in a pure strategy PBE are unique for almost all parameterisations.*

Proof. We can identify two reasons we may have multiple equilibra; if customer strategies depend upon their conjectures about firm effort, or if players do not have unique best responses holding fixed (other) customer strategies.

Suppose each customer C_t visits in period t and $\bar{t} = \underline{t}$.

Consider the set of customers who always purchase in equilibrium. Our first claim is that this set does not depend upon customer conjectures.

If $\max_t |\{C_j \in N(C_t) | j < t\}| = d_m$ gives the largest number of earlier neighbours that any customer ever observes, and $d_m < \underline{t}$, no equilibrium features high effort regardless of customer conjectures; if there is no customer on the network who can ever observe enough signals to be unwilling to purchase then there are no beliefs a customer can hold for which they would ever be unwilling to purchase in equilibrium, so the unique pure strategy PBE must feature all customers always purchasing.

Now consider networks such that $\bar{t} = \underline{t} > d_m$, where we cannot support all customers always purchasing. Again we can argue the set of customers who always purchase in equilibrium will not depend upon customer conjectures. $\bar{t} = \underline{t} < N$ tells us that if C_τ, who expects $\bar{e}_\tau = 0$, is willing to purchase believing they have observed n bad reviews generated when a high skill firm would have been exerting

high effort, they would also be willing to purchase expecting $\bar{e}_\tau = 1$ and believing they have observed n bad reviews generated when a high skill firm would have exerted low effort. It follows that if they are willing to purchase having seen n bad reviews holding maximially pessimistic and maximally optimistic beliefs, for any intermediate beliefs they are also willing to purchase.

Now turn to consider the set of customers who we have not identified as always purchasers (by implication this set is also independent of customer conjectures). $\bar{t} = \underline{t} < N$ tells us that regardless of customer conjectures these customers will only purchase if they believe $\theta = 1$ with certainty (as $\bar{t} = \underline{t}$ tells us that there is no information set for a customer at which flipping their conjecture about any of the firms' effort choices will change their purchase decision). Hence these customers will only purchase in equilibrium if they have either observed $s_j = 1$ for some neighbour C_j, or they have inferred $s_j = 1$ for some earlier customer C_j by observing the purchase decision of another customer C_k who only purchases when $s_j = 1$. It follows that the strategies of critical customers do not depend upon their conjectures about firm effort.

Hence if both the set of critical customers and the strategies of critical customers and always purchasers is independent of customer conjectures, equilibria are independent of customer conjectures. It remains to argue best responses are generically unique.

In our model customers and firms take discrete choices. If the optimal choice for customers and the firm at each information set is characterised by a strict inequality, it follows that all pure strategy PBE are unique. Our claim is that in the space of valid parameterisations for our model (comprising adjacency matrices for N nodes and values of c, w, p_{11}, p_{01} satisfying our definitions and assumptions), parameterisations in which either the firm or a customer are exactly indifferent at some information set are measure zero. To see this, firstly, fix a network, and consider the purchase decision of a customer C_k. In a pure strategy equilibrium, there

are a finite number of conjectures they could hold about firm strategies, and for each conjecture we can write their indifference condition as a linear equality (corresponding to a hyperplane in $c, w, p_{11}, p_{01}, u_0$). Hence for a fixed network the set of parameterisations for which a customer is indifferent corresponds to the union of a finite number of four-dimensional hyperplanes in five-dimensional space, and is measure zero. Similarly for firm effort choices, we can count parameterisations for which a firm is indifferent by starting in period $N-1$ and working backwards, and will obtain the union of a finite number of hyperplanes which is measure zero (note this relies on there being one customer per period, so that the firm is not taking effort choices simultaneously). As a result, the union of all parameterisations in which at least one player has an information set at which they are indifferent will be measure zero in the space of valid parameterisations, and so generically, best responses are unique. The proposition follows. □

B Proofs for Section 4 (Efficient Networks)

Proposition 4. *If $\mathcal{N}, \succeq_\alpha$ is a network-order pair such that equilibrium strategy profile $(\sigma, \boldsymbol{\rho})$ generates ex-ante distribution over histories $\Delta(H_t)$, there exists a bipartite network $\mathcal{N}_\mathcal{B}$ with node sets ${\mathcal{C}_\mathcal{B}}^1$ and ${\mathcal{C}_\mathcal{B}}^2$, such that for network-order pair $\mathcal{N}_\mathcal{B}, \succeq_\alpha$, we can find an equivalent equilibrium strategy profile $(\sigma_B, \boldsymbol{\rho}_B)$ which also generates ex-ante distribution over histories $\Delta(H_t)$, and in equilibrium all $C_i \in {\mathcal{C}_\mathcal{B}}^1$ always purchase on the equilibrium path.*

Proof. Let $\mathcal{N}, \succeq_\alpha$ be a network-order pair such that equilibrium strategy profile $(\sigma, \boldsymbol{\rho})$ generates ex-ante distribution over histories $\Delta(H_t)$.

Firstly, for $(\sigma, \boldsymbol{\rho})$ identify the set of customers who always buy ${\mathcal{C}_\mathcal{B}}^1 = \{C_j | P(a_j = 1 | \sigma, \boldsymbol{\rho})\}$. Now delete all links on $\mathcal{N}$ between nodes in ${\mathcal{C}_\mathcal{B}}^1$, to produce new network $\mathcal{N}'$, and (as the strategy space has now changed because customers in ${\mathcal{C}_\mathcal{B}}^1$ now have different possible information sets), define a new strategy profile $\boldsymbol{\rho}'$ such that $a_j = 1$

for all $C_j \in \mathcal{C}_{\mathcal{B}}{}^1$, and the remaining player strategies are unchanged.

We can see that $(\sigma, \boldsymbol{\rho}')$ must be an equilibrium strategy profile for $\mathcal{N}', \succeq_\alpha$, since all customers in $\mathcal{C}_{\mathcal{B}}{}^1$ take exactly the same decision as they were taking at every information set on $\mathcal{N}$, and so remain willing to always purchase, and the problems of the other customers and the firm remain unchanged (since the firm was unable to influence purchase decisions of customers in $\mathcal{C}_{\mathcal{B}}{}^1$ on $\mathcal{N}$ via a deviation in effort choice). Given our construction, we can also see strategy profile $(\sigma, \boldsymbol{\rho}')$ generates the same ex-ante distribution over histories $\Delta(H_t)$ as $(\sigma, \boldsymbol{\rho})$ (recall that a history H_t is defined as all purchase decisions and quality of service realisations up until period t, and so the set of possible histories depends only on $\mathcal{C}$ and $\succeq_\alpha$).

Now perform the following process: identify all C_k^t such that $C_k \notin \mathcal{C}_{\mathcal{B}}{}^1$ and C_k has only observed customers in $\mathcal{C}_{\mathcal{B}}{}^1$ when they visit the firm (formally, $\{C_j^{t-s} | C_j^{t-s} \in N(C_k^t), C_j^{t-s} \notin \mathcal{C}_{\mathcal{B}}{}^1, s \geq 1\} = \emptyset$). Note that these customers must exist, since by definition they do not purchase at some information set, and there must be some earliest period in which the first customer(s) who does not always purchase visits the firm.

For each C_k^t, identify the set $N(C_k^t) \cup \mathcal{C}_{\mathcal{B}}{}^1$, and for each $C_k^{t+s} \in N(C_k^t)$ $(s > 1)$, add $N(C_k^t) \cup \mathcal{C}_{\mathcal{B}}{}^1$ to $N(C_k^{t+s})$ and remove C_k^t from $N(C_k^{t+s})$, giving new network $\mathcal{N}_1'$. In other words, for each of the first 'wave' of critical customers C_k^t, let all of their later neighbours observe the elements of $\mathcal{C}_{\mathcal{B}}{}^1$ that C_k^t observed *instead* of observing C_k^t.

We can define a new strategy $\boldsymbol{\rho}_1'$ which preserves the strategies of players whose neighbours have not changed, and for each customer we have removed from $N(C_k^t)$, has them take the same action at each history as under $\boldsymbol{\rho}$, since if they know C_k^t's pure strategy they can effectively condition their purchase decision on a_k without observing C_k^t, because they can infer a_k from their information set (and they never conditioned their choice on C_k^t's quality of service in equilibrium, so they have lost no payoff relevant information). Given optimality of original strategies, these new

strategies will remain optimal (and all other players' strategies remain optimal). Again, we can see that $\mathcal{N}_1', \succeq_\alpha$ will be a network-order pair such that equilibrium strategy profile $(\sigma, \boldsymbol{\rho}_1')$ generates ex-ante distribution over histories $\Delta(H_t)$.

We can now repeat this 'cutting-out-the-middle-man' process on $\mathcal{N}_1'$ until we obtain a network where no critical customer has any neighbours not in ${\mathcal{C}_\mathcal{B}}^1$. Denote the set of critical customers ${\mathcal{C}_\mathcal{B}}^2$. The proposition follows. □

Lemma 1. *If all customers visit in $t = 1$ and $t = 2$, and all $t = 1$ customers receive high effort, aggregate expected benefit from effort for firms is maximised if $t = 2$ customers are critical and have degree m^*, where*

$$m^* = \in \left\{ \lfloor \frac{-1}{\ln(1 - p_{11})} \rfloor, \lceil \frac{-1}{\ln(1 - p_{11})} \rceil \right\}$$

Proof. Our argument will have the following shape: suppose $t = 1$ customers always purchase and are observed by critical customers who visit at $t = 2$, and that all $t = 1$ customers receive high effort. Fixing the number of customers who visit in each period, we will suppose that each $t = 2$ customer observes m $t = 1$ customers and that all $t = 1$ customers have equal degree (ignoring divisibility issues). We will then solve for the optimal value of m that maximises firm aggregate expected benefit from effort across all customers (equivalent to finding the level of m which gives the highest possible cost for which a firm is willing to exert effort for all $t = 1$ customers).

Note that divisibility aside, maximium aggregate expected benefit to effort must be maximised by a symmetric network where all customers $t = 1$ have the same degree and all customers at $t = 2$ have the same degree. To see this, note that if $t = 2$ have degrees that differ by more than 1 we could move a link from a $t = 2$ customer with a large degree to a $t = 2$ customer with a smaller degree and increase aggregate incentives for effort, since the increase in incentives as a result of making the high degree customer easier to influence will outbalance the fall in incentives

from making the low degree customer harder to influence.

If we have mn_2 links going out from $t = 2$ customers, we must have mn_2 links coming in to the n_1 $t = 1$ customers, meaning each $t = 1$ customer must have $\frac{mn_2}{n_1}$ links.

Suppose all customers conjecture that $e_k = 1$ for all C_k^1. Assuming the firm sets $e_k = 1$ for all C_k^1 for $k \neq j$, the firm's expected benefit from setting $e_j = 1$ rather than setting $e_j = 0$ is:

$$\frac{mn_2}{n_1}(p_{11} - p_{01})(1 - p_{11})^{m-1}\delta w - c$$

We can see that firm incentives to exert effort for C_j^1 are decreasing in m; if the customers watching C_j^1 see the results of the firm exerting effort for many other customers, the effort they exert for C_j is less likely to make a difference. Now let's aggregating these incentives across all C_j^1:

$$mn_2(p_{11} - p_{01})(1 - p_{11})^{m-1}\delta w - n_1 c$$

We want to maximise this with respect to m. Note that this is equivalent to maximising $m(1 - p_{11})^m$. Taking logs we have $ln(m) + mln(1 - p_{11})$, which is a concave function of m. Taking FOC we have $\frac{1}{m} + ln(1 - p_{11}) = 0$, maximised at $m = \frac{-1}{ln(1-p_{11})}$

Since this may not be an integer, the actual degree for $t = 2$ customers which maximises aggregate incentives will either be obtained by rounding up or down to give degree m^8. If it is possible to construct a bipartite network where all $t = 2$ customers have degree m^*, this will maximise aggregate incentives. □

Proposition 5. *In a bipartite network featuring n_1 customers visiting at $t = 1$ and receiving high effort, and n_2 customers visiting at $t = 2$, we must have $\frac{n_1}{n_2} \leq \frac{1}{a}$, where a is given by:*

$$a = -\frac{[ln(1 - p_{11})]c}{(p_{11} - p_{01})\delta w}(1 - p_{11})^{\frac{1+ln(1-p_{11})}{ln(1-p_{11})}}$$

Proof. We will argue as follows: ignoring divisibility issues, we will work out the largest possible ratio of $t = 1$ to $t = 2$ customers we can support such that all $t = 1$ customers receive high effort. We will then argue that this ratio allows us to bound above the largest possible ratio of $t = 1$ to $t = 2$ customers we can support on a *feasible* network (accounting for divisibility concerns).

Our previous lemma tells us that if we had no divisibility issues and could give each $t = 2$ customer degree $m^* = \frac{-1}{ln(1-p_{11})}$, this would maximise the firm's aggregate expected benefit to effort for a fixed ratio of $t = 1$ customers to $t = 2$ customers. This is equivalent to saying that each $t = 2$ customer makes the largest possible contribution to aggregate incentives in $t = 1$ when they have degree m^*. In turn, this is equivalent to saying that for any c, the number of degree m $t = 2$ customers who must be watching C_j^1 for a firm to set $e_j = 1$ is lowest when $m = m^*$. Recall we can write aggregate incentives as:

$$mn_2(p_{11} - p_{01})(1 - p_{11})^{m-1}\delta w - n_1 c$$

Evaluated at m^*:

$$-n_2(p_{11} - p_{01})\frac{(1 - p_{11})^{-\frac{1}{ln(1-p_{11})}-1}}{ln(1 - p_{11})}\delta w - n_1 c$$

Now note that to maximise $n_1 = N - n_2$ for fixed N we would like to set aggregate incentives equal to 0, such that there are just enough $t = 2$ customers to motivate effort for $t = 1$ customers to receive high effort. If $t = 2$ customers have degree m^*, this occurs when:

$$\frac{n_2}{n_1} = -\frac{[ln(1 - p_{11})]c}{(p_{11} - p_{01})\delta w}(1 - p_{11})^{(\frac{1+ln(1-p_{11})}{ln(1-p_{11})}}$$

We must have at least this many $t = 2$ customers for each $t = 1$ customer to incentivise high effort for each $t = 1$ customer, if we can give $t = 1$ customers non-

integer degrees. When we restrict ourselves to giving $t = 1$ customers an integer degree, it follows that we will need weakly more $t = 2$ customers for each $t = 1$ customers. Hence we have a lower bound on $\frac{n_2}{n_1}$ (and thus an upper bound on $\frac{n_1}{n_2}$) for any feasible network. □

Proposition 6. *If $\delta < 1$ and the optimal network $\mathcal{N}^*$ is a symmetric bipartite network, then any optimal order $\succeq^*$ allocates all customers to periods $t = 1$, $t = 2$, or $t = 3$. If $\pi p_{01} - w < u_0$, any optimal order $\succeq^*$ allocates all customers to periods $t = 1$ or $t = 2$.*

Proof. Firstly, note if the optimal network-order pair induces no trade or induces full trade with probability 1, for $\delta < 1$ the optimal order assigns all customers to $t = 1$.

The more interesting case is when the optimal network-order pair features critical customers on the equilibrium path. In this case, the proof for this result comprises three smaller claims:

1. An optimal order features at most one initial period in which customers purchase but none receive high effort on the equilibrium path.

2. An optimal order allocates all customers who receive high effort with positive probability to the same period, which is no later than $t = 2$.

3. An optimal order allocates all critical customers no later than the period immediately following customers who receive high effort with positive probability.

Claim 1: If $\pi p_{01} - w < u_0$, the first customers to purchase must receive high effort with positive probability, so there is no initial period of 'always buys' receiving low effort.

If $\pi p_{01} - w \geq u_0$ then customers are willing to purchase under their prior expecting low effort. If the planner chooses not to attempt to incentivise any high effort on the equilibrium path, they do best allocating 'guinea pigs' to $t = 1$ and followers

to $t = 2$; if high effort is impossible, there are no strategic concerns so discounting makes them move customers as early as possible.

If the planner does want to incentivise high effort it is possible they may find it hard (or impossible) to do so for $t = 1$ customers, if the firm is sufficiently confident they will attract later critical customers even if exerting low effort.

As a result, the planner may want to create the opportunity for the firm's reputation to decay, so that in bad histories, the firm does have incentives for high effort. A planner may be best off (or have no choice other than) giving up on high effort for the first customers, accepting that following good reviews for the first customers the firm may avoid ever exerting high effort, but following bad reviews from the first customers, the firm will have stronger incentives for high effort.

It remains to argue that if planner wants to provide this opportunity for reputations to decay, they need only provide one period for decay.

The only way that allocating customers receiving low effort to $t = 1$ can benefit the planner via the incentives channel is if all $t = 1$ customers viewed by a $t = 3$ critical customer receive bad service (such that the $t = 2$ customer receiving high effort remains willing to purchase if they expect high effort). Since the probability of doing so does not depend upon the number of periods across which the planner allows for decay, it follows that the planner needs at most one period before the first customer who receives high effort with positive probability.

Claim 2: We argue that for customers who receive high effort with positive probability, letting a firm take high effort choices sequentially rather than simultaneously can only weaken incentives with a symmetric bipartite network, reducing average effort on the equilibrium path.

To see this, suppose first that C_i and C_j are 'always buys' who both receive high effort at some info sets on the equilibrium path, and are both observed only by critical customer C_k. Ignore discounting, and suppose that C_i is served first, followed by C_j and then C_k.

Consider incentives when serving C_i. The firm knows that if they receive a bad review from C_i, they have a second chance at convincing C_k by exerting high effort when serving C_j. In other words, exerting high effort for C_i only matters if they would end up blowing their second chance and receiving a bad review from C_j. Hence the probability that effort for C_i ends up mattering is $(1 - p_{11})$; the probability that if they fail for C_i, they also fail for C_j. But note that this is exactly the same probability that high effort for C_i ends up mattering as in the case where C_i and C_j visit simultaneously. In other words, ignoring discounting, incentives for exerting high effort for C_i are no stronger than when C_i and C_j visit simultaneously (with discounting, they are actually weaker).

Now consider incentives when serving C_j. If $s_i = 1$, C_k contributes nothing to firm incentives for effort when serving C_j, as the firm knows that $a_k = 1$ regardless of s_j, so the firm will set $e_j = 0$. If $s_i = 0$, the firm should set $e_j = 1$, as they found $e_i = 1$ optimal so should definitely find $e_j = 1$ optimal given they no longer have a second chance. As such, with customers visiting sequentially, C_j receives high effort with probability $(1 - p_{11})$.

But now consider C_i and C_j visiting simultaneously. We know the firm finds $e_i = 1$ optimal as incentives are the same as when they visited sequentially, and we know we must have $e_j = 1$ since the firm's IC conditions are now symmetric. So now we have C_i receiving high effort with probability 1 as before but we have increased the probability C_j receives high effort by p_{11}.

We made this argument when C_i and C_k are observed by a single customer, but we can extend it to any network in which C_i and C_j receive high effort with positive probability and are observed by the same number of critical customers who have not yet seen a good review. As such, for a symmetric bipartite network, the social planner does best when all customers who might receive high effort visit simultaneously (note that symmetry must also be with respect to the order, in that we need it to be the case that if C_k and C_l observes C_i and C_j, and C_k observes some

earlier C_d, to give symmetric incentives at all histories we also need C_l to observe C_d).

Claim 3: Since a planner can restrict attention to bipartite networks in which critical customers observe only 'always buys' without loss, if $\delta < 1$ the latest the planner should ever allocate critical customers is the period after the final 'always buy' visits, as allocating them later is harms social welfare due to discounting, weakens incentives for earlier customers due to discounting and does not benefit the critical customers themselves in any history.

The proposition follows. □

Proposition 7. *In the class of bipartite networks with node sets* ${\mathcal{C}_{\mathcal{B}}}^1$ *and* ${\mathcal{C}_{\mathcal{B}}}^2$ *and* $|{\mathcal{C}_{\mathcal{B}}}^1| = n_1$, $|{\mathcal{C}_{\mathcal{B}}}^2| = n_2$, *where all members of* ${\mathcal{C}_{\mathcal{B}}}^1$ *visit the firm at* $t = 1$, *the complete bipartite network is optimal if* $n_1 \leq m^*$, $u_0 > 0$, *and* c *is sufficiently small.*

Proof. We can consider three possibilities a planner might want to achieve in this class of bipartite networks: allocating all customers in ${\mathcal{C}_{\mathcal{B}}}^2$ to $t = 1$ (for example if δ is very small), allocating ${\mathcal{C}_{\mathcal{B}}}^2$ customers to $t = 2$ without incentivising effort for any $t = 1$ customers, and allocating ${\mathcal{C}_{\mathcal{B}}}^2$ to $t = 2$ while trying to incentivise effort for at least some $t = 1$ customers. We can see that if the planner wants this first possibility, any network is optimal.

For the second and third possibilities, note that if $u_0 > 0$, then the social planner would like each customer to purchase only if the firm is a high type, so other things equal, the firm would like customers to receive a high quality of information, allowing them to avoid bad firms while reducing the probability they fail to purchase from good firms. Hence if we fix the number of customers visiting at $t = 1$, then for any customers allocated to $t = 2$ or later, the planner would benefit from letting the $t = 2$ or later customers observe an additional $t = 1$ customer if doing so does not cause the firm to switch from high to low effort for any $t = 1$ customer. It follows a bipartite network is optimal if the planner prefers the second possibility (for example, if effort is only marginally productive).

On the third possibility, we can see that in the class of networks which induce high effort for *all* customers in ${\mathcal{C}_{\mathcal{B}}}^1$, the complete bipartite network will be optimal, since the complete bipartite network maximises aggregate incentives for effort, and so must induce high effort for all ${\mathcal{C}_{\mathcal{B}}}^1$ if *any* network can, and holding fixed firm effort choices, the complete bipartite network maximises information quality for customers in ${\mathcal{C}_{\mathcal{B}}}^2$.

A potential problem case may occur if the firm can only incentivise effort for *some* $t = 1$ customers. Adding links between ${\mathcal{C}_{\mathcal{B}}}^1$ and ${\mathcal{C}_{\mathcal{B}}}^2$ will still increase *aggregate* incentives, but may do so by reducing some individual incentives and increasing other individual incentives, and as such, because we consider binary effort choices, the effect on total effort may be ambiguous. As a result, it may be that the complete bipartite network is *not* Pareto-superior to an incomplete bipartite network from the planner's perspective, if it reduces total expected effort. We can rule out this problem cases by assuming that c is not *too* large, in which case the proposition follows. □

Proposition 8. *There exists some n' and some $\beta \geq 0$ such that if $N > n'$ and $p_{11} - c \geq \beta$, if high effort is possible in equilibrium for some network-order pair, no complete bipartite network is an efficient network.*

Proof. To prove this result, we firstly argue that as $N \to \infty$, the expected effort of the firm averaged across all customers converges to 0 on any complete bipartite network. We then argue that the planner can incentivise effort (in expectation) for a positive proportion of customers and thus, if $p_{11} - c$ is sufficiently large, earn a higher expected payoff averaged across all customers with a non-complete bipartite network as $N \to \infty$.

Firstly, let's assume the planner designs a bipartite network with node sets ${\mathcal{C}_{\mathcal{B}}}^1$ and ${\mathcal{C}_{\mathcal{B}}}^2$ and $|{\mathcal{C}_{\mathcal{B}}}^1| = n_1$, $|{\mathcal{C}_{\mathcal{B}}}^2| = N - n_1$ and allocates all n_1 customers in ${\mathcal{C}_{\mathcal{B}}}^1$ to $t = 1$, and the other customers to either $t = 1$ or $t = 2$. We will explore how the planner's maximum expected payoff varies as N becomes large and they vary n_1.

With a complete bipartite network, the strongest possible incentives for effort for the firm to exert effort for a specific $t = 1$ customer occur if they plan zero effort for all other customers, with IC condition $\delta n_2(p_{11} - p_{01})(1 - p_{01})^{n_1 - 1} \geq c$. We can see that if we bound $\frac{N-n_1}{n_1}$ above then as $N \to \infty$ the LHS of this equation tends to 0, and so even with strongest incentives possible with a complete bipartite network the condition must fail to hold and no customers will receive high effort. If we let $\frac{N-n_1}{n_1}$ grow unboundedly, then we may be able to incentivise high effort for one (or all) $t = 1$ customers as $N \to \infty$, but the proportion of customers at $t = 1$ will converge to 0, so the average effort across all customers still converges to 0.

We can thus argue that as $N \to \infty$, with a complete bipartite network the proportion of customers receiving high effort must converge to 0, because either customers in ${\mathcal{C}_{\mathcal{B}}}^1$ receive zero effort for N sufficiently large, or the proportion of customers in ${\mathcal{C}_{\mathcal{B}}}^1$ converges to 0. Note this would also remain true if the planner used an order which spread customers across the first 3 periods, rather than 2.

A planner maximising average expected surplus thus has the following possible *optimal* choices for N large with a complete bipartite network: allocate all customers to $t = 1$ and have them take the decision they would under their prior, or let n_1 and $\frac{N-n_1}{n_1}$ grow unboundedly, allocating all ${\mathcal{C}_{\mathcal{B}}}^2$ to $t = 2$, letting almost all customers take the decision they would take under full information.

If customers always purchase under their prior, the first option gives average expected payoff per customer of πp_{01}, while the second gives $\delta(\pi p_{01} + (1 - \pi)u_0)$, so the first option is better if $u_0 < 0$ or if the firm is sufficiently impatient. In contrast if $\pi p_{01} < w + u_0$, so customers only purchase under their prior expecting high effort the first option gives average expected payoff of u_0, while the second still gives $\delta(\pi p_{01} + (1 - \pi)u_0)$.

We now note that if there is some finite network with which a planner can induce high effort, then as $N \to \infty$ the planner can induce a high effort for a positive proportion of customers by creating a new network featuring as many disjoint du-

plicates of this finite network as possible (the proportion will not be constant in N because there may be remainder customers, but we can see it is bounded away from 0).

Let $0 < d$ be the expected proportion of customers that a planner can induce high effort for from a high skill firm with an incomplete bipartite network. A (loose) lower bound on expected payoff from using the incomplete bipartite network described is thus $d\pi(p_{11} - c)$. We can see that if this is better than the maximum average expected payoff achievable with a complete bipartite network, the sender does better with our incomplete bipartite network. Depending on what order the planner would choose with a complete bipartite network, our condition will thus be $p_{11} - c > \frac{p_{01}}{d}$, $p_{11} - c > \frac{u_0}{\pi d}$, or $p_{11} - c > \frac{\delta}{d} p_{01} + \frac{1-\pi}{\pi} \delta u_0$.

Our proposition follows; if it is possible to achieve high effort for at least one customer with some finite network, above some lower bound on $p_{11} - c$, with sufficiently many customers a planner does better duplicating this finite network multiple times than designing a complete bipartite network. □

Chapter 5

Conclusions

We have presented each paper as self-contained. In this section, we provide a longer discussion of how features of our separate models might interact in a more general setting, and discuss possible extensions to our papers in greater depth than in the papers themselves. We also discuss how our results would be affected if customers were less sophisticated.

Seeding with Noisy Disclosure We have explored separately the problems of a sender making a disclosure via a noisy channel and a sender looking to seed a network with information in the presence of noisy transmission. For tractability, we did not explore the problem of a sender who first chooses a receiver to seed and then makes a verifiable disclosure to them when messages between receivers acquire noise. If we were interested in this problem, we can note that receivers on a network would need to have some action they took in the absence of any message, and given this, with an unbounded state there would exist some interval of sender types who preferred to send no message and receive the non-disclosure actions with certainty to the distribution of actions induced by noisy disclosure.

A consequence of this would be that receiver posterior beliefs should place zero probability on states for which the sender does not disclose, and so updating would be much more complicated. We can note, however, that (as in "Noisy Disclosure"),

partial disclosure in settings of noisy transmission may *benefit* rather than harm receivers, because the sender can noiselessly signal information via their disclosure decision. For example, if a sender only seeds the network with high or low disclosures, a receiver would know an intermediate message must be the result of noise. This is particularly important for receivers who observe very noisy messages; if a receiver is located a long way from the point of seeding on a network, such that their message is almost entirely noise, they definitely benefit ex-ante from any amount of signalling via the sender's disclosure decision. As such, it may be in at least some receivers' interests to allow the sender to sometimes choose to not disclose.

Turning to the sender, we can note that introducing a disclosure decision does not change the shape of the sender's underlying ex-ante incentives for information provision; though the sender *would* need to take into account how their equilibrium disclosure rule will vary with their seeding location. If we fix a sender's disclosure rule and consider optimal seeding given that disclosure rule, we can see the specific centrality measure which characterises optimal seeding will change. We show in "Strategic Information Release on a Communication Network" that providing receivers with access to a second informative signal about the state of the world has the effect of making their posterior variances 'less convex' in distance from the sender. In other words, when a receiver has access to no other information, doubling the amount of noise in their message affects their posterior variance more than if they have access to another informative signal. The implication is that optimal seeding should be more sensitive to the exact length of long paths in the presence of outside information for receivers. We can follow a similar line of reasoning if the sender sometimes does not disclose; since all receivers observe an informative signal (comprising the sender's disclosure decision) in addition to their messages, short path lengths are less likely to dominate optimal seeding decisions, meaning the relevant centrality measure is likely to be more similar to closeness centrality.

In addition to introducing a disclosure decision for the sender, we might also

want to introduce a disclosure decision for receivers as they pass information to one another. As we note in the paper, our base model offers no strict incentives for receivers to share information with one another, so for their disclosure decision to be interesting we would need to introduce some additional motivation (for example, receivers may play a repeated gossip game in which the seeds differ between rounds, and in which receivers punish neighbours who do not pass on messages by not passing on messages themselves in future).

Again, receivers could attempt to convey additional information via their disclosure decisions. Note, however, that the benefit from doing so is reduced when they are passing on a message that has already been censored. For example, if the sender only seeds the network in positive states, then if a receiver always passes a message on their neighbours can infer the state is negative when they observe no message; in contrast, if receivers only sometimes pass on the messages they receive, their neighbours do not know whether no message is the result of censorship by the sender or their neighbour. With aligned preferences, this may imply that receivers do not want to further censor messages given the sender has already optimally censored. With heterogeneous preferences among receivers Gieczewski (2022) shows that when there is uncertainty over who was seeded (but no noise in transmission), only information about moderate states of the world is likely to disseminate through a network via verifiable disclosures; if receivers cared about one another's actions we would expect a similar result to hold with noise in transmission and observed seeding.

If we introduced a disclosure decision into our noisy seeding model, we would have signals comprising a normal random variable with an interval censored, plus normal noise. Posteriors in this setting quickly become complicated at the cost of analytic tractability. We might thus want to look for a more tractable model in which to combine disclosure and noisy transmission. One alternative approach to modelling noisy information transmission is to model signals as binary, with a

positive probability of flipping value as they travel (as in Jackson, Malladi, and McAdams (2019)).

We note that in the absence of additional frictions, noisy disclosure games are much simpler with binary states. For example, a sender who is aligned with receivers can achieve perfect communication by only disclosing in one state, signalling the state noiselessly via their disclosure decision (if messages go missing with positive probability, as in Jackson, Malladi, and McAdams (ibid.), signalling via disclosure decision will also be noisy). In contrast if a sender is misaligned with receivers, as in "Noisy Disclosure" they will either engage in full or no-disclosure; since signalling via disclosure decision is cheap talk. In either of these cases, taking a seeding model with binary signals and noisy transmission and introducing a disclosure decision does not generate significant new insights[1]. Either the sender bypasses noise completely by signalling via their disclosure decision, or the sender always discloses, in which case we recover our base seeding model with no strategic disclosures. As such, we can note that while introducing disclosure to our base model may not give a tractable model, a model with a binary state may prove *too* tractable to explore the trade-offs a sender can face in a noisy disclosure problem.

Longer Message Chains with Networked Customers In "Reputational Incentives with Networked Customers", we model customers as able to learn from their immediate neighbours on a network, but not from their neighbours' neighbours. A more complicated model might allow customers to also learn about the purchase decisions or reviews of customers who are not their neighbours, either with additional noise or with a lower probability of observation.

We can see that introducing a positive probability that customers observe their neighbours' neighbours will have a similar effect to adding links to the customer

[1]With a binary state, to have an equilibrium in which a receiver observing a noisy message is unsure of the state, but in which the sender does not always disclose, we would need the sender to always disclose in one state and mix between disclosure and non-disclosure in the other state, such that a receiver observing a (noisy) message is uncertain about which state the sender observed.

network in our base model. Both from the perspective of learning and incentives, the effect of adding links is ambiguous. Holding incentives fixed, later customers do better when they can observe more early customers, but making it easier for early customers to observe one another could lead to earlier herding, harming later customers who benefit when early customers take independent purchase decisions and so experiment for longer. Turning to incentives, we know that increasing connectivity between customers means more later customers read the reviews of early customers, but each individual review customers read is less likely to determine their purchase decision; as such, letting customers learn about their neighbours' neighbours with positive probability may strengthen incentives with a sparse customer network but weaken them with a dense customer network.

Introducing noise, either in observations of neighbours or of neighbours' neighbours, would move our model away from a good news model; if false positives (good reviews of bad firms) are possible, then in general it will no longer be the case that one piece of good news is sufficient to convince customers to purchase. This does not affect our main insights; intermediate connectivity will still deliver the strongest incentives because as the number of reviews a customer observes grows, the cost of generating a distribution of posterior beliefs which is significantly different from the distribution induced by zero-effort grows unboundedly regardless of whether we have a good news model.

As a comparative static on a fixed network, we would expect making observations noisier to weaken incentives, since the 'gap' between the high-effort and low-effort review distributions will shrink. Note that in part this is due to our assumption that the 'bad' firm type is a non-strategic type who always provides bad service. In settings of reputational incentives, we can reason roughly that bad types are motivated to effort because they want to be pooled with good types, while good types are motivated to effort because they want to separate from bad types. In general, with reputational incentives we can see that a noiseless performance measure must

motivate no effort from the lowest type in equilibrium, while a noisy performance measure may motivate effort from the lowest type. As such, we can reason that there is a sense in which noise in customer observations would be more likely to strengthen incentives of a strategic bad type than a strategic good type, and so comparative statics on noise in observations will depend upon which type (if any) in a model is non-strategic.

Networked Customers with More Complex Information Structures In "Reputational Incentives with Networked Customers", we modelled customers as observing the purchase decisions and reviews of their neighbours on a network. Stepping back, we could explore reputational incentives with more complex information structures, for example with an information designer controlling more aspects of the information structure.

One question of interest is whether an information designer could improve welfare by changing what a customer can observe. In our model, a customer has separate information sets for when a neighbour has not purchased, and for when they have purchased but dislike a product; an alternative information structure might pool these (for example, if bad reviews get deleted such that later customers cannot observe the difference between a bad review and no purchase).

The key distinction between our base model and this information structure would be that only herds on 'don't purchase' would be possible; because customers can no longer observe their neighbours' purchase decisions, they cannot infer whether they neighbour observed a good review unless their neighbour *wrote* a good review (in which case what they observed is irrelevant). As a result, critical customers will only purchase if they observe a good review themselves. As before, it will remain possible that if after a certain date no good review has been observed, all remaining customers stop purchasing, but unlike in the base model, it is no longer the case that if, for example, customers only observe the previous period, there is an equilibrium

in which following a good review in the first period all remaining customers purchase.

The effect on incentives for a fixed network will be ambiguous; incentives for effort for a specific customer may become weaker, if it is no longer possible that they can start a 'herd' of customers who always purchase, but incentives may be more persistent over time, because if the network is not complete one success does not imply the firm will quit high effort for all remaining customers. If effort is sufficiently cheap this second effect should dominate. As such, if effort is both sufficiently cheap *and* productive, pooling non-purchases and bad reviews is likely to be welfare improving for many networks, since the benefit of more customers receiving high effort in equilibrium will outweigh customers taking less well informed decisions.

Naive Customers Throughout, we model customers as Bayesian; they understand that gossip may be unreliable, that they may miscomprehend messages written in a language they do not speak, and that whether a firm exerts effort for a customer affects the inference they should draw if that customer writes a bad review. We view these as plausible assumptions about the majority of customers, but in some circumstances there may be at least some customers who do not update beliefs in a rational way. As such, we offer a brief discussion of how our results would be affected if customers updated beliefs naively.

In "Strategic Information Release on a Communication Network" customers place more weight on their priors and less weight on their messages when their messages are noisier. Suppose instead that they were credulous, and believed that their messages were equal to the state, failing to take into account the noise accumulated via the 'broken telephone' effect. We can see that this will have a similar effect on seeding incentives to the case in which receivers do not observe seeding; receivers respond to the amount of noise they believe to be in their messages (here, none) rather than the amount of noise that is actually in their messages. As we

show, in this setting a sender now wants to provide all receivers with noiseless messages regardless of preference misalignment, because the sender only benefits from providing noisy messages to misaligned receivers when the receivers *understand* that their messages are noisy and put less weight on them accordingly. As a result, with homogeneous receiver preferences, a sender would want to seed the most central receiver (with credulous receivers, the closeness central receiver) regardless of preference misalignment. As such, with unobserved seeding, credulous receivers will affect the details but not the form of optimal seeding, while with observed seeding introducing credulous receivers will give qualitatively very different incentives to the sender.

In "Noisy Disclosure" the receiver again understands that following disclosure they observe a noisy message, and updates accordingly, while following non-disclosure the receiver updates based upon their conjecture of the sender's disclosure strategy. This suggests that there are two dimensions in which we might model a receiver as naive: they might be naive to noise and so not update correctly following a noisy message, even if they remain strategic in the sense that their updating rule does take into account their conjecture of sender strategy; they could also be *strategically* naive, and so fail to take into account that the sender's disclosure decision is informative about the state in equilibrium.

Firstly, let's suppose that a receiver continues to set their non-disclosure action based upon their conjecture, but is credulous on observing a disclosure featuring additive noise, and treats it at face value[2]. We would still expect to see partial disclosure in equilibrium if noise is not too large, since disclosure remains risky for the sender (but if receivers believe off-path messages are the result of deviations credu-

[2]Note that because we assume a bounded state, it is unclear how best to capture a credulous receiver who observes an extreme message; if $\theta \in [0, 1]$ and the receiver observes $s < 0$, are they sophisticated enough to only assign probability to possible states, even if they are not Bayesian? Are they sophisticated enough to only assign probability to states that match their conjecture about the sender's disclosure strategy? In other words, are they completely irrational, are they acting as though disclosure is noiseless, or are they updating as though noise is vanishingly small but unbounded?

lous updating would no longer give upper censorship with positive bias). In contrast, if noise were very large, it might be that all sender types with small enough bias find disclosure too risky with credulous receivers and strictly prefer non-disclosure. The effect of making receivers credulous on the exact partial disclosure equilibrium cutoff is ambiguous; disclosing now gives the cutoff type an action closer to the true state on average, but also a more volatile action as the receiver does not adjust the weight they place on messages to take account of the noise.

Now let's consider an even less sophisticated receiver who in addition to treating messages at face value, inherits their prior on observing no message. In this setting, we would expect to have partial disclosure in equilibrium for all bias levels with sufficiently small noise, since some sender types must strictly prefer to be believed to be the mean type to receiving a draw from the action distribution following disclosure. Note the sender's problem is reduced from a strategic communication problem to a non-strategic standard choice under uncertainty problem; the sender's problem is significantly simpler with receivers who are completely unsophisticated.

In "Reputational Incentives with Networked Customers", customers condition their updating both on the purchase decisions and reviews of their neighbours, and update differently depending upon their conjecture of firm effort. As such, there are several ways customers might be less sophisticated; they could update only based on reviews, ignoring purchase decisions, or they could fail to take into account firm effort when updating and instead set a fixed threshold for how many bad reviews deter a purchase. If customers did not update based on conjectured effort, we would not have multiple equilibria, but our findings would otherwise be qualitatively unchanged. If customers updated only based on reviews, rather than purchase decisions, in contrast, our results would be significantly affected, because there would be no herding in equilibrium. Similar to our discussion on alternative information structures, no herding on 'purchase' following good reviews would make incentives more persistent over time but weaker for each individual customer on many networks.

No herding on 'don't purchase' would imply that if a customer who observed all their friends not purchase simply inherited their prior, it would be possible for good firms to escape from a bad reputation; in contrast, in the base model, if all later customers observe some earlier customers and purchases stop after some period, it may be impossible for them to restart. Removing the possibility of herding on may weaken firm incentives for early customers (they are punished for bad reviews with a smaller 'stick'), but could benefit later customers who might now sometimes receive high effort.

Questions for Future Work We have discussed how our findings might differ if features of our models could be combined or extended. We conclude the thesis by discussing what we view as the most interesting open question relating to each of our papers.

"Strategic Information Release on a Communication Network" studied the problem of a sender choosing a single seeding point on a network in the presence of noisy transmission, where the transmission process featured a single noisy message travelling from the seeding point to each connected receiver. One of the most interesting open questions in settings of noisy seeding is: how should a sender optimally seed a network if a distinct noisy message travels between the seeding point and each receiver along *each* path? In such a setting a receiver observing multiple messages could learn about the noise in messages by comparing two messages which have travelled separate routes.

Given tractability issues with modelling multiple messages with a normal state and noise, it may be more fruitful to explore this question in a setting with a binary state, where messages flip with positive probability in transmission. We can see that the effect of giving a receiver access to multiple messages on seeding incentives will be ambiguous. For example, suppose a receiver takes a binary action which they seek to match to the state, and holds a prior that the state is equally likely to be

high or low. A receiver observing two noisy messages will always base their decision on the less noisy of the two messages; as such, if receivers have at most two paths between themselves and any possible seeding point, the sender's seeding problem with multiple messages is identical to their problem with a single message travelling a shortest path. In contrast, if three or more paths connect a receiver to the seeding point, the sender's seeding problem will differ.

"Noisy Disclosure" studied the problem of a sender who observes a random state of the world and can choose to disclose it to a receiver, understanding that if they disclose the receiver will observe a noisy message. One of our main insights was that with noisy disclosure there may exist equilibria in which a sender who wants to be *overestimated* discloses only in sufficiently *low* states. One open question is whether we can find equilibria of this form in other settings. In particular, we have noted that our noisy disclosure model can be viewed as a model of disclosure with an endogenous cost; we can thus ask: what other forms of disclosure cost can give us upper censorship with a positively biased sender?

We can reason that if disclosure is noiseless but requires paying a cost (which is independent of the state), we may be able to find upper censorship equilibria with positive bias for some priors. To illustrate, suppose that the sender has small positive bias, and consider what happens as we reduce disclosure costs. If costs are sufficiently large no type will find disclosure worthwhile, but as we reduce costs, eventually some type will have an incentive to start disclosing. If the state has a symmetric distribution and the sender has positive bias, the first sender to start disclosing will be the sender observing the highest state. However, if the state is left-skewed, such that the prior mean is very high, the sender observing the lowest state receives a very large overestimate from non-disclosure, compared to a correct estimate from disclosure, while the sender in the highest state receives a small underestimate from non-disclosure and a correct estimate from disclosure. If sender bias is small, it may thus be that as disclosure costs fall, the sender in the *lowest*

state starts disclosing first, because they experience the greatest loss from receiving the non-disclosure action. As such, we can see that upper censorship equilibria with positive bias are likely to exist in more general settings of costly disclosure, and that existence conditions may depend upon the prior.

"Reputational Incentives with Networked Customers" studies a setting in which customers observe the past reviews and purchase decisions of their neighbours on a network, and a firm can influence reviews by choosing a non-contractible costly effort. In the paper, we focus on a setting in which the order of customer visits and the customer observation network are both commonly understood, as a tractable first step to understanding firm incentives in more general settings. In general, we can note that reputational incentive problems are most tractable either when the environment is simple and commonly understood (as in our model) or when the environment features a large degree of uncertainty which firms and customers have limited ability to resolve. Intermediate models can be significantly harder to study; for example, if we modified our base model such that the firm did not observe customer identity or know the visiting order of customers, the firm would face potentially a very complicated updating problem trying to infer the identity of past customers from the purchase decisions of current customers.

As such, an interesting open question and next step is to consider reputational incentives with networked customers when firms and customers have *much less* information about the network structure. For example, we might consider reputational incentives for customers who visit a firm in a random order, are located on an infinite network with known topological properties, and observe only their immediate neighbours. The firm then has the following concern; conditional on a customer purchasing, how likely is it that they have neighbours that are critical customers who have not yet observed a good review? Their incentives will depend upon network topology; for example, if the probability of triads in the network is very high (such that two friends are likely to be friends with one another's friends), then the firm

may quit high effort earlier, since if a later customer purchases, this indicates they have seen a good review or informative purchase decision, and so the high probability of triads tells us that any of their friends who are yet to visit are likely to also have seen that review or purchase decision, making effort for that customer unlikely to affect revenue. Studying infinite networks with known topological properties gives us more scope to explore comparative statics on network topology, in contrast to our base model, which is better placed to help us understand specific interventions, such as cutting a link.

Aidan Smith

8/8/2022

References

Gieczewski, Germán (2022). "Verifiable communication on networks". In: *Journal of Economic Theory* 204, p. 105494. ISSN: 0022-0531. DOI: https://doi.org/10.1016/j.jet.2022.105494. URL: https://www.sciencedirect.com/science/article/pii/S0022053122000849.

Jackson, Matthew O., Suraj Malladi, and David McAdams (2019). "Learning through the Grapevine and the Impact of the Breadth and Depth of Social Networks". URL: http://dx.doi.org/10.2139/ssrn.3269543.