Bad Networks

Robert Akerlof[*], Richard Holden[†], DJ Thornton[‡]

September 4, 2025

Abstract

There is increasing evidence that social media is detrimental to mental health and self-esteem. A puzzle is why, in spite of this, people join these platforms. One possibility is that people feel trapped: they dislike these networks—in particular, the way in which they encourage self-comparison—but they need to be on them to socialize with peers. We refer to networks where people feel trapped as "bad networks." We model settings with network externalities and show that, surprisingly, bad networks are easy to establish. We also show that networks tend to be both bad and easy to establish when they create rat races—as social networks often do. Amplifying the rat race boosts network size which, while harmful to consumers, may benefit the platform.

Keywords: Social networks, self-comparison, miscoordination.

JEL Codes: D21, D26, D85.

[*] UNSW Business School, email: r.akerlof@unsw.edu.au.

[†] UNSW Business School, email: richard.holden@unsw.edu.au.

[‡] UNSW Business School, email: d.thornton@unsw.edu.au.

1 Introduction

The harmful effects of social media are becoming increasingly hard to ignore. In his recent book, *The Anxious Generation*, Jonathan Haidt argues that social media usage is fueling a mental health crisis among young people. Since 2010, rates of major depression among teens have risen by more than 150 percent, and the share of 8th, 10th, and 12th graders who report being satisfied with themselves has dropped by roughly 10 percentage points.[1] This decline in mental health began precisely when smartphones became widely adopted. Haidt illustrates the crisis through the story of Alexis, who joined Instagram at age 11. At first, she was thrilled, writing in her journal: "On Instagram I reach 127 followers. Ya! Let's put it this way, if I was happy and excited for 10 followers then this is just AMAZING!!!!" But her enthusiasm quickly faded. Her feed soon filled with images of models, dieting advice, and eventually pro-anorexia content promoted by the platform's algorithms. By eighth grade, she was hospitalized for anorexia and depression—struggles that continued throughout her teenage years.

While much of Haidt's evidence is correlational, there is growing causal evidence linking social media to mental health declines. For example, Braghieri et al. (2022) exploit the staggered rollout of Facebook across U.S. college campuses to show that access to the platform increased symptoms of poor mental health, particularly depression. Further evidence on mechanisms suggests these effects stem from Facebook's tendency to foster negative self-comparisons among users.

If social media has such deleterious effects, it raises the question: why are people using these platforms? One answer is that social media may be addictive. Addiction researcher Anna Lembke embraces this view, writing in *Dopamine Nation*, "the smartphone is the modern day hypodermic needle, delivering digital dopamine 24/7 for a wired generation."[2] Supporting this perspective, Allcott et al. (2022) provide causal evidence from a field experiment suggesting that addiction accounts for roughly 31% of social media use. They find that usage drops significantly when users can set limits on their future screen time.

However, another important aspect may be that people feel *trapped*: they dislike these platforms but need to be on them to socialize with their peers. According to

[1] The first statistic is based on data from the U.S. National Survey on Drug Use and Health; the second comes from the Monitoring the Future survey (see Haidt (2024)).

[2] See Lembke (2021), p.1.

this story, people are miscoordinated: they would be better off if they could socialize in another way, but no individual has the power to make that change. Parents seem to perceive this dilemma. As Jonathan Haidt puts it, "Most parents don't want their children to have a phone-based childhood, but somehow the world has reconfigured itself so that any parent who resists is condemning their children to social isolation." A recent survey of college students by Bursztyn et al. (2023) provides more concrete evidence. They find that the average student would need to be paid 59 dollars to get off of TikTok for four weeks. By contrast, the average student *would pay* 28 dollars to have TikTok deactivated for everyone.

We refer to a network as "bad" if it is welfare-reducing (along the lines of the "trapped" story). This paper has two aims. First, we analyze why bad networks arise. One might imagine that such networks are hard to establish—even when they are technically feasible. Why would people flock to a network that they intensely dislike—absent some form of irrationality? We show, perhaps contrary to intuition, that bad networks can get started easily—like parties that people do not wish to attend but feel obligated to when others are going. The second aim of this paper is to identify the features that make networks both bad and easy to establish. We show that this occurs when networks generate *rat races*—as many social networks do.

This paper is organized as follows. Section 2 provides an illustrative example that demonstrates the idea of a bad network, where a large number of agents join a network even though this is welfare-reducing.

Section 3 generalizes this example. It considers a setting where agents face network externalities whether they join a network (parameterized by a) or stay off (parameterized by b). We allow a and b to take arbitrary values. The interesting case arises when $0 > a > b$: the network is unpleasant for those who are on it, but even more unpleasant for those who are off it. "Instigators" get these networks established. These instigators then put pressure on other agents to join, creating a snowball effect. Because agents do not internalize the externalities they inflict—in particular, the pressure they put on other agents to join—these networks grow to suboptimally large sizes ($q^{NE} > q^*$). We also consider potential remedies, such as Pigouvian taxes. While "marginal" policies may be sufficient to induce the socially optimal outcome, more extreme policies are potentially needed to dislodge established networks.

Section 4 then asks whether there are networks with the property that $0 > a > b$. We demonstrate that networks tend to have this feature when they generate rat races. We provide explicit microfoundations for a social network with this property. In the case we consider, agents make two choices: whether to join a social network and whether to exert effort in a rat race. Agents on the network care more about the rat race (i.e. how they compare to others) than agents off the network, which we parameterize by α. We think of α as the extent to which the social network creates concern among agents about social comparison. Networks where α is large tend to have the property that $0 > a > b$. In addition, we show that the size of the network is increasing in α. Thus, amplifying the rat race may be beneficial to a platform even if it is harmful to consumers.

Relative to the existing literature, our contribution is twofold. First, while it is known that agents can miscoordinate on a bad network (see especially Bursztyn et al. (2023), who build a model with this property), existing work has not examined the ease or difficulty with which such networks get established. This paper shows why—perhaps surprisingly—it is easy to establish such networks.[3] Second, while networks with the feature $0 > a > b$ might seem counterintuitive, we show that they arise naturally in many settings. Rat races make networks bad and also create pressure to join.[4]

2 An Illustrative Example

Let us begin with an illustrative example. Consider a setting with a unit mass of agents who simultaneously decide whether to join a network. The utility of agent

[3] The contemporaneous work of Bursztyn et al. (2023) is perhaps the closest paper to ours. They study an environment with negative spillovers to non-users of a network which can lead to what they call "product market traps." In their model, the decentralized (rational expectations) equilibrium need not be unique nor socially optimal. They point out that the *introspective equilibrium* solution concept of Akerlof et al. (2023) permits them to select the bad equilibrium provided there is a large enough fraction of early adopters who want to use the product even when nobody else is using it. In our (Nash, as opposed to introspective) equilibrium it is an arbitrarily small mass of instigators that triggers the unraveling to a bad network involving full participation. By contrast, Bursztyn et al. (2023) require a "large enough" number of early adopters to reach the bad introspective equilibrium with everyone on the network.

[4] The rat race in Section 4 of our paper relates to Tirole (2021) who analyzes a model in which agents care about their image and choose whether to engage in activity in the public or private sphere. He finds that social networks move activity, at a cost, from the private sphere into the public sphere, which is consistent with our microfoundation.

$i \in [0, 1]$ is given by:

$$u(x_i) = \begin{cases} aq, & x_i = 1, \\ bq, & x_i = 0, \end{cases}$$

where $x_i = 1$ ($x_i = 0$) denotes the decision to join (stay off) the network, and $q \in [0, 1]$ is the fraction of agents who join. Network participation generates externalities for participants (captured by parameter a) and non-participants (captured by parameter b). We assume $a > b$, so that participants benefit more from the network than non-participants.

We will examine both the case where $a > 0$ and the case where $a < 0$. The case where $a < 0$ might not seem intuitive. Why would participation in a network generate negative externalities? However, we see such networks as common. In Section 4, we provide microfoundations for such networks and convey an intuition for why they can arise. We argue that, in cases where networks generate rat races—as tends to be true of social networks—negative externalities are endemic.

To begin our analysis, notice that agents strictly prefer to join the network when $q > 0$ and they are indifferent between joining and staying off when $q = 0$. Thus, the game has two Nash equilibria: full participation ($q^{NE} = 1$) and no participation ($q^{NE} = 0$).

The agents' aggregate welfare is given by

$$W(q) = \underbrace{(aq)q}_{\text{benefit to those on the network}} + \underbrace{(bq)(1-q)}_{\text{benefit to those off the network}} .$$

We refer to the network as a "good network" if $a > 0$. It is easy to show that, in this case, the welfare-maximizing value of q, denoted q^*, is equal to 1. We refer to the network as a "bad network" if $a < 0$. In this case, $q^* = 0$. Intuitively, when participation generates positive spillovers ($a > 0$), aggregate welfare rises as more agents join the network—whereas negative spillovers ($a < 0$) make participation socially harmful.

There are many prominent examples of networks that impose relatively small costs to those off the network and have considerable benefits for those on it. These might include services like Google Search, Spotify recommendations, or Tesla Autopark, where each additional user improves the underlying algorithm for everyone. Such networks are ones that we would see as "good." Social media, on the other hand, is a potential example of a bad network. These platforms can generate a "rat race" of social comparison that is welfare-reducing for participants ($a < 0$), while the experience for non-participants is made even worse by social exclusion ($b < a$). As we will discuss further in Section 4, this creates the very conditions for a bad network to thrive.

Putting the above findings together, we conclude that the following types of outcomes are possible.

Good outcomes:

1. Full participation on a good network ($q^{NE} = q^* = 1$) can occur when $a > 0$.
2. No participation on a bad network ($q^{NE} = q^* = 0$) can occur when $a < 0$.

Bad outcomes:

1. No participation on a good network ($q^{NE} = 0$ and $q^* = 1$) can occur when $a > 0$.
2. Full participation on a bad network ($q^{NE} = 1$ and $q^* = 0$) can occur when $a < 0$.

The first type of bad outcome—no participation on a good network—is a well-understood coordination failure, in which agents fail to realize mutual gains from participation.

In contrast, the second type of bad outcome—full participation on a bad network—has received relatively little attention. This second failure can occur when $a < 0$; however, in that case, a good outcome—no participation on a bad network—also remains possible. This raises a key question: when $a < 0$, which outcome is more likely to prevail—the good or the bad?

In the next section, we generalize our analysis to better understand the circumstances where bad outcomes prevail. One might think that bad networks would be difficult to establish, even if they are technically possible. Why would people join a network they strongly dislike? Yet, perhaps counterintuitively, such networks can form quite easily—much like parties that no one wants to attend but feel compelled to join once others start going.

3 Participation in Good and Bad Networks

To generalize the example, assume that the utility of agent $i \in [0,1]$ is given by:

$$u_i = \begin{cases} a\,\varphi(q) + \epsilon_i, & x_i = 1, \\ b\,\varphi(q), & x_i = 0, \end{cases} \tag{1}$$

where $\varphi(0) = 0$, $\varphi(1) = 1$, and $\varphi(q)$ is strictly increasing, twice differentiable, and weakly concave.[5] The ϵ_i's are distributed according to a unimodal pdf $f(\cdot)$ with support $[-c, c]$, where $c \in \overline{\mathbb{R}}$. For ease of exposition, we assume that f is symmetric about 0 and atomless.[6] We again assume that $a > b$.[7] Notice that the example from Section 2 corresponds to the case where $c = 0$ and $\varphi(q) = q$. In Section 4 we provide an explicit microfoundation for preferences of the form given in equation (1) (see Proposition 4).

To solve for the Nash equilibria of the game, notice that agent i prefers to join the network if and only if $\epsilon_i > (b-a)\varphi(q)$. Thus, in equilibrium, the mass of agents who join the network must be equal to $\mathbb{P}(\epsilon_i \geq (b-a)\varphi(q)) = 1 - F((b-a)\varphi(q))$. Hence, the Nash equilibrium must solve the following equation:

$$q^{NE} = 1 - F((b-a)\varphi(q^{NE})) \tag{2}$$

When $c = 0$, there are multiple Nash equilibria. However, when $c > 0$ (in the spirit of a trembling-hand refinement), there is a unique Nash equilibrium with $q^{NE} > \frac{1}{2}$.[8] We state this formally in the following lemma.

Lemma 1. *For all $c > 0$ there is a unique equilibrium. In this equilibrium, $q^{NE} > \frac{1}{2}$.*

To understand the intuition behind this result, let us refer to agents with $\epsilon_i > 0$ as "instigators," agents with $\epsilon_i < 0$ as "resistors," and agents with $\epsilon_i = 0$ as "neutral

[5] Taking $\varphi(1) = 1$ is without loss of generality since we can always rescale a and b. It is natural to assume that social networks have concave network externalities. While early adopters may bring substantial value to the network, network congestion, competition for attention, and over-saturation tend to reduce the marginal value of participation as network size increases.

[6] Atomic distributions over ϵ_i are easily accommodated and yield even sharper results.

[7] The case where $a < b$— although not of much economic interest—is easily handled and leads to similar types of inefficiency.

[8] We thank an anonymous referee for pointing out the similarity between our approach and trembling-hand equilibrium.

agents."[9] Instigators are agents who are inclined to join the network when there are no network participants ($q = 0$), while resistors are agents who are disinclined to join when there are no participants. When $c = 0$, all agents are neutral; but when $c > 0$, there are a combination of instigators and resistors (plus a zero-mass of neutral agents).

Notice that if no agents are on the network initially ($q_0 = 0$), all of the instigators will join. These instigators make up half of the population; thus, q rises to $q_1 = \frac{1}{2}$. When q increases to q_1, some resistors will also join, causing q to rise further: to $q_2 > q_1$. When q increases to q_2, yet more agents will join. The unique Nash equilibrium corresponds to the limit of this process: $q^{NE} = \lim_{n\to\infty} q_n$ (see Figure 1 for an illustration).[10]

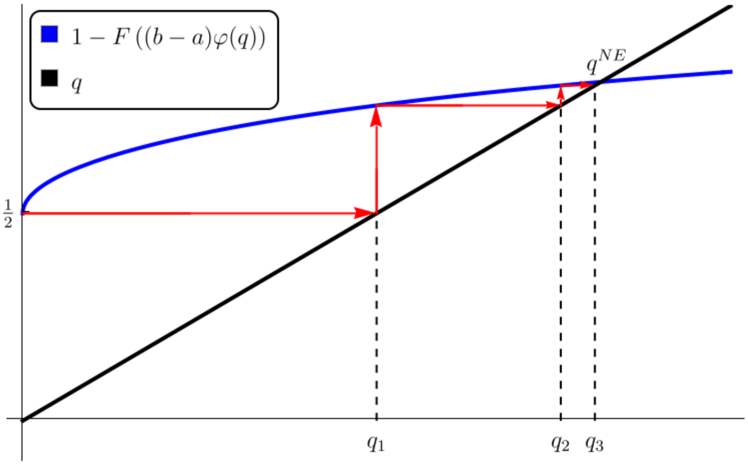

Figure 1: Starting from $q = 0$, all instigators join the network, taking us to $q_1 = \frac{1}{2}$. Because $a > b$, this induces some resistors to join the network, taking us to q_2, and so on until we reach the unique solution q^{NE} of equation (2).

Agents' idiosyncratic benefits/costs (ϵ_i) have an impact on the welfare analysis. Letting $\bar{\epsilon}(q) = F^{-1}(1-q)$, aggregate welfare is given by:[11]

$$W(q) = (a-b)q\varphi(q) + b\varphi(q) + \mathbb{E}\big(\epsilon_i \colon \epsilon_i > \bar{\epsilon}(q)\big). \tag{3}$$

[9] The term instigator is used by Granovetter (1978) to describe agents who have a "0% threshold" for taking an action—that is, agents who are willing to join a network in the absence of anyone else joining.

[10] Note that if consumers not only have heterogeneous preferences over joining the network (ϵ_i) but the network externalities (a and b) are also heterogeneous across consumers, there might be multiple equilibria.

[11] The notation $\mathbb{E}\big(\epsilon_i \colon \epsilon_i > \bar{\epsilon}\big)$ is equivalent to $\mathbb{E}\big(\epsilon_i \mathbb{1}_{\{\epsilon_i > \bar{\epsilon}(q)\}}\big)$ where $\mathbb{1}$ is an indicator function.

Let q^* denote the value of q that maximizes aggregate welfare. As in the illustrative example, when $c = 0$, so that $\epsilon_i = 0$ for all i, $q^* = 0$ if $a < 0$ and $q^* = 1$ if $a > 0$. However, q^* might take a value between 0 and 1 if $c > 0$. For instance, suppose there is one set of agents with ϵ_i large and positive and a second set with ϵ_i large and negative. It might be optimal to have the first set join the network and the second set stay off the network.

As before, we will use the terms "good network" and "bad network" to refer, respectively, to networks that exhibit positive externalities $(a > 0)$ and negative externalities $(a < 0)$. The following proposition compares the equilibrium level of network participation to the socially optimal level for good and bad networks.

Proposition 1.

1. *For good networks ($a > 0$), too few agents join the network relative to the social optimum ($q^{NE} < q^*$).*

2. *For bad networks ($a < 0$), too many agents join relative to the social optimum ($q^* < q^{NE}$).*

Intuitively, instigators get a bad network started. Other agents then join the bad network—even though they dislike it—because it is even worse to be off the network ($b < a$). It is like a party that people find unpleasant but feel obliged to attend. As people join the network, they both make the network more unpleasant *and* increase the pressure to join. That is, they make it a party where attendance is more obligatory. This externality leads to suboptimally high rates of network participation. The following corollary naturally follows.

Corollary 1. *For bad networks ($a < 0$), there is a positive mass of agents on the network who would be better off if the network did not exist.*

Proposition 1 shows that, for all bad networks, too many agents join relative to the social optimum. An extreme case—that can arise—is one where *all* agents join the network even though it is optimal to have *no* agents join. The following proposition provides conditions under which we see this outcome.

Proposition 2.

1. *If $c \leq a - b$, all agents join the network in equilibrium ($q^{NE} = 1$).*

2. *If $c < -a$, all agents are better off if there is no network ($q = 0$) than if there is a network ($q > 0$). This implies, moreover, that $q^* = 0$.*

Intuitively, for a bad network where c is small, resistors are not too resistant to joining the network. Thus, when instigators join the network, they create a snowball effect whereby all of the resistors join as well (hence, $q^{NE} = 1$). Moreover, for a bad network where c is small, agents' idiosyncratic tastes (ϵ_i's) are not very important from a welfare standpoint. The negative network externalities associated with having agents join are the dominant welfare consideration. Thus, $q^* = 0$.

Discussion

Propositions 1 and 2 explain why networks which are both socially undesirable and harmful to users can nonetheless sustain large amounts of participation in equilibrium. Such "bad networks" are remarkably easy to get going: even a small mass of instigators can trigger a cascade in which the pressure to join overwhelms any idiosyncratic dislike for the network. In the extreme, all agents may join the network and yet prefer that it did not exist.

Our findings are supported by empirical evidence. Bursztyn et al. (2023) report that the average student prefers to be on TikTok. They would need to be paid 59 dollars to get off of it for four weeks. However, they would *be willing to pay* 28 dollars to have TikTok deactivated for everyone. In this sense, these students are miscoordinated— trapped on a bad network (in line with Corollary 1).

Internal company research at Meta points to the same conclusion. As reported in the Wells et al. (2021) coverage of the "Facebook Papers," Meta's own analyses acknowledged that Instagram worsens body image issues for one in three teenage girls and that users themselves blamed the platform for increases in anxiety and depression. In our framework, this is a real-world instance of a bad network: widespread participation persists despite evidence of harm to many users. In Section 4, we demonstrate that social media platforms have incentives to exacerbate the harmful effects of their networks.

3.1 Policy

Of course, there are tools for correcting market failures.[12] It is natural to consider Pigouvian taxation as a potential remedy since the inefficiencies in the market arise due to externalities.[13] Here, we show that Pigouvian taxation may restore efficiency; however there are cases where it does not work.

Let τ denote the tax each agent pays when they join the network. Agent i's utility becomes:

$$u_i = \begin{cases} a\,\varphi(q) + \epsilon_i - \tau, & x_i = 1, \\ b\,\varphi(q), & x_i = 0, \end{cases}$$

Aggregate welfare is given by:

$$W^{\text{tax}}(q,\tau) = \underbrace{\int_{i=0}^{1} u_i di}_{\text{agents' utility}} + \underbrace{q \cdot \tau}_{\text{tax revenue}} .$$

A Pigouvian tax charges each agent based on the marginal externality they inflict:

$$\tau^P(q) = -(aq + b(1-q))\varphi'(q)$$

Suppose that, before agents choose whether to join the network, a social planner announces that the tax on the network will be $\tau^P(q^*)$, where q^* is the socially optimal level of network participation.[14] Does this tax maximize aggregate welfare?[15]

Proposition 3 gives conditions under which the Pigouvian tax induces the welfare-maximizing outcome q^*.

Proposition 3. *If $\varphi(q) = q$, a Pigouvian tax $\tau^P(q^*)$ on a bad network ($0 > a > b$) induces the welfare-maximizing outcome q^*.*

[12] There are also tools for addressing market failures arising from behavioral biases (e.g., see Bernheim and Taubinsky (2018)). Although we focus on purely rational agents, it would be an interesting question for future research to investigate the implications of an extension of our model which accounts for well-known behavioral biases or irrationalities.

[13] In this section we focus on bad networks, but the optimal policy for good networks mirrors Proposition 3 but with subsidies rather than taxes.

[14] If $q^* = 0$ it is without loss of generality to set $\tau(q^*) = +\infty$, which one can think of as "banning the network."

[15] We thank the editor for highlighting the implications of our model for optimal taxation.

The Pigouvian tax always induces a Nash equilibrium in which $q = q^*$. However, additional equilibria may also exist, so the Pigouvian tax does not necessarily guarantee the welfare-maximizing outcome. The proposition shows that the equilibrium with $q = q^*$ is unique, though, when $\varphi(q) = q$.

Discussion.
There are cases where we cannot rule out the possibility of multiple equilibria. This can occur, for example, when $\varphi(q) \neq q$ or the distribution of ϵ_i's is not unimodal. In these cases, the Pigouvian tax might induce an equilibrium $q' \neq q^*$.

One possibility is that $q' > q^*$. Intuitively, if a bad network has already formed, the Pigouvian tax—which is only calibrated to influence the marginal consumer properly—may not be sufficient to kill it. In such an instance, the optimal policy might involve first imposing an extreme tax (i.e. greater than $\tau^P(q^*)$) in order to kill the bad, focal equilibrium; once the network has been depleted, it might then be optimal to lower the tax to $\tau^P(q^*)$.[16]

On the flip side, the Pigouvian tax might induce an equilibrium $q' < q^*$. Here, the Pigouvian tax may not be sufficient to establish a network if a network has not already formed. In such an instance, the optimal policy might involve first imposing a low tax (i.e. lower than $\tau^P(q^*)$) in order to encourage the formation of a network; once the network has formed, it might then be optimal to raise the tax to $\tau^P(q^*)$.

4 When are Social Networks Bad?

In Section 3, we showed that networks get going easily when networks are bad ($0 > a > b$)—perhaps contrary to intuition. A remaining question is whether there are networks, such as social networks, that are prone towards being bad.

Here, we discuss a force that we see as important in making networks bad: rat races. Many networks generate competition between agents. Social networks, for instance, tend to generate competition by making users more aware of how they compare to one another. Agents may join the network because they feel the need to participate in the competition; however they may prefer to have no network, so the competition can be avoided. Here, we provide microfoundations for this idea,

[16] A formal analysis of optimal dynamic taxation is beyond our present scope, but we regard it as a promising direction for future research.

focusing on the case of social networks. We also show why it may be in the interest of platforms to promote competition—despite its negative consequences.

4.1 Model

Suppose there is a unit mass of agents and each decides whether to join a social network ($x_i = 0$ or 1) and whether to exert effort ($e_i = 0$ or 1). The effort the agent exerts is part of a zero-sum competition for esteem. We denote the agent's outcome in the esteem competition by $R_i \in [-1, 1]$ and refer to R_i as the agent's rank, where 1 denotes the highest rank and -1 denotes the lowest rank.

An agent's rank is determined by a combination of effort and luck (we assume, for simplicity, that agents have the same ability). Let q_e denote the fraction of agents who exert effort. We assume that the expected rank of an agent i who exerts effort e_i is $R(e_i, q_e) = e_i - q_e$. This assumption ensures that the esteem competition is zero-sum: for any q_e, $\int_0^1 R(e_i, q_e)\, di = 0$.[17]

Agent i is risk neutral and has a utility function which depends on the fraction of agents who exert effort (q_e) and the fraction of agents on the network (q_x):

$$U(x_i, e_i, q_e, q_x) = \underbrace{(1 + \alpha \cdot x_i)R(e_i, q_e)}_{\text{Esteem Component}} \quad + \quad \underbrace{\beta \cdot q_x \cdot x_i}_{\text{Connection Component}} \quad - \quad \underbrace{C \cdot e_i.}_{\text{Cost of Effort}} \quad + \epsilon_i \cdot x_i \tag{4}$$

The first component—the "esteem component"—captures the agent's concern about their rank (i.e. how they compare to others). The weight agents put on esteem depends upon whether they are on or off the social network. Parameter $\alpha \geq 0$ denotes the additional weight agents put on esteem when they are on the network. This captures the idea that social networks make self-comparisons more salient.[18,19]

The second component—the "connection component"—reflects the benefit agents

[17] We do not need, for our purposes, to pin down the exact distributions over ranks.

[18] In related theoretical work, Iyer and Katona (2016) consider a setting where intensifying competition among users for visibility can have negative effects on their welfare.

[19] It might be natural to assume that the salience of comparisons (α) is growing with the size of the network (q_x) as well. Our model easily accommodates this consideration, however it introduces an additional component in the payoff from joining the network. The resulting model can have multiple Nash equilibria, but qualitatively our results remain the same as every one of these equilibria is inefficient, so for simplicity we focus on a fixed increase in the size of salience, which induces a unique equilibrium.

on the network obtain from being able to connect with peers. We assume $\beta > 0$.

The third component of the utility function is the cost of exerting effort. We assume that $C > 1$, which ensures that agents who do not join the social network ($x_i = 0$) do not find it worthwhile to exert effort. This is a simple way of capturing the idea that agents who are off the network are less motivated to participate in the rat race for esteem.

The final component (ϵ_i) is agent i's idiosyncratic preference for joining the network.

4.2 Analysis

We separate our analysis into the case where esteem has low salience for agents on the network ($\alpha < C - 1$) and the case where it has high salience ($\alpha \geq C - 1$).

Case 1: Esteem has low salience on the network ($\alpha < C - 1$)

When α is low, for agents on the network, the returns to effort ($1+\alpha$), do not exceed the cost of effort (C), so $e_i = 0$. Similarly, for agents off the network, the returns to effort (1) do not exceed the cost (C), so $e_i = 0$. It follows that $q_e = 0$ ($e_i = 0$ for all i) and $R_i = 0$ for all i. Thus, the expected utility of agent i is given by:

$$\begin{aligned}\mathbb{E}(U_i) &= (1 + \alpha \cdot x_i) R_i(e_i, q_e) + \beta \cdot q_x \cdot x_i - C \cdot e_i + \epsilon_i \cdot x_i \\ &= \beta \cdot q_x \cdot x_i + \epsilon_i \cdot x_i\end{aligned}$$

We can rewrite the expected utility function as follows:

$$\mathbb{E}(U_i) = \begin{cases} \beta \cdot q_x + \epsilon_i, & x_i = 1, \\ 0, & x_i = 0, \end{cases} \tag{5}$$

This corresponds to the model in Section 3 with $a = \beta > 0$, $b = 0$, and $\varphi(q) = q$.

Notice that this network is a "good network": $a > 0$ and $a > b$. Intuitively, the network does not generate a rat race so its only function (connecting peers) is a positive one.

Case 2: Esteem has high salience on the network ($\alpha \geq C - 1$)

When α is high, for agents on the network, the returns to effort ($1 + \alpha$), exceed the cost of effort (C), so $e_i = 1$. For agents off the network, the returns to effort

(1) do not exceed the cost (C), so $e_i = 0$. It follows that $q_e = q_x$, $e_i = x_i$, and $R_i = e_i - q_x = x_i - q_x$. Thus, the expected utility of agent i is given by:

$$\begin{aligned}\mathbb{E}(U_i) &= (1 + \alpha \cdot x_i) R_i(e_i, q_e) + \beta \cdot q_x \cdot x_i - C \cdot e_i + \epsilon_i \cdot x_i \\ &= (1 + \alpha \cdot x_i)(x_i - q_x) + \beta \cdot q_x \cdot x_i - C \cdot x_i + \epsilon_i \cdot x_i\end{aligned}$$

We can rewrite the expected utility function as follows:

$$\mathbb{E}(U_i) = \begin{cases} (\beta - \alpha - 1) q_x + (\alpha - (C - 1)) + \epsilon_i, & x_i = 1, \\ -q_x, & x_i = 0, \end{cases} \tag{6}$$

This exactly corresponds to the model in Section 3—with $a = \beta - \alpha - 1$, $b = -1$, and $\varphi(q) = q$—provided $\mathbb{E}((\alpha - (C - 1)) + \epsilon_i) = 0$. It is a "bad network" ($0 > a > b$) if, additionally, $\beta - 1 < \alpha < \beta$. Intuitively, the negative aspect of the network (the rat race) outweighs the positive aspect of the network (connecting peers). Without the normalized expectation, equation (6) corresponds to the model from Section 3 but with an additional constant term.[20]

The following proposition summarizes.

Proposition 4.

1. *If esteem has low salience for network participants ($\alpha < C - 1$), the network is a good network.*

2. *If esteem has high salience for network participants ($\alpha \geq C - 1$), the network is a bad network if $\mathbb{E}((\alpha - (C - 1)) + \epsilon_i) = 0$ and $\beta - 1 < \alpha < \beta$.*

The salience of esteem α might be a strategic choice variable for a platform. We might ask how increasing the salience of esteem affects the overall size of the network (q_x). From equation (5), we see that when salience is low ($\alpha < C - 1$), increasing salience has no effect on the network's size.

[20] The model can easily be modified so that the social network not only reduces agents' utility but also their esteem. Suppose agents who stay off the network are able to hold motivated beliefs about their rank because they lack information about how they compare. We can model this in simple terms by assuming agent i's esteem is boosted by γ if they stay off the network. With this modification, the network lowers esteem since it prevents agents from holding motivated beliefs.

However, equation (6) shows that when salience is high ($\alpha \geq C - 1$), agent i's desire to join the network is increasing in α. Intuitively, increasing α makes the rat race more intense, which puts more pressure on agents to join the network and participate in the rat race. Because we cannot rule out the possibility of multiple equilibria, we focus on the effect of α in the largest equilibrium. In the largest equilibrium, increasing salience (α) increases the network size (q_x). The following proposition summarizes.

Proposition 5.

1. *When the salience of esteem is low ($\alpha < C - 1$), raising salience has no effect on network size (q_x).*

2. *When the salience of esteem is high ($\alpha \geq C - 1$), raising salience weakly increases network size (q_x) in the largest equilibrium.*

4.3 Social Comparison on Platforms

Proposition 5 suggests that a platform might try to increase the social-comparison aspects of its network (α) as a way of driving participation. There are a variety of design choices media platforms make that could raise α. Examples include prominently displaying engagement metrics such as "likes," shares, and follower counts, and algorithmic feeds prioritizing content that performs well according to these metrics. Experimental evidence shows that exposure to "upward comparison" (content—profiles depicting more attractive lifestyles, higher social activity, or healthier habits) lowers users' self-esteem (Vogel et al., 2014). By curating feeds to highlight such content, platforms make social comparison more salient (increase α), intensifying the competitive pressures in our model.

Recent empirical work suggests that platforms indeed have an incentive to promote these harmful design features. For instance, in a large-scale field experiment on Facebook, Twitter, and YouTube, Beknazar-Yuzbashev et al. (2025) find that reducing exposure to "toxic content" significantly lowers time spent on these platforms, as well as advertising impressions.

Several high-profile cases suggest that platforms recognize these harms yet preserve these features anyway. In 2019–2021, Instagram ran a global experiment hiding public "like" counts, with the stated aim "to make it less of a competition" (Booker, 2019). Independent evidence from Wallace and Buil (2021) and others shows that removing visible "likes" reduced negative affect and loneliness, consistent with lowering α. The change received positive user feedback, but Instagram ultimately made it optional rather than the default—maintaining the competitive pressure that fuels engagement.

Taken together, the evidence points to a structural misalignment, where features that cause widespread harm also make social networks more profitable.

5 Conclusion

There is significant evidence that social networks, despite their popularity, are harmful to users. In this paper, we ask why such networks arise in the first place, and what features make them "bad."

We show that networks with the feature $0 > a > b$ are not only harmful *if they get established* but also get established *easily*. Effectively, these are parties that people do not like to attend but feel more and more pressure to attend as others choose to do so. A few "instigators" is all it takes to get such networks started.

While networks with the feature $0 > a > b$ might seem counterintuitive, we argue that they arise naturally in many settings. Rat races make networks bad—yet they also create pressure to join. We argue that rat races are a pervasive feature of social networks. Moreover, amplifying the rat-race nature of social networks boosts network size which, while harmful to consumers, may benefit the platforms.

This paper (see Proposition 3) suggests that traditional policies, such as Pigouvian taxation, can serve as helpful remedies. However, once networks are established, "marginal" policies may be insufficient to induce socially optimal outcomes.

References

Akerlof, Robert, Richard Holden, and Luis Rayo, "Network externalities and market dominance," *Management Science*, 2023.

Allcott, Hunt, Matthew Gentzkow, and Lena Song, "Digital addiction," *American Economic Review*, 2022, *112* (7), 2424–2463.

Beknazar-Yuzbashev, George Jiménez-Durán, Rafael McCrosky, Jesse Stalinski, and Mateusz, "Toxic Content and User Engagement on Social Media: Evidence from a Field Experiment," Working Paper 11644, CESifo GmbH, Munich 2025.

Bernheim, B. Douglas and Dmitry Taubinsky, "Chapter 5 - Behavioral Public Economics," in "Handbook of Behavioral Economics: Applications and Foundations 1," Vol. 1, North-Holland, 2018, pp. 381–516.

Booker, Brakkton, "Instagram Will Test Hiding 'Likes' On Some U.S. Accounts Starting Next Week," NPR November 2019.

Braghieri, Luca, Ro'ee Levy, and Alexey Makarin, "Social media and mental health," *American Economic Review*, 2022, *112* (11), 3660–3693.

Bursztyn, Leonardo, Benjamin R. Handel, Rafael Jimenez, and Christopher Roth, "When Product Markets Become Collective Traps: The Case of Social Media," Working Paper 31771, National Bureau of Economic Research 2023.

Granovetter, M., "Threshold models of collective behavior," *American Journal of Sociology*, 1978, *83*, 1420–1443.

Haidt, Jonathan, "End the phone-based childhood now," *The Atlantic*, 2024, *March 13.*

Iyer, Ganesh and Zsolt Katona, "Competing for Attention in Social Communication Markets," *Management Science*, August 2016, *62* (8), 2149–2455.

Lembke, Anna, *Dopamine nation: Finding balance in the age of indulgence*, Penguin, 2021.

Tirole, Jean, "Digital dystopia," *American Economic Review*, 2021, *111* (6), 2007–2048.

Vogel, Erin A, Jason P Rose, Lindsay R Roberts, and Katheryn Eckles, "Social comparison, social media, and self-esteem.," *Psychology of popular media culture*, 2014, *3* (4), 206.

Wallace, Elaine and Isabel Buil, "Hiding Instagram Likes: Effects on negative affect and loneliness," *Personality and Individual Differences*, February 2021, *170*, 110509.

Wells, Georgia, Jeff Horwitz, and Deepa Seetharaman, "Facebook Knows Instagram Is Toxic for Teen Girls, Company Documents Show," *The Wall Street Journal*, September 2021.

6 Appendix: Proofs

6.1 Proof of Proposition 1

To begin, we prove a Lemma.

Lemma 2. *For any* $q \in (0,1)$,

$$\frac{d}{dq}\mathbb{E}\big(\epsilon_i : \ \epsilon_i > F^{-1}(1-q)\big) = F^{-1}(1-q).$$

Proof. Letting $\bar{\epsilon} = F^{-1}(1-q)$, we can write

$$\mathbb{E}\big(\epsilon_i : \ \epsilon_i > \bar{\epsilon}\big) = \int_{\bar{\epsilon}}^{c} \epsilon f(\epsilon)\, d\epsilon.$$

So we have

$$\frac{d\bar{\epsilon}}{dq} = -\frac{1}{f(F^{-1}(1-q))} = -\frac{1}{f(\bar{\epsilon})}.$$

Hence by Leibniz integral rule,

$$\frac{d\left(\mathbb{E}\big(\epsilon_i : \ \epsilon_i > \bar{\epsilon}\big)\right)}{dq} = -\bar{\epsilon} f(\bar{\epsilon})\frac{d\bar{\epsilon}}{dq} = \bar{\epsilon},$$

which completes the proof. □

We now prove the proposition.

Proof. Recall that welfare is given by

$$W(q) = a\varphi(q)q + b\varphi(q)(1-q) + \mathbb{E}(\epsilon_i : \ \epsilon_i > \bar{\epsilon}).$$

To begin, suppose $a < 0$. Then welfare cannot be maximized at 1 since $W(0) = 0 > a = W(1)$. So either welfare is maximized at 0, or welfare is maximized at an interior point. For now, suppose that the optimum is interior. The first derivative of welfare is given by

$$\begin{aligned} W'(q) &= a(q\varphi'(q) + \varphi(q)) + b((1-q)\varphi'(q) - \varphi(q)) + F^{-1}(1-q) \\ &= (a-b)\varphi(q) + (aq + b(1-q))\varphi'(q) + F^{-1}(1-q) \end{aligned}$$

where the $F^{-1}(1-q)$ term comes from Lemma 2.

Now, in any interior *equilibrium*, equation (2) implies that

$$(a-b)\varphi(q^{NE}) + F^{-1}(1-q^{NE}) = 0.$$

So it follows that

$$\begin{aligned} W'(q^{NE}) &= (a-b)\varphi(q^{NE}) + (aq^{NE} + b(1-q^{NE}))\varphi'(q^{NE}) + F^{-1}(1-q^{NE}) \\ &= (aq^{NE} + b(1-q^{NE}))\varphi'(q^{NE}) \qquad (7) \\ &< 0, \end{aligned}$$

where the inequality comes from the fact that $q^{NE} \in (0,1) \implies \varphi'(q^{NE}) > 0$ and $0 > a > b$ implies $aq + b(1-q) < 0$ for all q. So in any interior equilibrium, welfare can be improved by decreasing the number of agents on the network. It remains to show that there cannot be some other "global maximum" of the welfare function at $q^* > q^{NE}$. Since the Nash equilibrium is unique, for any $q > q^{NE}$ we must have $(a-b)\varphi(q) + F^{-1}(1-q) < 0$, and therefore

$$W'(q) = \underbrace{(a-b)\varphi(q) + F^{-1}(1-q)}_{<0} + \underbrace{(aq + b(1-q))\varphi'(q)}_{<0} < 0,$$

which shows that the *global* maximum q^* to the planner's problem can never be larger than q^{NE}. Hence if q^* is interior then $q^* < q^{NE}$. Finally, if $q^* = 0$ then clearly the NE is still too large since $q^{NE} > 0$ (in particular, $q^{NE} > \frac{1}{2}$). This proves Proposition 1 for bad networks.

Now suppose $a > b > 0$ so the network is good. Then by exactly the same reasoning, the slope of welfare at the NE is given by equation (7), which, for $a, b > 0$ is strictly positive. Moreover, since $(a-b)\varphi(q) > F^{-1}(1-q)$ for $q < q^{NE}$ it follows that the smallest solution to the planner's problem must be larger than q^{NE}, since in particular

$$(a-b)\varphi(q) + (aq + b(1-q))\varphi'(q) > (a-b)\varphi(q) > -F^{-1}(1-q).$$

That is, $W'(q) > 0$ for $q < q^{NE}$. Hence the welfare maximizing quantity is always larger than q^{NE}, which completes the proof of the proposition. □

6.2 Proof of Corollary 1

Proof. The corollary is immediate from the argument preceding Lemma 1. In particular, since $q^{NE} > \frac{1}{2}$, but the mass of agents with $\epsilon_i > 0$ is $\frac{1}{2}$, any agent with $\epsilon_i < 0$ who joins the network is strictly worse off than if the network never existed. These agents receive a strictly negative payoff in equilibrium but would receive 0 if $q = 0$. □

6.3 Proof of Proposition 2

We begin by proving part 1. of the proposition.

Proof. As argued in the text of Section 3 and depicted in Figure 1, strictly more than $\frac{1}{2}$ of all agents must join the network in any NE. This is because it is a dominant strategy to join for all agents with $\epsilon_i > 0$, which leads at least some agents with $\epsilon_i < 0$ to join. Note that the equilibrium condition can be written as

$$(a-b)\varphi(q^{NE}) = -F^{-1}(1-q^{NE}). \tag{8}$$

Since $F^{-1}(1-q^{NE}) < F^{-1}(\frac{1}{2}) = 0$, (as f is symmetrically distributed around 0), it follows that at any q satisfying equation (8), the RHS $-F^{-1}(1-q^{NE})$ is strictly positive. But for $q > \frac{1}{2}$ the RHS is also strictly convex. Indeed, it is easily shown that

$$-\frac{d^2}{dq^2}F^{-1}(1-q) = \frac{f'(F^{-1}(1-q))}{[f(F^{-1}(1-q))]^3}.$$

So for any $q > \frac{1}{2}$, we have $F^{-1}(1-q) < 0 \implies f'(F^{-1}(1-q)) > 0$, and the denominator is always positive, thus we conclude $-\frac{d^2}{dq^2}F^{-1}(1-q) > 0$. On the other hand, the LHS $(a-b)\varphi(q^{NE})$ is weakly concave by assumption, and $\varphi(0) = 0$.

Taken together, this implies that there is a unique equilibrium— a weakly concave and strictly convex function on $[\frac{1}{2}, 1]$ can have at most one intersection. Hence either there is an intersection at a point strictly less than 1, or else $-F^{-1}(0) = c \leq (a-b)$, in which case the unique equilibrium is $q^{NE} = 1$, which proves part 1. of Proposition 2. □

Before we prove part 2. of the proposition, it is convenient to prove the following Lemma.

Lemma 3. *For all* $q \in [0,1]$,

$$\mathbb{E}\big(\epsilon_i \colon \epsilon_i > F^{-1}(1-q)\big) \leq cq. \tag{9}$$

Proof. For notational simplicity, let $\epsilon_q = F^{-1}(1-q)$ and define

$$T(q) \equiv \int_{\epsilon_q}^{c} \epsilon f(\epsilon)\, d\epsilon.$$

Then since $\epsilon \leq c$ over the range of integration,

$$T(q) \leq \int_{\epsilon_q}^{c} c f(\epsilon)\, d\epsilon = c(F(c) - F(\epsilon_q)) = cq.$$

□

We now prove part 2. of the proposition.

Proof. First, since φ is concave with $\varphi(0) = 0$ and $\varphi(1) = 1$, observe that we have the elementary bounds

$$q \leq \varphi(q) \leq 1. \tag{10}$$

Using the lower bound and the fact that $a, b < 0$, we have

$$a\varphi(q)q + b\varphi(q)(1-q) \leq aq^2 + bq(1-q) = (a-b)q^2 + bq. \tag{11}$$

Combining equation (11) with Lemma 3 gives

$$W(q) \leq (a-b)q^2 + bq + cq.$$

Since $c < -a$, we have

$$\begin{aligned} W(q) &\leq (a-b)q^2 + bq - aq \\ &= -(a-b)q(1-q). \end{aligned}$$

This proves that $W(q) < 0$ for all $q \in (0,1)$. Finally, $W(1) = a < 0$, hence $W(q) < 0$ for all $q > 0$ and so $q^* = 0$ is uniquely optimal for the planner. Since an agent with the largest possible idiosyncratic benefit c from joining the network receives $-a + c < 0$ in the Nash equilibrium, it follows that all agents are worse off on the network in equilibrium. □

6.4 Proof of Proposition 3

Proof. Define

$$B(q) \equiv (a-b)q + F^{-1}(1-q),$$

With $\varphi(q) = q$ and a constant tax τ, agent i joins the network if

$$\epsilon_i \geq (b-a)q + \tau,$$

hence the equilibrium condition is

$$q = 1 - F\big((b-a)q + \tau\big),$$

which is equivalent to $B(q) = \tau$. Suppose the planner's optimal participation $q^* \in (0,1)$ is interior. Then the constant (Pigouvian) tax is given by

$$\tau^* = \tau^P(q^*) \;\equiv\; -\big(aq^* + b(1-q^*)\big) \;>\; 0.$$

Step 1 (Quantile–tail identity and concavity). Define

$$H(q) \;\equiv\; \mathbb{E}[\epsilon_i \colon \epsilon_i > F^{-1}(1-q)] \;=\; \int_{F^{-1}(1-q)}^{c} x\, f(x)\, dx.$$

We know from lemma 2 that

$$H'(q) = F^{-1}(1-q), \qquad H''(q) = -\frac{1}{f\big(F^{-1}(1-q)\big)} \ < \ 0,$$

so H is strictly concave on $(0,1)$.

With $\varphi(q) = q$, welfare is

$$W(q) = (a-b)q^2 + b\,q + H(q).$$

Recall that the planner's optimum can never be $q^* = 1$, since $a = W(1) < W(0) = 0$. Moreover, if the planner's optimum is $q^* = 0$, the optimal policy bans the network, $\tau^* = +\infty$, and the unique equilibrium is $q = 0$, so the result is immediate. Henceforth assume $q^* \in (0,1)$. Then the first-order condition (FOC) is

$$W'(q^*) = 0 \quad \iff \quad 2(a-b)q^* + b + H'(q^*) = 0. \tag{12}$$

Step 2 (Equilibrium at the optimal tax). Under the Pigouvian tax τ^*, the equilibrium condition for any $q \in (0,1)$ is $B(q) = \tau^*$. Since $\tau^* = B(q^*)$, this implies

$$(a-b)q + H'(q) \ = \ (a-b)q^* + H'(q^*). \tag{13}$$

Combining (13) with the FOC (12) yields the identity

$$(a-b)(q+q^*) + H'(q) + b \ = \ 0, \tag{14}$$

which holds for every equilibrium q.

Step 3 (Strict concavity gives a strict lower bound on $W(q) - W(q^*)$). Suppose for a contradiction that $q \neq q^*$ is an equilibrium under the tax τ^* (i.e. q satisfies (13)). By strict concavity of H,

$$H(q) - H(q^*) \ > \ H'(q)\,(q - q^*).$$

Therefore

$$\begin{aligned} W(q) - W(q^*) = &\ (a-b)\big(q^2 - q^{*2}\big) + b\,(q - q^*) + \big(H(q) - H(q^*)\big) \\ > &\ (a-b)\big(q^2 - q^{*2}\big) + b\,(q - q^*) + H'(q)\,(q - q^*) \\ = &\ (q - q^*)\Big[(a-b)(q + q^*) + b + H'(q)\Big] \\ = &\ 0, \end{aligned}$$

where the final equality follows from (14). Hence

$$W(q) - W(q^*) \ > \ 0,$$

which contradicts the optimality of q^* for the planner. Therefore no $q \neq q^*$ can solve (13), i.e. $B(q) = \tau$ admits the unique solution $q = q^*$. □

6.5 Proof of Proposition 4

Proof. Proposition 4 is proved in the text of Section 4. □

6.6 Proof of Proposition 5

Proof. First, suppose $\alpha < C - 1$. Then by equation (5), utility does not depend on α. Hence raising the salience has no effect on incentives, and therefore on the network size q_x.

Now suppose $\alpha \geq C - 1$. Then by equation (6), agent i joins the network when

$$\epsilon_i \geq -\beta q_x - \alpha(1 - q_x) + C - 1.$$

Since the RHS is strictly decreasing in α, the probability $\mathbb{P}(\epsilon_i > -\beta q_x - \alpha(1 - q_x) + C - 1)$ is weakly increasing in α. Hence the largest intersection of the line q_x with $\mathbb{P}(\epsilon_i > -\beta q_x - \alpha(1 - q_x) + C - 1)$ is also weakly increasing in α. Therefore, the largest equilibrium network size q_x (which is defined by the largest solution to $q_x = \mathbb{P}(\epsilon_i > -\beta q_x - \alpha(1 - q_x) + C - 1))$ is weakly increasing in α, as claimed. □