Breaking Bad News*

Moritz Meyer-ter-Vehn† ⓡ Simon Board‡

January 25, 2025

Abstract

We study a model of social learning on networks where agents learn about product quality when their neighbors are harmed by the product. Adding links to the network may benefit some agents at the expense of others; nevertheless, we show that all agents agree that the complete network, and hence full transparency, is uniquely optimal. This insight extends to a general class of social learning games: A sufficiently powerful information designer can Pareto-improve outcomes by publicly releasing bad news.

1 Introduction

We study a simple social learning model in which agents can adopt a new product and learn from prior agents' experiences. For example, consider doctors learning about new drugs (e.g. Coleman, Katz, and Menzel (1957)), consumers learning about new gadgets (e.g. Goolsbee and Klenow (2002)) or entrepreneurs learning about new financial products (e.g. Banerjee, Chandrasekhar, Duflo, and Jackson (2013)). In such settings a social planner (e.g. medical association, industry watchdog, government

*Thank you to Dilip Abreu, Dirk Bergemann, V Bhaskar, Alex Bloedel, Daniel Clarke, Ben Golub, Daniel Haanwinckel, Marina Halac, Matt Jackson, Jay Lu, Bryony Reich, Fedor Sandomirskiy, Yuliy Sannikov, Karl Schlag, Vasiliki Skreta, Leeat Yariv seminar audiences at Games 2024, NSEC2024, Princeton, NYU, UCLA, USCS, UT-Austin, Washington University for helpful comments. Keywords: Network, Social Learning, Transparency, Bad News, Information Design; JEL D83, D85

†UCLA, http://www.econ.ucla.edu/mtv/

‡UCLA, http://www.econ.ucla.edu/sboard/

regulator) may wish to improve on market outcomes by reporting the outcomes of early movers to agents who choose later. How should such policies be designed?

This paper focuses on learning from "perfect bad news" events that proves the product is low quality (e.g. side effects of drugs, exploding gadgets, and predatory financial practices). Such bad news events are often the focus of transparency laws that, in the US, started with the 1906 Pure Food and Drug Act and the 1933 Securities Exchange Act. For example, the FDA requires drug companies report adverse health events, the CPSC requires manufacturers report unsafe products, and the SEC requires customers be notified of any events that could cause financial harm like fraud (e.g. Fung, Graham, and Weil (2007)).

Our model follows Bikhchandani, Hirshleifer, and Welch (1992) in studying informational externalities among agents who can consume a product of unknown quality. Agents move at exogenous times and observe the outcomes of some earlier-moving "neighbors"; we interpret the resulting information structure as a network. One of the the outcomes, "harm", is proof that product quality is low.

We wish to compare welfare across networks and understand how a planner should design such networks. This enterprise faces two theoretical challenges. First, it is complicated to solve for Bayesian equilibrium on a given network, let alone solve for the optimal network. Second, adding a particular link can make one agent better off and another worse off. Nevertheless, we will see that the optimal network is remarkably simple: Full transparency.

The case against transparency in models of social learning is that choices made by early agents can lead to herding, hindering further information generation and social learning by later agents. For example, Che and Hörner (2018) show that withholding "perfect good news" signals raises welfare by inducing socially beneficial experimentation in the converse event where such signals fail to materialize.[1] In our setting with perfect bad news learning, social learning induces a "guinea pig" effect that suggests a strong conflict of interest. To see this, consider three agents (Ruby, Sven, Tara) who can try the product in order and learn when a neighbor is harmed. Assume Tara observes Sven's outcome but not Ruby's. Allowing Sven to observe

[1]Relatedly, Sgroi (2002) and Kremer, Mansour, and Perry (2014) show transparency is suboptimal in models with, respectively, observable actions and imperfect outcome signals.

Ruby's outcome helps Sven avoid low-quality products. This is good for Sven, but the reduced chance of seeing Sven harmed lowers Tara's learning. Sven's information (about Ruby's outcome) makes him a less effective guinea pig for Tara.

In Section 3 we show that, despite this conflict of interest, every type of every agent prefers the complete network to all other networks. In our three agent example, the key idea is that, if Tara also observes Ruby's outcome directly, she only cares about learning from Sven in the event that Ruby has not been harmed. And since the absence of bad news is good news, observing Ruby makes Sven more optimistic and inclined to try the product, rendering him a more effective guinea pig for Tara. Thus, while agents differ in their preferences over adding a single link, they all agree on the optimal network. This result can be extended: A *transitive* network, in which each agent sees everything her neighbors see, Pareto-dominates all of its subnetworks.

In Section 4, we ask whether full transparency about perfect bad news outcomes also constitutes a Pareto-improvement in models with richer outcomes, which provide inconclusive evidence about low or high quality (e.g. agents' actions). Simply releasing bad news can make some agents worse off. However, a sufficiently powerful information designer, who can directly communicate with the agents and control their information, can fix this problem and render full transparency about bad news a Pareto-improvement using a "survey mechanism" that is robust and detail-free. Taken together, these results provide a logical foundation for the transparency policies (such as those run by the FDA, CSPS, and SEC), when agents learn via bad news signals.

1.1 Literature

We contribute to a large literature studying social learning on networks starting with Bikhchandani, Hirshleifer, and Welch (1992) and Smith and Sørensen (2014). Our network model follows Acemoglu et al. (2011) in that agents move in deterministic order and observe information from their neighbors in addition to a private signal. The key difference is that our agents see their neighbors' outcomes instead of their actions, as in Wolitzky (2018) and Board and Meyer-ter-Vehn (2021), and that observing a neighbor harmed constitutes perfect bad news. This allows us to Pareto-rank networks.

There is also a large literature on information design (e.g. Bergemann and Morris (2019)) that includes models where communication is dynamic (e.g. Ely (2017)) or is constrained by a network (e.g. Galperti and Perego (2022)). The typical paper in this literature characterizes the optimal information policy for a particular game. In contrast, we are interested in whether a particularly simple policy − full disclosure of perfect bad news − is Pareto optimal across a wide range of environments.

More closely related to our result, Knoepfle and Salmi (2024) show that a social planner would release perfect bad news signals without delay in a game of endogenous timing. Our setting with exogenous timing is simpler, allowing us to study a broader range of individualized information policies. Indeed, with endogenous timing, asymmetric policies that reveal Sven's outcome to Tara (but not vice versa) raise utilitarian welfare by inducing Sven to move immediately, which is optimal for Tara.

2 Model

Players and Actions. Agents $\{1, \ldots, r, s, t, ..., T\}$ (Ruby, Sven, Tara) consider a new product of uncertain, binary quality $q \in \mathcal{Q} := \{L, H\}$. At time t, agent t decides (once and for all) whether to consume the product, $c_t \in C_t := \{0, 1\}$ resulting in a binary outcome: "harm" $\xi_t = \dagger$ if she consumed low quality, $c_t = 1$ and $q = L$, and "null" $\xi_t = \emptyset$ otherwise.

Information. Tara observes a type θ_t and the outcomes of agents $G_t \subseteq \{1, ..., t-1\}$. We interpret $\{G_t\}$ as a directed network; when $s \in G_t$, we write $s \leftarrow t$ and call s a *neighbor* of t.

Payoffs and Beliefs. Consuming the product yields utility $u(\theta_t, \xi_t)$; not consuming yields 0. Agents dislike harm

$$u(\theta_t, \dagger) < u(\theta_t, \emptyset). \tag{1}$$

The state space is $\Omega = \mathcal{Q} \times \Theta$. Tara has type-dependent beliefs $\Pr_{\theta_t}(q, \theta_{-t})$ over quality and others' types. If θ_t assigns probability one to an event, we say she "knows"

the event.

Solution Concept. Assume throughout the paper that agents consume the product when indifferent. Proceeding by induction over t, we solve for the unique rationalizable strategy profile, and hence the unique (Bayesian-Nash) equilibrium of this game. Given this unique equilibrium, we rank networks $G \succcurlyeq G'$ according to the interim Pareto-ranking of all types θ_t of all agents t in all type spaces.

Remarks. Our key assumption is that the high-quality product harms nobody, so harm constitutes perfect bad news proving that quality is low.[2] Thus Tara only faces two relevant classes of histories: the empty history where she observes no harm and the histories in which she observes harm, in which she concludes quality is low. This distinguishes our model from traditional social learning models (e.g. Bikhchandani, Hirshleifer, and Welch (1992)) in which agents learn from actions rather than outcomes, and a neighbor's consumption choice constitutes imperfect news. We turn to such models in Section 4.

Our specification of types $\theta = (\theta_1, ..., \theta_T)$ is very general. Tara's private type θ_t may influence her private tastes about consuming the product $u(\theta_t)$ (as in Goeree, Palfrey, and Rogers (2006)), and private signals about quality $p(\theta_t) = \Pr_{\theta_t}(q = H)$ (as in Bikhchandani, Hirshleifer, and Welch (1992)). The model does not require a common prior over Ω and, if there is a common prior, it allows for arbitrary correlations. Our results easily extend to settings with timing uncertainty (where t moves at some time $\tau(\theta_t)$), imperfect harm (where consuming low-quality only harms some agents), and random networks.[3]

3 Optimal Network

The conditional probability that θ_t observes harm

$$y(\theta_t) = \Pr{}_{\theta_t}(\xi_s = \dagger \text{ for some } s \in G_t | L) \tag{2}$$

[2]The converse assumption that low-quality harms all consumers is for convenience.

[3]The latter extensions require that the occurrence of harm upon low-quality consumption and the realization of the network are independent events.

is a (Blackwell)-sufficient statistic for her social information about quality. The higher is $y(\theta_t)$, the higher is θ_t's information and hence welfare. This formalizes the notion that t uses her neighbors as guinea-pigs, indicating a conflict of interest.

When θ_t observes no harm, she updates her beliefs upwards to

$$p^{\emptyset}(\theta_t, y(\theta_t)) := \frac{p(\theta_t)}{1-(1-p(\theta_t))y(\theta_t)} \tag{3}$$

and consumes the product if (3) exceeds the threshold $p^*(\theta_t)$ that uniquely solves her indifference condition $pu(\theta_t, \emptyset) + (1-p)u(\theta_t, \dagger) = 0$.[4]

We first show that "indirect links" can make agents worse off.

Example 1. Suppose we start with a network in which Tara observes Sven, and add a link from Sven to Ruby (see Figure 1, left). Clearly, the new link is beneficial for Sven who may learn from Ruby's experience and thus has Blackwell-more information. However, adding Ruby presents a tradeoff for Tara: If Ruby is harmed, Sven gets pessimistic and less likely to try, which is bad for Tara; if Ruby is unharmed, Sven gets optimistic and more likely to try, which is good for Tara.

On average, the new link may make Tara worse off. To understand the tradeoff, assume that types capture private consumption costs $\theta_t = \kappa_t \sim U[0,1]$, independent across agents and independent of q; the common belief of high-quality is $p(\theta_t) \equiv p$. Assume utility $u(\theta_t, \dagger) := -\kappa_t < 1-\kappa_t =: u(\theta_t, \emptyset)$, inducing cutoff $p^*(\theta_t) = \kappa_t$. Sven consumes the low-quality product and is harmed if his random posterior P_s (which equals $p^{\emptyset}(\theta_s, y(\theta_s))$ with probability $y(\theta_s)$ and 0 otherwise) exceeds κ_s (which is uniform on $[0,1]$). All told, Tara's chance of seeing Sven harmed is $y(\theta_t) = \Pr_{\theta_t}[P_s \geq \kappa_s | L] = E_{\theta_t}[P_s|L]$. Any rise in Sven's information, such as the link to Ruby, helps Sven avoid low-quality products and lowers $E_{\theta_t}[P_s|L]$; this lowers Tara's social learning and welfare.[5,6] △

[4]The solution $p^*(\theta_t) \in \mathbb{R}$ is unique by (1). If more strongly $u(\theta_t, \dagger) < 0 < u(\theta_t, \emptyset)$ we have $p^*(\theta_t) \in (0,1)$.

[5]With other distributions of costs κ_s, Sven's link to Ruby may benefit Tara. For example, suppose that Sven's costs κ_s exceeds p for sure, so he doesn't consume given his prior beliefs; the link to Ruby may induce him to consume, benefiting Tara.

[6]Looking across agents, Tara's loss from Sven's link to Ruby is balanced by Sven's gain, so the overall welfare effect is ambiguous. But if Tara has many clones who all observe the same Sven, then the link $r \leftarrow s$ reduces utilitarian welfare.

Figure 1: **Networks from Examples 1-2.** The **left** panel adds a link from Sven to Ruby as in Example 1. The **right** panel adds a correlating link as in Example 2.

The next example shows how the complete network resolves Tara and Sven's conflict of interest.

Example 2. Suppose that Tara initially observes Ruby as well as Sven, and consider the effect of adding a link from Sven to Ruby (see Figure 1, right). Since Tara observes Ruby directly and seeing one neighbor harmed is proof of low-quality, she cares about learning from Sven only in the event that Ruby is unharmed. And in this event the link $r \leftarrow s$ makes Sven unambiguously more optimistic, raising his chance of consumption and thereby Tara's social learning and welfare. $\triangle$

Full transparency for Sven is thus optimal for Sven and Tara. More broadly, Tara wants Sven to observe any proof of low quality that she observes herself, since this makes him the most effective guinea pig by directing his experimentation to the event where it maximizes Tara's learning, namely where Ruby is unharmed.[7]

To generalize Examples 1 and 2, call network G *transitive* if whenever Tara observes Sven and Sven observes Ruby then Tara also observes Ruby directly; formally, $r \in G_s$ and $s \in G_t$ imply $r \in G_t$. This condition describes hierarchical organizations where agents observe information from all direct and indirect reports; it also captures settings where agents see prior agents with delay. One instance of a transitive network is the complete network $\bar{G}$, where $\bar{G}_t := \{1, ..., t-1\}$ for all $s < t$.

Theorem 1. *A network G Pareto-dominates all of its subnetworks $G' \subseteq G$ iff G is transitive. Thus, the complete network $\bar{G}$ Pareto-dominates all other networks.*[8]

[7]While Tara and Sven agree on full transparency, they would not agree on other policies. Indeed, if Tara could force guinea pig Sven to consume the product, as in Smith, Sørensen, and Tian (2021), she would do so, which is clearly not in Sven's interest. It is thus important that the only "design variable" is the social information agents learn from one another.

[8]Mild assumptions about the support of beliefs guarantee that the Pareto-ranking is strict.

Proof. Define the set of states $\omega = (q, \theta)$ where t observes harm, or equivalently, a neighbor of t consumes the low-quality product

$$B_t = \{\omega : \xi_s = \dagger \text{ for some } s \in G_t\} = \{(L, \theta) : c_s = 1 \text{ for some } s \in G_t\}.$$

By definition $y(\theta_t) = \Pr_{\theta_t}(B_t|L)$, so θ_t's information and hence welfare rise in B_t. Writing $B'_t, y'(\theta_t)$ for the respective variables in network G', we will show that $B'_t \subseteq B_t$ for all $G' \subseteq G$ iff G is transitive.

That transitivity is necessary follows from Example 1: If $s \in G_t$ and $r \in G_s$, but $r \notin G_t$, we can construct a state $\omega = (L, \theta)$ where (i) Ruby tries and is harmed, (ii) Sven tries unless he sees Ruby harmed, and (iii) no other agent tries. So constructed, $\omega \notin B_t$ but $\omega \in B'_t$ for subnetwork $G' \subseteq G$ with a sole link $s \leftarrow t$.

We argue sufficiency by induction. The anchor $B_1 = B'_1 = \emptyset$ is trivial. Assume $B_s \supseteq B'_s$ for all $s < t$, and fix $\omega = (L, \theta) \in B'_t$, so Tara's neighbor Sven consumes and is harmed in G'. Now consider network G; we wish to show $\omega \in B_t$. If some neighbor Ruby of Sven's is harmed, then by transitivity Tara also observes that Ruby is harmed. If none of Sven's neighbors are harmed, then the induction hypothesis renders Sven more optimistic in G, $p^{\emptyset}(\theta, y(\theta_s)) > p^{\emptyset}(\theta, y'(\theta_s))$ since

$$y(\theta_s) = \Pr{}_{\theta_s}(B_s|L) \geq \Pr{}_{\theta_s}(B'_s|L) = y'(\theta_s).$$

Thus, Sven consumes and is harmed in G. In either case, $\omega \in B_t$, concluding the induction step. □

Since observing harm perfectly reveals product quality and nothing remains to be learnt, one might wonder why the case for full transparency even requires a proof. And yet, Che and Hörner (2018) show that signals that perfectly reveal high quality may lower ex-ante welfare: Such information makes agents more pessimistic in the absence of good news signals and reduces socially beneficial experimentation. In contrast, the absence of bad news signals makes agents more optimistic. Indeed, in the perfect good news version of our model[9] the optimal network for all types θ_t of Tara is the *t-star,* where she observes all other agents, and other agents observe nobody.[10] This,

[9] That model flips inequality (1), swaps "high" and "low" quality, and replaces "harm" with "success".

[10] *Proof:* Fix a state ω and let $s < t$ be the first agent who succeeds in any given network G.

of course, is the worst network for all other agents, meaning there is maximal conflict over the network.[11]

An immediate corollary of Theorem 1 is that the complete network remains optimal when harm occurs with delay. Indeed assume that Sven's harm from consuming the low-quality product is only observed after time $\tau(s) \geq s$, so only $G \subseteq G_t^\tau = \{s : \tau(s) < t\}$ are feasible. Since G^τ is transitive, full transparency Pareto-dominates.

Theorem 1 shows that the clique is optimal for the agents. Is it optimal for the product's seller?

Corollary 1. *The complete network $\bar{G}$ maximizes demand $\sum_t c_t$ for the high-quality product for all types θ.*

Proof. The proof of Theorem 1 shows that whenever type θ_t consumes in network G, she also consumes in the clique $\bar{G}$ when harm has not yet occurred as is the case for $\theta = H$. □

Moving beyond high-quality sellers, suppose agents' beliefs are derived from a common prior $\Pr(q, \theta)$ and types capture costs $\theta_t = \kappa_t$ and are uniformly distributed, as in Example 1. Tara's expected consumption equals the prior probability of high-quality $E[P_t] = p$, so unconditional expected demand equals $E\left[\sum_t c_t\right] = Tp$, which is independent of the network. For these beliefs, the complete network then minimizes expected demand for low-quality sellers, as well as maximizing it for high-quality ones.[12]

Since s has not observed a success in G, she is less optimistic than in the t-star. Thus, she also tries and succeeds in the t-star. All told, t observes more successes, and hence has Blackwell-superior information in the t-star.

[11] Transparency is optimal for perfect bad news signals but not for perfect good news signals. How about other signals? By continuity, Theorem 1 extends to "mostly harmless" high-quality products that cause harm with some small probability ϵ. But for larger values of ϵ the clique is no longer optimal: The correlating link $r \leftarrow s$ in Example 2 may dissuade Sven from consumption when Ruby is harmed, depriving Tara of a second harm-signal. Inferences are subtle with imperfect bad news signals: While an exogenous harm outcome constitutes bad news, the fact that some neighbor chose to consume the product constitutes countervailing good news (as in Wolitzky (2018)).

[12] Corollary 1 has interesting implications for settings where the seller knows the quality of its good and designs the network G before the game. In such a signaling game, the natural equilibrium is for the high-quality seller to signal high-quality by choosing $\bar{G}$, forcing the low-quality seller to pool and also choose full transparency.

4 Transparency with Observable Actions

In Section 3, we considered a simple model in which agents learn from each other via perfect bad news outcomes and showed that transparency is Pareto optimal. In practice, Tara may learn from her neighbor Sven in many ways: She may observe his actions, or his outcomes may constitute "imperfect signals". For example, a doctor may see whether colleagues prescribe a new drug and the effect it has on their patients' symptoms. In such a context, is it still desirable to transparently release "perfect bad news" events? There are many example of such events in the context of our applications: The anti-nausea drug Thalidomide led to birth defects; faulty lithium-ion batteries in Samsung's Galaxy Note 7 caused some phones to explode; and investors in Bernie Madoff's Ponzi scheme discovered they could not withdraw money. It seems intuitive that one would want to publicly release such bad news. In this section we ask: Do such transparency laws robustly lead to Pareto improvements?[13]

So motivated, consider a model with general, finite, latent outcomes $\tilde{\xi}_t \in \Xi_t$; these include the outcomes from the baseline model, harm $\dagger$ and null $\emptyset$, and may additionally include observable actions or other imperfect signals about quality. Tara's actual outcome equals her latent outcome $\xi_t = \tilde{\xi}_t$ if she consumes the product $c_t = 1$, and $\xi_t = \emptyset$ if $c_t = 0$.

As in the baseline model, high-quality prevents harm $\Pr_{\theta_t}(\tilde{\xi}_s = \dagger | q = H) = 0$ for all s, t, so observing harm perfectly reveals $q = L$ which is bad news in that

$$E_{\theta_t}[u(\theta_t, \xi_t)|q = L] < E_{\theta_t}[u(\theta_t, \xi_t)|q = H],$$

generalizing (1). Formally, the state space now consists of quality, (latent) outcomes, and types, $\Omega = \mathcal{Q} \times \Xi \times \Theta$. Social learning is fully captured by q in that beliefs $\Pr_{\theta_t}(q, \theta_{-t}, \xi)$ impose independence between ξ_t and (θ_{-t}, ξ_{-t}) conditional on q.

Our leading example of non-binary outcomes are observable actions as in Bikhchandani, Hirshleifer, and Welch (1992) and Acemoglu et al. (2011). Here, Tara's trinary outcome is $\xi_t = \emptyset$ when she doesn't consume, $\xi_t = c$ when she consumes but is not harmed, and $\xi_t = \dagger$ when she is harmed.

[13] We only consider releasing "perfect bad news" events. We know from Sgroi (2002), Kremer, Mansour, and Perry (2014), and Che and Hörner (2018) that information about actions and good news events cannot Pareto-improve outcomes.

4.1 Broadcasting Harm

The simplest transparency policy G^+ augments a given network G with a harm broadcast, whereby t observes outcomes ξ_s of all neighbors $s \in G_t$ and binary "harm indicators" ϕ_s for all predecessors $s < t$ that indicate whether $\xi_s = \dagger$ or $\xi_s \in \Xi \backslash \dagger$.[14] The following example shows that this policy doesn't necessarily constitute a Pareto-improvement over G:

Example 3. Consider a network G with two links $r \leftarrow s \leftarrow t$ and a type of Tara θ_t who knows that (i) no agent before Ruby tries, (ii) Ruby doesn't try based on her prior, but is induced to try by the absence of harm in G^+ and has outcome $\xi' \notin \{\emptyset, \dagger\}$, (iii) Sven interprets ξ' as bad news: he consumes when he sees $\xi_r = \emptyset$ and is harmed if quality is low, but doesn't consume when he sees $\xi_r = \xi'$. Type θ_t learns q perfectly in G, but learns nothing in G^+.[15] △

The problem with broadcasting harm in Example 3 is that while it fixes the observation network for outcomes other than harm, the harm broadcast alters agents actions and outcomes. Specifically, in the example, Ruby's outcomes ξ' dissuades Sven from trying, compared to outcome $\emptyset$ in G.

We now consider an *adjusted harm broadcast* G^* that addresses this flaw and Pareto-improves on the original network, G. Consider an Information Designer who observes outcomes and informs Tara at time t whether any previous agent $s < t$ has been harmed, ϕ^*_s; absent harm, the Designer additionally informs Tara of the hypothetical outcomes $\xi_{G_t} = (\xi_s)_{s \in G_t}$ of her neighbors in the original network G (but not of the actual outcomes $\xi^*_{G_t}$ induced by G^*). This information trivially Blackwell-dominates her information in G and hence raises Tara's utility.

The critical question is how the Designer can infer the hypothetical outcomes ξ_s from the observed actual outcomes ξ^*_s. Analogous to Theorem 1, the key idea is that in the absence of harm the new policy G^* makes agents more optimistic, induces

[14] This can be viewed as a multi-layer network, where observability of outcomes is governed by G and observability of harm indicators by $\bar{G}$.

[15] Example 3 shows that in most networks, broadcasting harm does not raise the welfare of all types of all agents. This is clearly an exacting welfare measure, and it seems likely that for many natural type spaces and networks G, the harm broadcast G^+ does raise expected surplus.

more experimentation, and thus provides the Designer with weakly more information than the original network G. Formally, suppose inductively that, conditionally on $\phi_s^* \neq \dagger$ for $s < t$, the Designer has learned the hypothetical outcomes ξ_s for $s < t$. The induction step requires him to learn Tara's hypothetical outcome, ξ_t. To do so, the Designer can simply ask Tara for her consumption $c_t = c(\xi_{G_t}, \theta_t)$ in the hypothetical scenario that she doesn't learn whether $\phi_s^* \neq \dagger$ for $s < t$. Assuming that (absent incentives to misrepresent) Tara is truthful, the Designer can then deduce ξ_t as follows. If $c_t = 0$, trivially $\xi_t = \emptyset$. If instead $c_t = 1$, then she also consumes when the additional information of no harm, $\phi_s^* \neq \dagger$ for $s < t$, makes her more optimistic $c_t^* = c(\xi_{G_t}, \phi^*_{\bar{G}_t}, \theta_t) = 1$, and so the hypothetical outcome equals the actual outcome, $\xi_t = \xi_t^*$, which the Designer observes.

To implement the adjusted harm broadcast G^*, the Designer does not need to know quality q or agents' types θ, but must be able to (i) observe past outcomes, (ii) control agents' access to information, both by providing neighbors' hypothetical outcomes ξ_s and by withholding their actual outcomes ξ_s^*, and (iii) survey agents to elicit their hypothetical consumption c_t. The construction is "robust" (Wilson, 1987) in that it achieves the Pareto-improvement in any type space, and "detail-free" (Dasgupta and Maskin, 2000) in being formulated in terms of actual choices instead of abstract types. However, the need to conceal information makes it unclear how to decentralize G^*.

4.2 Information Design

The prior section considered two ways of supplementing social learning in network G. The harm broadcast G^+ simply adds information about past agents' harm, but fails to Pareto-improve G. The adjusted harm broadcast G^* fixes this problem, and Pareto-improves G by altering the observability of outcomes; in particular concealing the actual outcomes $\xi^*_{G_t}$ of Tara's neighbors. Unlike Theorem 1, which Pareto-improves any "network mechanism" G with the complete "network mechanism" $\bar{G}$, the construction of G^* raises the question whether any mechanism in the same (yet to be defined) class of mechanisms as G^* can be Pareto-improved by a mechanism in the same class that broadcasts harm.

To formalize this, we follow and expand on different classes of mechanisms in the information design literature (e.g. Bergemann and Morris (2019)). Specifically, we consider mechanisms M that send each agent t a deterministic message $m_t = m(\xi^{t-1}, \theta^t)$ based on the outcomes ξ_s of agents $s < t$ and types θ_s of agents $s \leq t$.[16] We consider five classes of progressively more powerful mechanisms:

(a) *Networks:* The Designer chooses a network G, and agents learn the outcomes of their neighbors, $\mathcal{M}^{networks} = \{M : m_t = \xi_{G_t}\}$. This coincides with the observation structure in Section 3, albeit with richer outcomes.

(b) *Unknown types:* The Designer does not know agents' types but can release any information based on past agents' outcomes, $\mathcal{M}^{unknown} = \{M : m_t = m(\xi^{t-1})\}$.

(c) *Survey mechanisms:* Motivated by the adjusted harm broadcast, the Designer surveys past agents' types, and can release any information based on past agents' outcomes and types, $\mathcal{M}^{survey} = \{M : m_t = m(\xi^{t-1}, \theta^{t-1})\}$.

(d) *Elicited types:* The Designer asks agent t to report her type $\hat{\theta}_t$ before sending her a message in a way that truth-telling $\hat{\theta}_t = \theta_t$ is incentive compatible, $\mathcal{M}^{elicit} = \{M : m_t = m(\xi^{t-1}, \hat{\theta}^t) \text{ s.t. } \hat{\theta}_t = \theta_t\}$.[17]

(e) *Known types:* The Designer can release any information based on past agents' outcomes and types and the current agent's type $\mathcal{M}^{known} = \{M : m_t = m(\xi^{t-1}, \theta^t)\}$.

Each class gives the Designer more power than the last,

$$\mathcal{M}^{networks} \subset \mathcal{M}^{unknown} \subset \mathcal{M}^{survey} \subset \mathcal{M}^{elicit} \subset \mathcal{M}^{known}.$$

[16] Standard information design would allow for stochastic messages that can also depend on quality, $m_t = m(\xi^{t-1}, \theta^t, q, \epsilon_t)$. We rule out direct dependence on q to stay true to the notion that the planner must learn quality from agents' outcomes, and avoid the trivial result that full transparency about q is first-best in our model of pure informational externalities. Insisting on deterministic messages is for simplicity and avoids distinctions between our subjective probabilities $\Pr_{\theta_t}(q, \theta_{-t}, \xi)$ and objective probabilities about ϵ_t.

[17] In referring explicitly to types θ, these mechanisms are not detail-free. To remain detail-free like the adjusted harm broadcast G^* one could consider mechanisms that survey hypothetical choices c_t based on alternative information instead of surveying types $\mathcal{M}^{hypo} \subset \mathcal{M}^{survey}$, or mechanisms that let agents choose from a menu of signals instead of eliciting their types $\mathcal{M}^{menu} \subset \mathcal{M}^{elicit}$

As in the baseline model, subject to breaking ties in favor of trying the product, any mechanism $M \in \mathcal{M}$ admits a unique equilibrium, and we Pareto-rank mechanisms based on the outcomes of this equilibrium.

Returning to the question whether full transparency about harm is optimal, we say that a class of mechanisms $\mathcal{M}$ satisfies the *Breaking Bad News Improvement Principle (BBN)* if for any $M \in \mathcal{M}$ there exists $M^* \in \mathcal{M}$ where any agent t learns whenever any $s < t$ has been harmed and $M^* \succcurlyeq M$. For example, the construction of the adjusted harm broadcast G^* shows that any mechanism $M \in \mathcal{M}^{networks}$ can be Pareto-improved by some $M^* \in \mathcal{M}^{survey}$. BBN would require this Pareto-improving construction more strongly for every $M \in \mathcal{M}^{survey}$.

Theorem 2. *The five classes have the following properties:*

(a) : *For some* $(\Xi, \Theta, \Pr)$, $\mathcal{M}^{networks}$ *does not satisfy BBN.*

(b) : *For some* $(\Xi, \Theta, \Pr)$, $\mathcal{M}^{unknown}$ *does not satisfy BBN.*

(c) : *For all* $(\Xi, \Theta, \Pr)$, $\mathcal{M}^{survey}$ *satisfies BBN.*

(d) : *For some* $(\Xi, \Theta, \Pr)$, $\mathcal{M}^{elicit}$ *does not satisfy BBN.*

(e) : *For all* $(\Xi, \Theta, \Pr)$, $\mathcal{M}^{known}$ *satisfies BBN.*

Proof. For parts (a) and (b), consider the mechanism M associated with the two-link network $r \leftarrow s \leftarrow t$ from Example 3 and the following two types: Sven's type θ_s knows that (i) he cannot learn from agents before Ruby,[18] and (ii) Ruby consumes, resulting in outcome $\xi_r = \emptyset$ if $q = L$ and $\xi'_r \neq \emptyset, \dagger$ if $q = H$; thus, θ_s learns perfectly in mechanism M.

Tara's type θ_t knows that (i) she cannot learn from agents before Sven, (ii) in mechanism M Ruby does not consume and her outcome $\xi_r = \emptyset$ induces Sven to consume and experience an outcome ξ_s that fully reveals quality, (iii) if Ruby learns that none of her predecessors has been harmed, she consumes and her outcome is ξ'_r, and (iv) if Sven observes outcome ξ'_r he does not consume.

We now argue that a mechanism M^* where Ruby learns whether a predecessor has been harmed cannot make both θ_s and θ_t better off. If the message to Sven m^*_s pools ξ^{s-1}, where all predecessors have outcome $\emptyset$, and $\hat{\xi}^{s-1}$, where Ruby's outcome is switched to ξ'_r, then θ_s learns nothing in M^* and so prefers M. If m^*_s does not pool

[18] E.g. because their types and outcomes are independent of quality.

ξ^{s-1} and $\hat{\xi}^{s-1}$, then by construction θ_t learns nothing in M^* but learns perfectly in M.

For part (c), pick an arbitrary mechanism $M \in \mathcal{M}^{survey}$. To construct $M^* \in \mathcal{M}^{survey}$ that improves on the original mechanism we follow the "adjusted harm broadcast" from Section 4.1. In particular, the Designer tells any agent, Tara, whether harm has taken occurred; if it has not, he additionally sends her the original "hypothetical" message $m_t = m(\xi^{t-1}, \theta^{t-1})$. He can deduce the "hypothetical" outcomes ξ^{t-1} because consumption is higher under the new mechanism, $c_t^* = c(m_t, \phi^*_{\bar{G}_t}, \theta_t) \geq c_t$ when $\phi_s^* \neq \dagger$ for all $s < t$, and he can deduce who consumes in the original mechanism $c_t = c(m_t, \theta_t)$ from the surveyed type θ_t.[19]

For part (d) we show there exists $(\Xi, \Theta, \Pr)$ such that $\mathcal{M}^{elicit}$ does not satisfy BBN. Intuitively, elicitation is a powerful tool that allows the Designer to separate types with different beliefs; broadcasting harm ties the Designer's hands. Specifically, suppose there are events $\{a, b, c, d, \dagger\}$. Type θ_s's beliefs over these events and quality are:

State	a	b	c	d	$\dagger$
$q = H$	1/12	1/12		2/12	
$q = L$	1/12		1/12		6/12

and his indifference belief is $p^*(\theta_s) = 1/2 + \epsilon$. Mechanism M lets Sven choose between signals $\{\{a, b\}, \{c, d, \dagger\}\}$ and $\{\{a, c\}, \{b, d, \dagger\}\}$. Since the latter is uninformative about quality, $\Pr_{\theta_s}(H|\{a, c\}) = \Pr_{\theta_s}(H|\{b, d, \dagger\}) = \Pr_{\theta_s}(H) = 1/3$, he prefers the former signal, which can yield $\Pr_{\theta_s}(H|\{a, b\}) = 2/3$. There is another type θ_s' with beliefs

State	a	b	c	d	$\dagger$
$q = H$	1/4		1/4		
$q = L$		1/4		1/4	

for whom signal $\{\{a, c\}, \{b, d, \dagger\}\}$ constitutes perfect learning.

[19] To appreciate the power of survey mechanism, recall that no mechanism $M^* \in \mathcal{M}^{unknown}$ Pareto-improves on the mechanism M from the proof of parts (a) and (b), since mechanisms that pool ξ^{s-1} and $\hat{\xi}^{s-1}$ make θ_s worse off, and those that separate ξ^{s-1} and $\hat{\xi}^{s-1}$ make θ_t worse off. The Pareto-improving mechanism $M^* \in \mathcal{M}^{survey}$ avoids this dichotomy by separating ξ^{s-1} and $\hat{\xi}^{s-1}$ iff they are separated in M, which requires surveying agents' types or hypothetical consumption.

We claim that there is no mechanism M^* that reveals $\{\dagger\}$ and partitions $\{a, b, c, d\}$ in a way that (i) both types $\{\theta_s, \theta_s'\}$ are better off than in the original M, (ii) Sven has higher consumption for all states, which is required to make future agents no worse off than in M. To see this, first note that M^* must refine $\{\{a, b, c, d\}, \{\dagger\}\}$. To make θ_s' no worse off, M^* must refine $\{\{a, c\}, \{b, d\}, \{\dagger\}\}$. This latter information structure allows θ_s to achieve his favorite decisions, namely consume in $\{b, d\}$ but not in $\{a, c, \dagger\}$. However, this means θ_s no longer consumes in state $\{a\}$, which reduces the learning of some subsequent agents.

Part (e) follows as in part (c). The only difference is that Tara's message in the absence of harm now also depends on her type $m_t(\xi^{t-1}, \theta^t)$. Critically, the Designer can still reconstruct the outcomes of the agents from the original mechanism. □

5 Conclusion

This paper has considered a model in which agents learn about a new product when their neighbors are harmed by low-quality. While this guinea pig effect suggests a strong conflict of interest, full transparency is best for everyone and makes Sven the optimal guinea pig for Tara. This insight generalizes to rich models of incomplete information when an Information Designer can survey agents in order to refine messages given to subsequent agents. We believe that such mechanisms with surveys, which sit between those with unknown types and elicited types, are new to the information design literature and we hope that they will be useful also in other applications.

We have framed our model in terms of learning about product quality. Bad news learning also naturally arises when agents learn about the consequences of potential misconduct. For example, tax payers learn whether a law is enforced, employees learn whether their boss is paying attention, and financial firms learn whether a regulator's inspections are meaningful. To spell out the mapping, a tax payer "tries the product" of tax evasion, a competent tax authority renders tax evasion "low quality", and the sentencing of a tax cheat constitutes "harm". So reinterpreted, the discussion after Corollary 1 means that transparency minimizes expected tax evasion under a competent tax authority.

We conclude our paper by returning to the Pareto-ranking of networks $G \succcurlyeq G'$

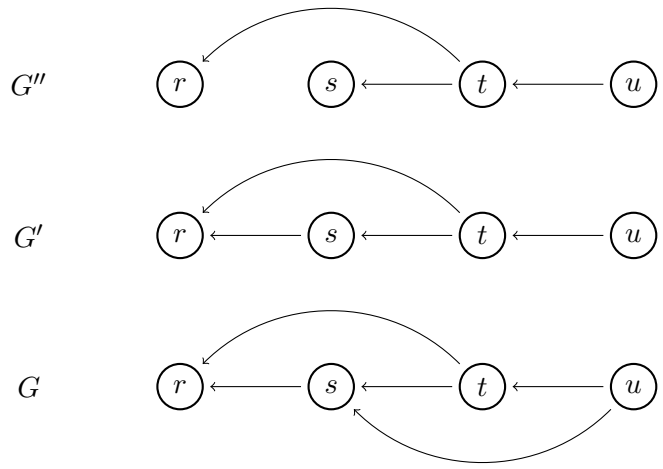

Figure 2: **Networks from Example 4.**

in the baseline model of Section 2. Our main result, Theorem 1, provides a sufficient condition: $G \supseteq G'$ and G is transitive. Conversely, Example 1 provides a necessary condition: Suppose $s \leftarrow t$ in G' and G adds a link $r \leftarrow s$; then $G \succcurlyeq G'$ only if G includes the link $r \leftarrow t$. Example 4 shows that neither condition characterizes $G \succcurlyeq G'$, in that the sufficient condition is not necessary and the necessary condition is not sufficient. We leave a characterization of the Pareto-ranking for future work.

Example 4. (Pareto-ranking): Figure 2 shows three networks, $G \supset G' \supset G''$. First, we show that $G' \not\succ G''$ even though G' includes the link $r \leftarrow t$. By Theorem 1, Tara is more informed in G'. By Example 1, this can lower Ulric's utility. Thus, the necessary condition is not sufficient for Pareto ranking.

Next, we show that $G \succ G''$ even though G is not transitive. By Theorem 1, Ruby, Sven and Tara are better off under G; to show that Ulric is better off under G, consider three cases. First, when $\xi_r = \dagger$, then $c_t = c''_t = 0$ but Ulric learns more in G. Second, when $\xi_s = \dagger$, then Ulric learns perfectly in G. Third, when $\xi_r = \xi_s = \emptyset$, then the proof of Theorem 1 implies $c_t \geq c''_t$; again, Ulric learns more in G. $\triangle$

References

Acemoglu, D., M. A. Dahleh, I. Lobel, and A. Ozdaglar (2011): "Bayesian Learning in Social Networks," *Review of Economic Studies*, 78(4), 1201–1236.

Banerjee, A., A. G. Chandrasekhar, E. Duflo, and M. O. Jackson (2013): "The Diffusion of Microfinance," *Science*, 341(6144), 1236498.

Bergemann, D., and S. Morris (2019): "Information Design: A Unified Perspective," *Journal of Economic Literature*, 57(1), 44–95.

Bikhchandani, S., D. Hirshleifer, and I. Welch (1992): "A Theory of Fads, Fashion, Custom, and Cultural Change as Informational Cascades," *Journal of Political Economy*, 100(5), 992–1026.

Board, S., and M. Meyer-ter-Vehn (2021): "Learning Dynamics in Social Networks," *Econometrica*, 89(6), 2601–2635.

Che, Y.-K., and J. Hörner (2018): "Recommender Systems as Mechanisms for Social Learning," *Quarterly Journal of Economics*, 133(2), 871–925.

Coleman, J., E. Katz, and H. Menzel (1957): "The Diffusion of an Innovation among Physicians," *Sociometry*, 20(4), 253–270.

Dasgupta, P., and E. Maskin (2000): "Efficient Auctions," *Quarterly Journal of Economics*, 115(2), 341–388.

Ely, J. C. (2017): "Beeps," *American Economic Review*, 107(1), 31–53.

Fung, A., M. Graham, and D. Weil (2007): *Full Disclosure: The Perils and Promise of Transparency*. Cambridge University Press.

Galperti, S., and J. Perego (2022): "Information Systems," Discussion paper.

Goeree, J. K., T. R. Palfrey, and B. W. Rogers (2006): "Social Learning with Private and Common Values," *Economic Theory*, 28, 245–264.

Goolsbee, A., and P. J. Klenow (2002): "Evidence on Learning and Network Externalities in the Diffusion of Home Computers," *Journal of Law and Economics*, 45(2), 317–343.

Knoepfle, J., and J. Salmi (2024): "Dynamic Evidence Disclosure: Delay the Good to Accelerate the Bad," .

Kremer, I., Y. Mansour, and M. Perry (2014): "Implementing the "Wisdom of the Crowd"," *Journal of Political Economy*, 122(5), 988 – 1012.

Sgroi, D. (2002): "Optimizing Information in the Herd: Guinea Pigs, Profits, and Welfare," *Games and Economic Behavior*, 39(1), 137–166.

Smith, L., and P. Sørensen (2014): "Rational Social Learning with Random Sampling," Working paper, MIT.

Smith, L., P. Sørensen, and J. Tian (2021): "Informational Herding, Optimal Experimentation, and Contrarianism," *Review of Economic Studies*, 88(5), 2527–2554.

Wilson, R. (1987): *Game-theoretic Analyses of Trading Processes*pp. 33–70, Econometric Society Monographs. Cambridge University Press.

Wolitzky, A. (2018): "Learning from Others' Outcomes," *American Economic Review*, 108(10), 2763–2801.