Bad Networks

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

July 10, 2024

Abstract

We analyze a model of social networks where there are benefits to being both on or off the network. Multiple equilibria naturally arise, including an equilibrium in which all agents choose to join a socially sub-optimal or *bad* network. Focusing on bad networks, we highlight the role of *instigators* who receive a private benefit from joining the network. Even if this private benefit is arbitrarily small, for any positive mass of instigators the unique equilibrium is full participation on a bad network. Finally, we offer a micro-foundation for the forces that give rise to bad social networks based on a competition for *esteem*. The social network can amplify the benefits of esteem on the network, and this is consistent with empirical findings of people feeling "trapped" on social networks.

Keywords: Esteem, instigators, social networks.

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.

"There is only one thing in life worse than being talked about, and that is not being talked about." Oscar Wilde.

1 Introduction

The harmful effects of social media are becoming harder and harder to deny. Haidt (2024a) points to four categories of harm to adolescents emanating from the use of phones or phone-like devices—many of which involve the use of social media, including "social deprivation", "sleep deprivation", and "attention fragmentation." There is also a widely-held view that many young people feel compelled to be on social media because others are on social media. As Haidt (2024b) puts it: "Social media has trapped an entire generation in a collective-action problem."

The time-series evidence pointing to the correlation between the use of mobile devices/social media and increases in depression and various forms of self harm is unmistakable (Haidt (2024b)). Even the timing of falls in standardized-test scores and entrepreneurship look suspicious.

But there is also mounting causal evidence. For instance, Braghieri et al. (2022) utilize the staggered rollout of Facebook across U.S. college campuses to identify a causal impact of social media exposure on self-reported mental-health outcomes and the likelihood of students' reporting a negative effect of mental health on academic performance. Beknazar-Yuzbashev et al. (2022) conduct a field experiment where the treatment group is exposed to a smaller amount of toxic content by way of a browser extension. Less exposure to toxic content translated into 23% less Facebook content consumption, a 9% drop in Twitter advertisement consumption, and a reduction in the toxicity of social media posts by those in the treatment group. Allcott et al. (2022) present evidence from a field experiment suggesting that addiction accounts for 31% of social media use. Moreover, it is arguably in the interests of social-media platforms to expose users to toxic content.[1]

Social media platforms may be harmful not merely to adolescents, or indeed just individuals, but to society. For instance, Facebook has been blamed for, among

[1] Beknazar-Yuzbashev et al. (2024) analyze a model in which advertisement-driven platforms in markets with network externalities can find it optimal to expose users to toxic content which users prefer not to see if, conditional on being exposed to the content, those users engage more on the platform.

other things: facilitating foreign influence in U.S. politics, and providing a platform for polarizing and repugnant views. It is routinely suggested that Twitter has destroyed civilized public discourse.

Of course there are also many benefits of social networks. Among other things they may democratize political debate and potentially play an important role in connecting individuals in an age when we might otherwise be "bowling alone."

Depending on the relative size of these costs and benefits, then, social networks may be *good* or *bad*. And, as in any setting with network externalities, coordination problems naturally arise. Moreover, the vast majority of participants seem to feel trapped in a bad equilibrium. They don't want to be on the network, but they feel they have to be.

There remains, however, the puzzle of how bad networks that everyone wants to get off came to exist. How did they get going in the first place? There is no collective-action problem to get off the network before the network exists. One answer is that young people stumbled into this problem before they understood it. Everyone ended up on social networks before realizing they didn't want to be, and now they're stuck. A second, more nefarious explanation is that people were conned. This is Haidt's (2024b) explanation: "Early app developers deliberately and knowingly exploited the psychological weaknesses and insecurities of young people to pressure them to consume a product that, upon reflection, many wish they could use less, or not at all."

It is harder to see how a bad network can get started if all the participants are purely rational, and immune to psychological exploitation. The purpose of this paper is twofold. First, we show that, in fact, it is surprisingly easy for a bad network to get started. Second, we provide an explicit micro-foundation for agents preferences which give rise to bad social networks, and we show how the social network itself can manipulate these preferences by making image considerations salient.

In our model we emphasize an under-appreciated downside of social networks: that they may create a rat race. Individuals face a binary decision to be either on or off the network. The good thing about being on the network is the ability to participate in it, and the value of this is increasing in network size. The bad thing about being on the network is the inability to enjoy the benefits of being off the network, which are also increasing in network size.

When all agents are homogeneous and the benefits to being on the network are larger than those from being off the network then there are two equilibria: one in which no network forms, and one in which all agents join the network.

We then introduce a small mass of *instigators* who derive a private benefit from joining the network.[2] Our main result is that even if the mass of instigators is arbitrarily small and the private benefits to instigators are also arbitrarily small the unique equilibrium is full participation. Thus, even when no network is socially optimal, the unique equilibrium is full participation in the bad network. This is redolent of Haidt's (2024b) observation: "Social media has trapped an entire generation in a collective-action problem."

We go on to show that the presence of *anti-instigators* who have the opposite preference to instigators does not alter the structure of the equilibrium. Of course, an appropriately chosen Pigouvian tax permits a planner to achieve the optimal network size.

Finally, we offer an explicit micro-foundation for the forces that give rise to bad social networks. Here, agents make two choices. First they decide whether to join a social network, and then whether to exert (costly) effort. Effort affects an agent's performance in a competition for *esteem*, and esteem improves the agent's payoff. Importantly, how much esteem matters to agents depends on whether they are on or off the network. Moreover, the social network can amplify this esteem component when agents are on the network. As we discuss later, this is consistent with empirical findings regarding how social-image considerations increase network engagement.

This concept of rat-race costs of social networks connects to Tirole (2021) who analyzes a model in which activity can take place either in private or in the public sphere and in which agents care about their "image". The idea that there is a costs from activity being pushed from the private sphere into the public sphere 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

[2]See Granovetter (1978) for an early depiction of such instigators.

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 is the role of instigators as the trigger mechanism to get the bad network to form, and our explicit micro-foundation. In our (Nash, as opposed to introspective) equilibrium it is an arbitrarily small measure of instigators with arbitrarily small private benefits that trigger 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.

The remainder of the paper is organized as follows. Section 2 states our baseline model and preliminary welfare analysis. Section 3, which is the heart of the paper, introduces instigators into the model, and provides conditions under which a benevolent social planner would restrict the size of—or even ban—a network. Section 4 provides a micro-foundation for the model analyzed in Sections 2 and 3. Section 5 concludes. Proofs not contained in the main text are relegated to the appendix.

2 Good and Bad Networks

Suppose there is a unit mass of agents who simultaneously decide whether to join a network. The utility of agent i is given by:

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

where $x_i = 1$ (0) denotes the choice to join (remain off) the network, and q denotes the fraction of agents that join the network. Notice that there are benefits of the network both to 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; in the latter case, we can think of a and b as costs rather than benefits. For ease of exposition, we assume $a \neq b$. Lemma 1 characterizes the Nash equilibria of the 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 this, 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. *When there are benefits to those on the network ($a > 0$) and the benefit to those off the network is 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 $b < \max\{a, 2a\}$.*

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 a well understood form 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 (recall, both are 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 such 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 from joining the network. For simplicity, we initially focus on the case where there are just instigators. In particular, we assume that a fraction μ_I 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."[3]

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. Non-instigators strictly prefer to join whenever $q > 0$. Thus, non-instigators will join if the instigators join. This gives us the following lemma.

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

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

[3]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.

($b < a$). The network is, effectively, a party people feel compelled to attend but do not want to attend.[4]

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 so; and non-instigators join the network whenever instigators do. Thus, anti-instigators do not prevent the instigators or non-instigators from joining the network. Moreover, the anti-instigators end up joining as well provided there are few of them, or the private cost of joining is small compared to the cost of staying off when other agents join:

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

In sum, anti-instigators do not keep other agents from joining a bad network. 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.[5]

4 When are Social Networks Bad?

Here, we consider the forces that make social networks good ($q^* = 1$) or bad ($q^* = 0$). As we mentioned above, Haidt (2024a) delineates a number of different harms stemming from social networks: social deprivation (the loss of physical play and synchronous in-person interactions), sleep deprivation (due to spending

[4]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 $\left(\epsilon_i > b + \mu_I(a - b) \text{ and } \mu_I < -\frac{a}{a-b}\right)$, mixed participation is welfare maximizing ($q^* = \mu_I$). If both private benefits and the mass of instigators are large $\left(\epsilon_i > -\frac{a}{\mu_I} \text{ and } \mu_I > -\frac{a}{a-b}\right)$, full participation is welfare maximizing.

[5]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.

large amounts of time on digital devices), and attention fragmentation (the shortening of attention spans due to constant alerts). These costs largely occur because of social pressure to be on the network given that others are on the network. Hence the sense of feeling "trapped." Of course, he duly acknowledges there are benefits from social connections. The preferences we specify in equation (1) capture these costs and benefits.

To that end, 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 they decide whether to exert effort ($e_i = 0$ or 1).

The effort an agent exerts affects their 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. An agent who exerts effort ($e_i = 1$) receives a random rank between 1 and $1 - q_e$ while an agent who does not exert effort ($e_i = 0$) receives a random rank between $1 - q_e$ and 0. Notice that if $e_i = 1$, an agent's expected rank is $1 - \frac{q_e}{2}$; and if $e_i = 0$, an agent's 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:

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

The first 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 "esteem component" reflects the concern agents have about esteem. Esteem is equal to the difference between the agent's rank (R_i) and the average rank ($1/2$). 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 it more salient how one is viewed by peers and compares with peers.

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.

Agents who do join the network ($x_i = 1$) find it optimal to exert effort if $C \leq \frac{1+\alpha}{2}$, i.e. $\alpha \geq 2C-1$. We separate our analysis below 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. Moreover, 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 social network to exert effort. When agents exert effort (do not exert effort) when they join (do not join) the network, agent i's expected 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 1, if agent i does not join the social network, 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 esteem) of agents off the network.

According to equation 1, if agent i does join the social 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: $\alpha = 2C-1$.

In this case, the model corresponds 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.

Recall from Sections 2 and 3 that, the network is a "bad network" ($q^* = 0$) if $a, b < 0$; however, full participation on the bad network is only likely to occur if $0 > a > b$, or:

$$0 > \beta - \frac{1+\alpha}{2} > -\frac{1}{2}$$

This condition can be rewritten as the following two conditions:

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 big. Otherwise, the network would have positive value to those on it ($a > 0$). 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 it would never have a chance of getting established. In this intermediate range, though, the network is a bad network that is also likely to get established.

The effect of α.

Suppose a social network can control the extent to which image considerations 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 $x_i = 1$). Thus, raising social image considerations is potentially an effective tool for increasing engagement on the network. At the same time, raising social image concerns may turn the network into a bad network.

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 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.

Beknazar-Yuzbashev, George, Rafael Jiménez Durán, and Mateusz Stalinski, "A Model of Harmful yet Engaging Content on Social Media," *Available at SSRN*, 2024.

_ , _ , Jesse McCrosky, and Mateusz Stalinski, "Toxic content and user engagement on social media: Evidence from a field experiment," *Available at SSRN 4307346*, 2022.

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.*

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

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 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. Then by the first order condition it 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, if $b > a$ then the unique equilibrium is $q = 0$. Combining this with lemma 2 immediately gives us cases

1. and 2. In cases 3. and 4. both $q^* = 0$ and $q^* = 1$ are equilibria, so full participation can, but does not necessarily occur. □

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.[6] 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

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

$q = \mu_I$ 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

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

[6]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.

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

$$\mu_I < \min\{-\tfrac{b+\epsilon_I}{a-b}, -\tfrac{a}{\epsilon_I}\} = -\max\{\tfrac{b+\epsilon_I}{a-b}, \tfrac{a}{\epsilon_I}\}$$

We have $q = \mu_I$ is optimal when

$$-\tfrac{b+\epsilon_I}{a-b} < \mu_I < -\frac{a}{a-b},$$

and finally that $q = 1$ is optimal when

$$\mu_I > \max\{-\tfrac{a}{a-b}, -\tfrac{a}{\epsilon_I}\} = -\min\{\tfrac{a}{a-b}, \tfrac{a}{\epsilon_I}\}.$$

This proves proposition 2. □