Bad Networks

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

May 23, 2024

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 develop an explicit model of social networks that provides micro-foundations for why they may be bad; and we show that social media platforms have an incentive to increase the self-comparison aspect, making them worse for users.

Keywords: Social networks, self comparison, miscoordination.

JEL Codes: D21, D26, D85.

[*] University of Warwick, email: r.akerlof@warwick.ac.uk.

[†] 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 harder and harder to deny. In his recent book, *The Anxious Generation*, Jonathan Haidt makes the case that social media usage is responsible for a mental health crisis among young people. Since 2010, there has been more than a 150 percent increase in major depression among teens, and a roughly 10 percentage point drop in the share of 8th, 10th, and 12th graders who say they are satisfied with themselves.[1] This decline in mental health begins exactly at the moment when smartphones start to be adopted. Haidt illustrates this crisis with the story of Alexis, who joins Instagram at the age of 11. Initially, she is excited, 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!!!!" However, her enthusiasm quickly wains. Alexis found her feed increasingly populated with images of models, dieting tips, and eventually, pro-anorexia posts, pushed on her by the platform's algorithms. By the time she reached eighth grade, she required hospitalization for anorexia and depression, struggles that persisted throughout her adolescence.

While much of Haidt's evidence is correlational in nature, there is mounting causal evidence. For instance, Braghieri et al. (2022) utilize the staggered rollout of Facebook across U.S. college campuses to show that Facebook increased symptoms of poor mental health and especially depression. Additional evidence on mechanisms suggests that the findings are due to Facebook's tendency to promote negative self-comparisons among users.

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

Another possibility is that people feel *trapped*: they dislike these platforms but

[1] The first number is based on data from the US National Survey on Drug Use and Health. The later number is based on data from the Monitoring the Future survey. (see Haidt (2024b))

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

need to be on them to socialize with their peers. According to 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 use the term “bad network” to refer to a network on which people feel trapped. If we take the bad network story seriously, a key question is how such networks get established. Why, absent some form of irrationality, would people flock to a network that they intensely dislike?

The purpose of this paper is twofold. First, we show that, in fact, it is surprisingly easy for bad networks to get going. Second, we develop an explicit model of social networks that provides micro-foundations for why they may be bad. The feature that makes them bad is the self-comparison aspect, which accords with the findings of Braghieri et al. (2022). In addition, we show that social media platforms have an incentive to increase self-comparison, making them worse for users.

We first consider 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. Two types of bad outcomes that can occur are (i) all agents stay off the network even though it is welfare maximizing for them to join, and (ii) all agents join the network even though it is welfare maximizing for them to stay off. The latter case, which is our main focus, is possible when $0 > a > b$. In other words, bad networks can arise when it is unpleasant to join but even worse to stay off the network.

We next show that bad networks get going very easily. We allow some agents to be “instigators” and others “anti-instigators.” Instigators receive private benefits from joining the network while anti-instigators face private costs from joining. We show that an arbitrarily small number of instigators, who receive arbitrarily small private benefits, are sufficient to start a bad network. Intuitively, the instigators get the party started; and once there is a party, everyone feels obligated to be there.

Anti-instigators, by contrast, do not prevent the bad network from forming.

We next develop a micro-founded model of social networks that explains why they may have $0 > a > b$. An important feature of social networks is that they make it salient how one compares to others. For example, the pictures Instagram feeds Alexis cause her to question her own appearance. We build a model in which joining a social network has two effects. On the one hand, joining allows an agent to establish social connections (which are valuable); on the other hand, joining makes self-comparisons more salient. We show that this latter feature creates a rat race among agents and simultaneously creates incentives to join the network. Importantly, the platform benefits from making self-comparisons salient since this gets agents to join.

The rat-race in our framework 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 micro-foundation.

The paper closest to ours is contemporaneous work by Bursztyn et al. (2023). 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. The key differences between the model in Bursztyn et al. (2023) and our paper is the role of instigators as the trigger that gets bad networks started and our explicit micro-foundation. In our (Nash, as opposed to introspective) equilibrium it is an arbitrarily small mass of instigators that triggers the unravelling 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.

2 Good and Bad Networks

Consider a setting with a unit mass of agents who simultaneously decide whether to join a network. The utility of agent i is given by:[3]

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

where $x_i = 1$ (0) denotes the choice to join (remain off) the network, and q denotes the fraction of agents that join. There are externalities of the network for agents who join (captured by parameter a) and agents who remain off (captured by parameter b). We allow a and b to be either positive or negative. For ease of exposition, we assume $a \neq b$. Lemma 1 characterizes the Nash equilibria of this game.

Lemma 1.

1. *When the benefit of the network to those on it is small compared to those off it ($a < b$), the unique equilibrium is no participation ($q^{NE} = 0$).*

2. *When the benefit of the network to those on it is large compared to those off it ($a > b$), both no participation ($q^{NE} = 0$) and full participation ($q^{NE} = 1$) are equilibria.*

To see why, notice that if $a < b$, agents strictly prefer to remain off the network whenever $q > 0$. As a result, the unique equilibrium is no participation. If $a > b$, agents strictly prefer to join when $q > 0$ and only weakly prefer to join when $q = 0$. As a result, both full participation and no participation are equilibria.

Lemma 2 characterizes the optimal q from an aggregate welfare perspective (which we denote by q^*).

Lemma 2.

1. *When there are no benefits of the network ($a, b < 0$), no participation is welfare maximizing ($q^* = 0$).*

2. *When there are large benefits to those off the network ($b > \max\{0, 2a\}$), mixed participation is welfare maximizing $\left(q^* = \frac{b}{2(b-a)}\right)$.*

[3] Our main results hold for a more general class of utility functions which induce qualitatively similar aggregate welfare, but we focus on linear externalities for expositional simplicity.

3. *When there are benefits to those on the network ($a > 0$) and the benefits to those off the network are small ($b < 2a$), full participation is welfare maximizing ($q^* = 1$).*

To see why Lemma 2 holds, observe that aggregate welfare is given by

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

If there are no benefits of the network ($a, b < 0$), $W(q)$ is maximized by setting $q = 0$. If there are benefits of the network (a or $b > 0$), there may be an interior solution that balances the benefit to those on the network with those off the network; but if b is sufficiently low, $W(q)$ is maximized by putting everyone on the network.

If we examine Lemmas 1 and 2 together, we obtain conditions under which a good outcome ($q^{NE} = q^*$) can occur. We also obtain conditions under which a bad outcome ($q^{NE} \neq q^*$) can occur—or does occur (recall there may be two equilibria). Proposition 1 specifies these conditions.

Proposition 1.

Good outcomes:

1. *Full participation on a good network ($q^{NE} = q^* = 1$) can occur when $a > b$ and $a > 0$.*

2. *No participation on a bad network ($q^{NE} = q^* = 0$) can occur when $a, b < 0$.*

Bad outcomes:

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

2. *No participation on a mixed network ($q^{NE} = 1$ and $q^* = \frac{b}{2(b-a)}$) does occur when $\max\{a, b\} > 0$, and $b > \max\{a, 2a\}$.*

3. *Full participation on a bad network ($q^{NE} = 1$ and $q^* = 0$) can occur when $0 > a > b$.*

The first two bad outcomes—no participation on a good network and no participation on a mixed network—are well understood forms of miscoordination. The third bad outcome—full participation on a bad network—has been less studied.

This type of outcome can arise when $0 > a > b$—that is, when it is costly to be on the network but even more costly to be off it.

Note that when $0 > a > b$, the bad outcome is not guaranteed; a Nash equilibrium also exists with no participation. In the following section, we focus squarely on the case where $0 > a > b$ and ask whether there are forces that make the bad outcome with full participation more or less likely.

3 Participation in Bad Networks

When $0 > a > b$, an important question is whether the good outcome ($q^{NE} = q^* = 0$) or the bad outcome ($q^{NE} = 1$ and $q^* = 0$) is more likely to arise (both are Nash equilibria). To explore this question, we expand the model by assuming that some agents receive a private benefit from joining the network ("instigators") or face a private cost from joining the network ("anti-instigators"). We examine how these agents affect the equilibrium.

Formally, we assume that agents have utility

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

where ϵ_i is a private benefit or cost from joining the network. For simplicity, we initially focus on the case where there are no anti-instigators. In particular, we assume that a fraction $\mu_I > 0$ of agents are instigators and receive a private benefit $\epsilon_I \geq 0$. For all other agents, $\epsilon_i = 0$. We refer to them as "non-instigators."[4]

Notice that instigators strictly prefer to join the network if the private benefit is large enough: $\epsilon_I > (b - a)q$. Since $b < a$, $\epsilon_I > 0$ is, in fact, a sufficient condition for instigators to join independently of q. Non-instigators strictly prefer to join whenever $q > 0$. Thus, non-instigators will join if the instigators join. Importantly, the equilibrium behavior of both instigators and non-instigators does not depend on the mass μ_I of instigators. This gives us the following lemma.

Lemma 3. *If $\epsilon_I > 0$, full participation ($q^{NE} = 1$) is the unique equilibrium.*

[4] 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 the network in the absence of anyone else joining.

The lemma shows that even an arbitrarily small mass of instigators—with an arbitrarily small preference to join the network—is sufficient to kill the no-participation equilibrium.

The private benefits that instigators receive, of course, have an impact on the welfare analysis. Aggregate welfare is now given by:

$$W(q) = (a-b)q^2 + bq + \epsilon_I \min\{q, \mu_I\}.$$

However, as the following lemma shows, provided instigators' private benefits are small, it is still optimal to have no participation ($q^* = 0$).

Proposition 2. *If private benefits are small* $\left(\epsilon_I < -\max\{\frac{a}{\mu_I}, b + \mu_I(a-b)\}\right)$*, no participation is welfare maximizing* ($q^* = 0$)*.*

We see from Proposition 2 that instigators, provided their private benefits are relatively small, simply have the effect of killing the good equilibrium. Intuitively, even though agents dislike the network ($a < 0$), they dislike being off it even more ($b < a$). The network is, effectively, a party people feel compelled to attend but do not want to attend.[5]

Anti-instigators. Suppose now that, in addition to a mass μ_I of instigators, there is a mass μ_A of anti-instigators. For anti-instigators, $\epsilon_i = -\epsilon_A$, where $\epsilon_A > 0$ is a private cost associated with joining the network.

Notice that if $\epsilon_I > 0$, instigators join the network regardless of whether non-instigators or anti-instigators do; and non-instigators join the network whenever instigators do. Thus, anti-instigators do not prevent instigators or non-instigators from joining. Moreover, the anti-instigators end up joining as well if the private cost of joining is small compared to the cost of staying off when instigators and non-instigators join:

$$\epsilon_A < (b-a)(1-\mu_A).$$

Anti-instigators are also more likely to end up joining if there are few of them. In sum, anti-instigators do not keep other agents from joining a bad network.

[5] Proposition 2 focuses on the case where no participation is welfare maximizing ($q^* = 0$). There are also cases where mixed participation or full participation are welfare maximizing. If private benefits are large and the mass of instigators is small $\left(\epsilon_I > -b - \mu_I(a-b)\right.$ and $\left.\mu_I < -\frac{a}{a-b}\right)$, mixed participation is welfare maximizing ($q^* = \mu_I$). If both private benefits and the mass of instigators are large $\left(\epsilon_I > -\frac{a}{\mu_I}\right.$ and $\left.\mu_I > -\frac{a}{a-b}\right)$, full participation is welfare maximizing.

Moreover, aggregate welfare may be strictly worse than in a world without anti-instigators given that anti-instigators face an additional cost when they join the network compared to non-instigators.[6]

4 When are Social Networks Bad?

We have shown that when $0 > a > b$, bad networks get going easily. A key remaining question is why social networks might have this property.

Haidt (2024a) argues that the self-comparison aspect of social networks is a key reason why they are harmful to mental health. Since 2010, the share of 8th, 10th, and 12th graders who say they are satisfied with themselves has dropped by 10 percentage points. Indeed, Braghieri et al. (2022) find that negative self-comparisons are the primary driver of poor mental health among college Facebook users.

With this is mind, we build a model that provides micro-foundations for network externalities a and b. There are two key forces at work. On the one hand, joining a social network increases the salience of self-comparisons. On the other hand, joining creates social connections, which agents value.

4.1 Model

Suppose there is a unit mass of agents and each agent makes two choices. They decide whether to join a social network ($x_i = 0$ or 1) and whether to exert effort ($e_i = 0$ or 1).

Effort affects the agent's performance in a competition for *esteem*. Agent i receives a rank $R_i \in [0, 1]$ in the competition for esteem, where 1 is the highest rank and 0 is the lowest rank. Agents who exert effort always receive higher ranks than those who do not. Specifically, an agent who exerts effort receives a random rank between 1 and $1 - q_e$, where q_e is the fraction of agents who exert effort; an agent

[6] The approach we take to equilibrium selection in Section 3 is to add small private benefits and costs of joining the network. An alternative approach is to use a focality concept such as introspective equilibrium (see Akerlof et al. (2023) for a discussion). Introspective equilibrium is based upon level-k thinking. It assumes that agents are endowed with a level-0 behavior—or "impulse"—and defines introspective equilibrium as the limiting case where $k \to \infty$. It is easy to show that if even a small fraction of agents have an impulse to join the network, the unique introspective equilibrium is full participation. The agents with an impulse to join play an analogous role to instigators.

who does not exert effort receives a random rank between $1 - q_e$ and 0. Notice that if the agent exerts effort, their expected rank is $1 - \frac{q_e}{2}$; if the agent does not exert effort, their expected rank is $\frac{1}{2} - \frac{q_e}{2}$. Therefore, exerting effort increases an agent's expected rank by $\frac{1}{2}$.

Agent i is risk neutral and has a utility function with three components:

$$U_i = \underbrace{(1 + \alpha \cdot x_i)(R_i - \tfrac{1}{2})}_{\text{Esteem Component}} \quad + \quad \underbrace{\beta \cdot q_j \cdot x_i}_{\text{Connection Component}} \quad - \quad \underbrace{C \cdot e_i.}_{\text{Cost of Effort}} \tag{1}$$

The first component—the "esteem component"—captures the agent's concern about how they compare to others. The agent's self esteem is equal to $R_i - \frac{1}{2}$, which is the difference between their own rank (R_i) and the average rank (1/2). In other words, the agent's self esteem is based upon how they perform relative to other agents. 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 reflects the idea that a social network makes self-comparisons more salient.

The second component—the "connection component"—reflects the benefit to agents on the network from being able to connect with peers on the network. We assume $\beta > 0$.

The final component of the utility function is the cost of exerting effort. The benefit of exerting effort is that it increases an agent's expected rank by $\frac{1}{2}$. We assume that $C > \frac{1}{2}$, which ensures that agents who do not join the social network ($x_i = 0$) do not find it worthwhile to exert effort.

4.2 Analysis

Agents who do join the network find it optimal to exert effort if $C \leq \frac{1+\alpha}{2}$, i.e. $\alpha \geq 2C - 1$. We separate our analysis into the case where $\alpha < 2C - 1$ and $\alpha \geq 2C - 1$.

Case 1: $\alpha < 2C - 1$

First consider the case where the social network has a small effect on the salience of esteem ($\alpha < 2C - 1$). In this case, no agent has an incentive to exert effort. When no agent exerts effort, expected utility is given by:

$$E(U_i) = \beta \cdot q_j \cdot x_i$$

Notice that this corresponds to the model in Section 2 with $a = \beta > 0$ and $b = 0$. According to Lemma 2, this is a "good network" where full participation is optimal ($q^* = 1$). Intuitively, this type of social network has the beneficial effect of connecting peers, and it does not induce a rat race among agents where they compete for esteem.

Case 2: $\alpha \geq 2C - 1$

When esteem is relatively salient on the social network ($\alpha \geq 2C - 1$), agents on the network exert effort. In this case, agent i's utility is given by:

$$E(U_i) = \beta \cdot q_j \cdot x_i + \left(\frac{(1+\alpha) - \alpha q_j}{2} x_i - \frac{q_j}{2}\right) - C \cdot x_i \tag{2}$$

According to equation 2, if agent i does not join the social network ($x_i = 0$), their expected payoff is $-\frac{q_j}{2}$. Thus, agents who do not join the network are hurt by the network. Intuitively, the agents on the network put effort into raising their rank; this lowers the rank (and therefore esteem) of agents off the network.

According to equation 2, if agent i does join the network, their expected payoff is:

$$\left(\beta - \frac{1+\alpha}{2}\right) q_j + \left(\frac{1+\alpha}{2} - C\right).$$

The first term is a network externality, while the second term is a benefit/cost unrelated to network size. To keep the exposition simple, let us focus on the case where this second term is equal to zero: $C = (1+\alpha)/2$.

In this case, the model corresponds exactly to the model from Section 2, with $a = \beta - \frac{1+\alpha}{2}$ and $b = -\frac{1}{2}$. Notice that, if β is low, the network is harmful to agents on the network as well as agents off the network (i.e. $a < 0$). Agents on the network are harmed by the network because it generates a rat race where they are forced to compete for esteem.

The network is a bad network that gets established easily ($0 > a > b$) when:

1. $\beta < \frac{1+\alpha}{2}$

2. $\beta > \frac{\alpha}{2}$

The first condition says that the benefit from connecting agents (β) cannot be too large. Otherwise, the network would have positive value to those on it (i.e. $a > 0$). The second condition says that the benefit from connecting agents (β) cannot be too small. Otherwise, agents would not be tempted to join the network and so it would never get going in the first place. In this intermediate range, the network is a bad network that gets established easily.[7]

The effect of α.

Suppose a social network can control the extent to which self-comparisons are salient. That is, they can increase or decrease the value of α. From equation 2, we see that an increase in α increases the agent's desire to join the network (i.e. choose $x_i = 1$).[8] Thus, raising social image considerations is potentially an effective tool for increasing engagement on the network. At the same time, raising the salience of self-comparisons may turn the network into a bad network.

Indeed, there are a variety of design choices that social media platforms make that appear to be consciously geared toward making self-comparison more salient (see, for example, Vogel et al. (2014)). For instance, most platforms prominently display metrics such as “likes” and follower counts. Algorithmic feeds tend to prioritize content that performs well according to such metrics. Additionally, platforms tend to push content from outside users' immediate peer groups, showcasing idealized images and lifestyles.

5 Conclusion

There is significant evidence that social networks are harmful to individuals but that people feel compelled to be on them because others are on them. We provide a framework for analyzing this social media rat race.

We show that an equilibrium where every agent participates on a bad network can arise naturally. Indeed, an arbitrarily small number of instigators who receive

[7] 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.

[8] In particular, the difference between utility on and off the network can be written as $\beta q_j + \frac{\alpha}{2}(1 - q_j) - C$, which is strictly increasing in the salience of esteem α whenever $q_j < 1$.

an arbitrarily small private benefit from being on network leads to full participation on the bad network. Finally, our microfoundation emphasizes the way in which social networks themselves can exploit an individual desire for esteem to increase network engagement. This amplifies the rat race and deepens the extent to which people feel trapped on social networks.

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.

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, *The Anxious Generation*, Penguin Random House, 2024.

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

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.

6 Appendix: Proofs

6.1 Proof of Lemma 2

Proof. Recall that welfare is given by

$$\mathbb{W}(q) = (a-b)q^2 + bq.$$

If $(a-b) > 0$ then this is a convex function and the optimum must be at the boundary, i.e. either $q = 0$ or $q = 1$. Since $\mathbb{W}(0) = 0$ and $\mathbb{W}(1) = a$, it follows that $q = 1$ is welfare maximizing if $a > 0$. Hence if $b < a < 0$, it must be that $q = 0$ is welfare maximizing. This proves the half of case 1. Moreover, if $b < 0 < a$ then $q = 1$ must be welfare maximizing, this proves the first part of case 3. Now suppose $(a-b) < 0$ so that the welfare function is concave. If $a < b < 0$ then welfare is strictly decreasing and so $q = 0$ is welfare maximizing, this completes the proof of case 1. If $b > 0$, then by the first order condition welfare has a unique interior maximum at $q = \frac{b}{2(b-a)} \geq 0$. Since q must be in $[0, 1]$, this interior maximum is only valid if $\frac{b}{2(b-a)} < 1$, that is, if $b > 2a$. On the other hand, if $b > a$ and $b < 2a$ then full participation must be uniquely welfare maximizing. Finally, if $b > 2a$ and $b > 0$, then it's necessarily true that $b > a$, since for $a > 0$ we have $b > 2a > a > 0$ and for a negative we have $b > 0 > a$. This proves cases 2. and the final part of case 3. □

6.2 Proof of proposition 1

Proof. By lemma 1, $q = 0$ is always an equilibria regardless of the values of b and a. So "Good outcome" 2. and "Bad outcome" 1. follow by combining this fact with lemma 2. Similarly, $q = 1$ is an equilibrium whenever $a > b$, and so "Good outcome" 1. and "Bad outcome" 3. follow by combining this fact with lemma 2. Finally, when $b > a$ the unique equilibrium is $q = 0$, and since mixed networks can only arise when $b > a$, it follows that no participation is always the unique equilibrium on mixed networks, which gives us the final case– "Bad outcome" 2. □

6.3 Proof of proposition 2

Proof. Here we provide a full characterization of socially optimal network sizes including those discussed in Footnote 4. Recall that welfare is given by

$$\mathbb{W}(q) = (a-b)q^2 + bq + \epsilon_I \min\{q, \mu_I\}.$$

Since $a > b$, welfare is the minimum of two strictly convex functions and therefore there are 3 possible maxima: $0, \mu_I$ and 1.[9] We have

$$\begin{aligned} \mathbb{W}(0) &= 0, \\ \mathbb{W}(\mu_I) &= (a-b)\mu_I^2 + (b+\epsilon_I)\mu_I, \\ \mathbb{W}(1) &= a + \epsilon_I \mu_I. \end{aligned}$$

The optimal q depends in on which of these three values is the maximum. Notice that $\mathbb{W}(\mu_I) > 0$ if and only if $(a-b)\mu_I + (b+\epsilon_I) > 0$, that is, if and only if $\mu_I > -\frac{b+\epsilon_I}{a-b}$. Finally to see where $\mathbb{W}(\mu_I) > \mathbb{W}(1)$, notice that

$$\begin{aligned} & \mu_I < -\frac{a}{a-b} \\ \implies & a + (a-b)\mu_I < 0 \\ \implies & b\mu_I > a(1+\mu_I) \\ \implies & b\mu_I(1-\mu_I) > a(1-\mu_I^2) \\ \implies & (a-b)\mu_I^2 + b\mu_I > a \\ \implies & (a-b)\mu_I^2 + (b+\epsilon_I)\mu_I > a + \epsilon_I\mu_I, \end{aligned}$$

where of course this final line is equivalent to $\mathbb{W}(\mu_I) > \mathbb{W}(1)$. The converse of the above chain of implications also holds. Hence we conclude that $q = 0$ is optimal when

$$0 > \max\{(a-b)\mu_I + (b+\epsilon_I), a + \epsilon_I\mu_I\},$$

[9] This is precisely the reason that we can extend our analysis to an arbitrary number of heterogeneous masses of instigators (as we mention in section 3)– the welfare function is still the lower envelope of some number of convex functions.

$q = \mu_I$ is optimal when

$$(a-b)\mu_I^2 + (b+\epsilon_I)\mu_I > \max\{0, a+\epsilon_I\mu_I\},$$

and finally, $q = 1$ is optimal when

$$a+\epsilon_I\mu_I > \max\{0, (a-b)\mu_I^2 + (b+\epsilon_I)\mu_I\}.$$

Now we rewrite these conditions in terms of ϵ_I and μ_I. We have $q = 0$ optimal when

$$\epsilon_I < -b-(a-b)\mu_I \text{ and } \epsilon_I < -\frac{a}{\mu_I}$$

We have $q = \mu_I$ is optimal when

$$\epsilon_I > -b-(a-b)\mu_I \text{ and } \mu_I < -\frac{a}{a-b}$$

and finally that $q = 1$ is optimal when

$$\epsilon_I > -\frac{a}{\mu_I} \text{ and } \mu_I > -\frac{a}{a-b}.$$

This proves proposition 2. □