Influencer Cartels*

Marit Hinnosaar† Toomas Hinnosaar‡

April 4, 2024§

First draft: February 12, 2021

Abstract

Influencers are often paid based on past engagement rather than their marketing campaigns' success, which incentivizes fraud. We study how influencers collude to inflate engagement, improving their market outcomes. Our theoretical model shows such influencer cartels mitigate the free-rider problem and may increase or decrease welfare, depending on engagement quality. Using machine learning to analyze texts and photos from Instagram and combining this with novel data from influencer cartels, we find that general interest cartels generate lower-quality engagement than topic-specific ones, which are closer to natural engagement. Narrow topic-specific cartels may be welfare-improving, whereas general interest cartels hurt everyone.

JEL: L41, C72, L86, M31, D26

Keywords: influencers, collusion, cartels, free-riding, commitment, marketing, Natural Language Processing, Large Language Models, Latent Dirichlet Allocation, cosine similarity

*We thank John Horton, Dina Mayzlin, Stephen P. Ryan, Stephan Seiler, and seminar participants at University of East Anglia, University of Leicester, University of Nottingham, Collegio Carlo Alberto, Mannheim Virtual IO Seminar, UK Competition and Markets Authority, 12th Paris Conference on Digital Economics, IIOC, Munich Summer Institute, Baltic Economics Conference, Australasian Meeting of the Econometric Society, NBER Summer Institute on IT and Digitization, Econometric Society European Meeting, EARIE, NIE Annual Conference, 13th Workshop on the Economics of Advertising and Marketing, Bristol-Warwick Empirical IO Workshop, Berlin IO Day, 6th Monash-Warwick-Zurich Text-as-Data Workshop, and CESifo Economics of Digitization Conference for helpful comments and suggestions.

†University of Nottingham and CEPR, marit.hinnosaar@gmail.com

‡University of Nottingham and CEPR, toomas@hinnosaar.net

§The latest version: https://marit.hinnosaar.net/influencer_cartels.pdf

1 Introduction

Collusion between a group of market participants to improve their market outcomes is typically considered an anti-competitive behavior. While some forms of collusion, such as price-fixing, are illegal in most countries, new industries provide new collusion opportunities for which regulation is not yet well-developed. In this paper, we study one such industry—influencer marketing. Influencer marketing combines paid endorsements and product placements by influencers. It allows advertisers a fine targeting based on consumer interests by choosing a good product-influencer-consumer match. Influencer marketing is a large and growing industry, with 31 billion U.S. dollars in ad spending in 2023.[1]

Many influencers are not paid based on their marketing campaigns' success, instead, their prices are based on past engagement (likes and comments). This gives incentives for fraudulent behavior—for inflating their influence. Inflating one's influence is a form of advertising fraud. It allows fraudulent market participants to steal advertising budgets and leads to market inefficiencies by directing ads to wrong eye balls. An estimated 15% of influencer marketing spending was misused due to exaggerated influence.[2] There are many ways to exaggerate influence. Regulators have started to address some of these. In 2023, the U.S. Federal Trade Commission proposed a new rule that would prohibit selling and buying false indicators of social media influence, such as, fake followers or views.[3] In this paper, we are studying a different way of exaggerating influence—influencer cartels. While there is substantial literature in economics on fake consumer reviews (Mayzlin et al., 2014; Luca and Zervas, 2016; He et al., 2022; Glazer et al., 2021; Smirnov and Starkov, 2022) and other forms of advertising fraud (Zinman and Zitzewitz, 2016; Rhodes and Wilson, 2018), the economics of this fraudulent behavior has not been studied.

In an influencer cartel, a group of influencers collude to inflate their engagement in order to increase their prices. Like in traditional industries, influencer cartels involve a formal agreement to manipulate the market for members' benefit. In traditional industries, the agreement typically involves price fixing or allocating markets. Influencer cartels involve a formal agreement to inflate the engagement measures to increase the prices influencers can get from advertisers. Influencer cartels operate in online chat rooms or

[1] In 2023, influencer marketing ad spending was almost as large as the print newspaper ad spending (35 billions U.S. dollars) according to Statista. Source: `https://www.statista.com/outlook/amo/advertising/influencer-advertising/worldwide#ad-spending`, accessed March 17, 2024.

[2] Source: `https://en.wikipedia.org/wiki/Influencer_marketing`, accessed June 5, 2023.

[3] Source: Federal Trade Commission, June 30, 2023, "Federal Trade Commission Announces Proposed Rule Banning Fake Reviews and Testimonials", `https://www.ftc.gov/news-events/news/press-releases/2023/06/federal-trade-commission-announces-proposed-rule-banning-fake-reviews-testimonials`, accessed March 18, 2024.

discussion boards, where members submit links to their content for additional engagement. In return, they are required to engage with other members' content through likes and comments. An algorithm enforces the cartel rules.

In this paper, we study how influencers collude to inflate engagement, to improve their market outcomes, and what are the welfare implications of the influencer cartels. To do that, we first build a theoretical model. Then we use machine learning to analyze text and photos from Instagram combined with a novel dataset of influencer cartels.

Our model shows that in this market, the key distortion is the free-rider problem. Engaging with other influencers' content brings attention to someone else's content, creating a positive externality. In equilibrium, there would be too little engagement compared to the social optimum. A cartel could lessen the free-rider problem by internalizing the externality. By joining the cartel, influencers agree to engage more than the equilibrium engagement. They get compensated for this additional engagement by receiving similar engagement from other cartel members. If the cartel only brings new engagement from influencers with closely related interests, this could benefit cartel members but also consumers and advertisers. However, the influencer cartel can also create new distortions. The cartel may overshoot and create too much low-quality engagement. Our theoretical results show that this may hurt all involved parties, consumers, advertisers, and indirectly even the influencers themselves.

The key dimension to separate socially beneficial cooperation from welfare-reducing cartels is the quality of engagement, i.e., whether the additional engagement comes mostly from influencers with similar interests. The idea is that influencers are typically used to promote the product among people with similar interests, e.g., vegan burgers to vegans. If a cartel generates engagement from influencers with other interests (e.g., meat-lovers), this hurts consumers and advertisers. Consumers are hurt because the platform will show them irrelevant posts, and advertisers are hurt because their ads are shown to badly targeted consumers. Whether or not a particular cartel is welfare-reducing or welfare-improving is an empirical question.

In our empirical analysis, we combine data from two sources: cartel interactions from Telegram and data from Instagram. Our cartel data allows us to directly observe (not predict or estimate) which Instagram posts are included in the cartel and observe which engagement originates from the cartel (via cartel rules). Our dataset includes two types of cartels: three topic-specific cartels and six general cartels with unrestricted topics. We use machine learning to analyze text and photos from Instagram to measure engagement quality. Our goal is to compare the quality of natural engagement to that originating from the cartel. We measure the quality by the topic match between the cartel member and the Instagram user who engages. To quantify the similarity of Instagram users, first,

we use a large language model (Language-agnostic BERT Sentence Embedding) and an analogous large neural network for text and photos (Contrastive Language Image Pre-training model) to generate numeric vectors (embeddings) from the text and photos in Instagram posts. Then we calculate cosine similarity between the users based on these numeric vectors. To further analyze the topic match of influencers and users who engage with their content, we use the Latent Dirichlet Allocation (LDA) model to map each Instagrammer's content to a probability distribution over topics.

Using this data, we show that the engagement that originates from general cartels compared to topic-specific cartels is of lower quality in terms of the topic match. Furthermore, the engagement originating from the general cartels is almost as bad topic match as coming from a completely randomly selected Instagram user. While engagement originating from the topic-specific cartels is closer to natural engagement.

Our empirical and theoretical results have two policy implications. Cartels that lead to limited added engagement from closely related influencers are socially beneficial, whereas cartels that increase engagement indiscriminately are socially undesirable. Therefore, policies that reduce large-scale cartel formations are likely to be welfare-improving. A good starting point could be, for example, shutting down influencer cartels that advertise themselves, can be found via search engines, and are open to the general public.

Second, monetary incentives based on the follower count and engagement tend to give incentives for fraud and unproductive collusion. Therefore the advertising market could be better off by using contracts that offer influencers a fraction of the added sales rather than payments related to the engagement. Alternatively, instead of simply measuring engagement quantity (for example, number of comments), the platform could improve the outcomes by reporting match-quality-weighted engagement measures, using methods such as in this paper. Both approaches reduce the incentives to create the lowest-quality engagement.

Similar trade-offs rise in other settings, for example, academic citations. Researchers who cite other papers create a positive externality on the authors of the papers. Since the ones who cite don't internalize the externality, in equilibrium, there isn't enough citations. Forming a group that agrees to cite each others papers, helps the group participants. Whether this agreement is helpful to readers depends on the closeness of the topic match of the group members. Agreeing to cite papers that aren't topic-specific cannot direct readers to relevant works. There is evidence of such agreements in academic journals (Franck, 1999). Thomson Reuters regularly excludes journals from the Impact Factor listings due to anomalous citation patterns.[4] Van Noorden (2013) and Wilhite and Fong (2012) have studied citation cartels. In contrast to citation cartels,

[4] `http://help.prod-incites.com/incitesLiveJCR/JCRGroup/titleSuppressions`.

where data of explicit agreements is limited, in influencer cartels, the collusion and outcomes are directly observable. More generally, trade-offs similar to our model arise also in patent pools and record sharing. Building a product on someone else's patent creates a positive externality. To internalize the externality, firms have formed patent pools already since 1856 (Moser, 2013; Lerner and Tirole, 2004). But patent pools can easily be anti-competitive (Lerner et al., 2007; Lerner and Tirole, 2004, 2015). Another example is record-sharing, for example, by hospitals (Miller and Tucker, 2009). Hospitals who share their records create positive externality to patients and other hospitals, which they are not able to fully able to internalize. Indeed, Grossman et al. (2006) find that competition between hospitals is one of the main barriers to data sharing and suggest methods for cooperation.

This paper adds to a small but growing literature in economics on influencer marketing. The empirical literature has analyzed advertising disclosure (Ershov and Mitchell, 2020), while the theoretical literature has studied the relationship between followers, influencers, and advertisers, optimal platform design, as well as the benefits of mandatory advertising disclosure (Fainmesser and Galeotti, 2021; Pei and Mayzlin, 2022; Mitchell, 2021; Berman and Zheng, 2020; Szydlowski, 2023). In contrast to these papers, our focus is on collusion between the influencers.

The paper adds to the literature on social media and attention (for an overview see Aridor et al. (2024)). While the literature on social media has extensively studied consumption and production of social media content, there is less work on strategic engagement: the strategic choice which content to engage with. Our work is most closely related to (Filippas et al., 2023) who using Twitter data study what they call attention bartering. Similarly to this paper they model social media users' decision to engage (in their setting, whether to follow other users). Different from their work, our paper focuses on the users agreement to collude when deciding whether to engage.

The paper also relates to the empirical literature on the operation of cartels.[5] As nowadays cartels typically are illegal, most studies use either historical data on known cartels from the time they were legal (Porter, 1983; Genesove and Mullin, 2001; Röller and Steen, 2006; Hyytinen et al., 2018, 2019) or data from the court cases (Clark and Houde, 2013; Igami and Sugaya, 2022), including of bidding-rings in auctions (for example, Porter and Zona (1993); Pesendorfer (2000); Asker (2010); Kawai et al. (2021)). The literature shows that collusion in cartels doesn't always take place via fixing prices or output (Genesove and Mullin, 2001). We describe a novel type of collusion to affect market outcomes in a new and yet unregulated industry. Instead of smoky backroom deals, in this industry, communication takes place in a chat room and agreements are

[5]For overviews, see Harrington (2006) and Marshall and Marx (2012).

enforced by an algorithm.

The paper also contributes to the theoretical literature on cartels. While the conventional wisdom is that cartels reduce welfare, in some settings, cartels can be socially desirable. Fershtman and Pakes (2000) showed that sometimes collusion might lead to more and higher-quality products, which benefits the consumers more than the price increases hurt them. Deltas et al. (2012) found that in trade, collusion could help to coordinate the resources and therefore, benefits the consumers. We are providing another reason why collusion may help to internalize a positive externality.

In our empirical analysis, we build on the recent literature in economics that uses text and images as data.[6] In particular, we are using Large Language Models and large neural networks to generate embeddings from text and photos. We are also using the LDA model (Blei et al., 2003), which has been recently used in economics, for example, to extract information from Federal Open Market Committee meeting minutes (Hansen et al., 2018). We are also using the cosine similarity index. This and other similarity indexes have been used as quality measures in economics by, for example, by Chen et al. (2023) and Hinnosaar et al. (2022).

The rest of the paper is organized as follows. In the next section, we provide some institutional details of influencer marketing and influencer cartels. Section 3 introduces the theoretical model and gives the welfare implications of influencer cartels. Section 4 describes the dataset. Section 5 presents the empirical results. Section 6 discusses the policy implications. Section 7 concludes.

2 Influencer marketing and influencer cartels

In influencer marketing, firms pay influencers for product placement and product endorsement. Compared to TV or newspaper advertising, influencer marketing allows fine targeting, generating a great product and influencer match, and hence, a great product and consumer match. Influencer marketing is a large industry: in 2023, influencer marketing ad spending was about $31 billion, which close to the ad spending for print newspaper ads. While the most influential influencers are athletes, musicians, and actors, but most Instagram users involved in influencer marketing have only a few thousand followers. According to ANA (2020), 74% of the firms used mid-level influencers (25,000–100,000 followers) and 53% micro-influencers (up to 25,000 followers). Instagram is one of the main platforms for influencer marketing. It is a platform where users share photos and videos, and engage with other users' content by liking and commenting their posts

[6]For a recent surveys of the uses of text as data in economics, see Gentzkow et al. (2019); Ash and Hansen (2023).

and can follow other users to see more of their content. Instagram algorithm is more likely to show posts that the user's social network has engaged with, that is, posts that the users who the user follows have commented on or liked. This implies that an influencer engaging with another user's post, increases the likelihood of it being shown to its followers.

Many influencers are not paid based on the actual success of the current marketing campaign, instead, they are being paid based on past engagement—comments and likes on previous posts. This gives rise for fraudulent behavior: for inflating one's influence. An estimated 15% of the influencer marketing spending is misused due to exaggerated influence. The influencers with a large following, typically, are paid based on the success of the marketing campaign, by tracking the sales originating from the influencer, using personalized links or coupons. But as of 2020, only 19% of the firms using influencer marketing were tracking the sales induced by influencers (ANA, 2020). Instead, most smaller influencers are paid before the start of the campaign, based on their characteristics. Initially, Instagram influencers were paid for the number of followers. This led to influencers getting fake followers (bots). The industry then moved to detect fake followers and measure and compensate engagement—likes and comments. There are alternative ways how to generate fake engagement. Some fake engagement is generated by automatic bots, which is relatively easy to detect. In this paper we study Instagram cartels, where the engagement is generated by humans and is, therefore, more difficult to separate from the natural engagement.

Instagram influencer cartels. In Instagram influencer cartels, influencer collude to inflate each others engagement, and they do that to increase the price that they can get from the advertisers.

How do the influencer cartels operate? They operate in other online platforms, either in a chat room or a discussion board (typically, on Telegram or Reddit).[7] In the chat room, members of the cartels submit links to their Instagram content that they would like to receive additional engagement. In order to receive that engagement, they themselves must engage with a fixed set of links submitted by other users. Specifically, before submitting a link themselves, they are required to like and write meaningful comments to previous N posts from other members. The rules of the cartel are enforced autotically by an algorithm.

The cartel increases engagement first via the direct effect as the cartel members engage with each other's posts. But the cartel also increases engagement indirectly. The Insta-

[7]For more details, see a computer science overview of Instagram cartels operating on Telegram (Weerasinghe et al., 2020) or for example: Apr 9, 2019 "Instagram Pods: What Joining One Could Do For Your Brand", Influencer Marketing Hub. `https://influencermarketinghub.com/instagram-pods/`

gram algorithm gives higher exposure to posts with higher engagement, which leads to even more engagement. More specifically, the Instagram algorithm is more likely to show posts that the user's social network has engaged with, that is, posts that the users who the user follows have commented on or liked. This implies that an influencer engaging with another user's post, increases the likelihood of it being shown to its followers.

As the cartels' activity of artificially increasing engagement is fraudulent, the groups are secret. Instagram considers the groups as violating Instagram's policies.[8]

The cartels in our sample operate in Telegram chatrooms and advertise themselves as a way to "attract lucrative brand partnerships" (see screenshots in Online Appendix A). The cartels in our sample have the requirement that before submitting a post, the member must like and write comments to the last five posts submitted by other members. The process ensures that each post receives five likes and comments each time it is submitted. Online Appendix A appendix shows an example of the post submitted to the cartel received the required comments. The rules are enforced by an algorithm that deletes submissions by users that don't follow the rules. The cartels in our sample have entry requirements: either thresholds for the minimum number of followers (ranging from 1,000 followers to 100,000 followers) or restrictions on the topics of the posts.

3 Theoretical Model

To build intuition, we present the theoretical results in three steps. We start with a basic model of influencer engagement without collusion and the advertising market. We then add collusion and, finally, the advertising market. Our focus is solely on engagement between influencers, we abstract away from all other aspects of influencer marketing, including content creation. All proofs are in appendix B.

3.1 Basic Model

We assume that there is an infinite sequence of players (influencers), indexed by $t \in \{-\infty, \ldots, -1, 0, 1, \ldots, \infty\}$. Player t is characterized by two-dimensional type (α_t, R_t).[9] The first parameter α_t captures the topic, which we model as Salop (1979) circle, it is an angle from 0° to 360° on the circle. The second parameter is the player's reach

[8]Source: Devin Coldewey, Apr 29, 2020, "Instagram 'pods' game the algorithm by coordinating likes and comments on millions of posts", TechCrunch. `https://techcrunch.com/2020/04/29/instagram-pods-game-the-algorithm-by-coordinating-likes-and-comments-on-millions-of-posts/`.

[9]Our treatment of player types is inspired by conventional wisdom in influencer marketing practice (Burns, 2020), which emphasizes the importance of "three R's": (1) Relevance: how relevant is the content to the audience, (2) Reach: the number of people the content could potentially reach, and (3) Resonance: how engaged is the audience. We model the first one as α_t and combine the latter two into R_t, which we call reach for brevity.

$R_t \geq 1$, which measures how many people the player's content regularly reaches (number of followers and typical search traffic). The distribution of topics is assumed to be uniform and reach has a power law distribution with mean 2. That is, the probability density function is $f(R_t) = 2R_t^{-3}$.[10] Both parameters are independent draws from corresponding distributions.

In this analysis, we focus on engagement. Player t has a piece of content and chooses between two actions $a_t \in \{0, 1\}$: to engage with the previous player's content $a_t = 1$ or not to engage $a_t = 0$. This can be thought of as a wall of content and the is whether to like and comment the previous post or not. We normalize all payoffs without engagement to zero.

Player t's choice to engage creates a social benefit and a social cost. We think of the benefit as providing information and entertainment to the audience and the cost as attention by the audience. The cost and benefit are both proportional to reach R_t, which measures the size of the audience. Both are also function of the topic difference of the influencer and the content she is engaging with. If the topics are similar, then the benefit is high and cost is low. If the topics are different, then the cost is high and the benefit is low. We model the cost and benefit separately, because we need to model the externality.

Specifically, we assume that the social benefit is $R_t \cos(\Delta_t)$, where $\Delta_t = |\alpha_t - \alpha_{t-1}|$ is the difference between players' t and $t-1$ topics.[11] If Δ_t is close to $0°$ (so that $\cos(\Delta_t) = 1$), the players' content is on similar topics, whereas if the difference is close to $90°$ their content is unrelated ($\cos(\Delta_t) = 0$), and if it is close to $180°$ the content is contradictory (e.g. political content, then we can have $\cos(\Delta_t) = -1$). The engagement generates also a social cost $R_t C(\Delta_t)$, where $C(\Delta) = \sin(\Delta)$ for all $\Delta \leq 90°$ and 1 otherwise.[12] Formally, the social welfare generated by action a_t by player t with reach R_t and topic difference Δ_t with the previous player $t-1$, is

$$W_t(a_t) = a_t \underbrace{R_t \cos(\Delta_t)}_{\text{Benefit}} - a_t \underbrace{R_t C(\Delta_t)}_{\text{Cost}}. \tag{1}$$

The costs and benefits of engagement are divided asymmetrically between players. We assume that players capture a constant fraction β of social costs and benefits. As factor β multiplies all payoffs related to engagement, without loss in generality we normalize β

[10] The uniform assumption for the topic is the standard in literature since Salop (1979). Power law distribution is a natural assumption for reach as it is the prevalent distribution for the number of readers, followers, comments (Gabaix, 2016). The mean 2 assumption is for tractability.

[11] Distance $|\alpha_t - \alpha_{t-1}| \in [0°, 180°]$ denotes the shortest angle difference on a circle. Formally, $|\alpha_t - \alpha_{t-1}| = \min\{\text{abs}(\alpha_t - \alpha_{t-1}), 360° - \text{abs}(\alpha_t - \alpha_{t-1})\}$, where $\text{abs}(x)$ is the absolute value.

[12] The assumption that the cost function is 1 beyond the $90°$ threshold simply accounts for the fact that $\sin(\Delta)$ function would be decreasing. Any weakly increasing function in this region would give similar results.

to one. If a player engages with previous content, then only she pays the cost, but the benefits are divided between the creator of the content and the one who engages. These assumptions capture the long-term relationships with the audience. If the influencer engages with content that her audience is not interested in, she incurs the full cost of misdirected attention. It captures the idea that her followers will pay less attention to her future content, they may even stop following her. On the other hand, the influencer who engages gets only a fixed fraction $\gamma < 1$ of the benefit, the remaining $1 - \gamma$ goes to the content creator. Thus, engagement creates a positive externality to the content creator.

In summary, the payoff of player t depends only on actions a_t and a_{t+1} as follows:

$$u_t(a_t, a_{t+1}) = a_t \underbrace{\gamma R_t \cos(\Delta_t)}_{\text{Internalized benefit}} - a_t \underbrace{R_t C(\Delta_t)}_{\text{Cost}} + a_{t+1} \underbrace{(1-\gamma) R_{t+1} \cos(\Delta_{t+1})}_{\text{Externality}}. \tag{2}$$

We assume that players' actions are not observable to the following players.[13] We also assume that player t observes the topic α_{t-1} of preceding player $t-1$, but does not know the follower's type. We consider Bayes-Nash equilibria, where players choose optimal action a_t, observing their own and previous player's type, and taking an expectation over the follower's type.

The free-riding problem. The positive externality of engagement creates a free-riding problem and therefore in equilibrium, there is less engagement than socially optimal. Specifically, in equilibrium player engages only with players whose topic is sufficiently similar (Δ_t is small enough). We can therefore say that only high quality engagement (in terms of match value) occurs in equilibrium. On the other hand, when taking into account the positive externality, it would be socially optimal to also engage with players whose topic is less similar. We can therefore say that in equilibrium there is too little engagement and even somewhat lower-quality engagement is socially optimal. The following proposition formalizes this intuition.

Proposition 1. *There is more engagement in social optimum than in non-cooperative equilibrium, but the additional engagement is of lower quality. In particular,*

1. *in non-cooperative equilibrium,* $a_t = \mathbf{1}_{\Delta_t \leq \tan^{-1}(\gamma)}$,
2. *in social optimum,* $a_t = \mathbf{1}_{\Delta_t \leq 45^\circ}$.

The comparison between socially optimal and equilibrium engagement shows that there is room for improvement from cooperation. If players could commit to engage somewhat more and get more engagement in return, they would be happy to do so.

[13] This assumption eliminates equilibria, where players engage conditional on past engagement.

3.2 Influencer Cartels

Our model of influencer cartels is simple, as the rules in those cartels are enforced by an algorithm. Therefore, we do not need to model the incentives of the cartel members to follow the rules as tacit collusion or repeated game. Instead, we model a cartel as an entry game, where they choose whether to enter a given cartel agreement. After learning their own types (α_t, R_t), but before learning other players' types, players simultaneously choose whether to join the cartel. A player who does not join the cartel gets outside option, which we normalize to 0. Players who join, form a subsequence $(\dots, s_{-1}, s_0, s_1, s_2, \dots)$, where s_t is the t'th member of the cartel.

We model cartel as a simple agreement defined by a single parameter Λ. The cartel member s_t must engage with the content of previous member s_{t-1} whenever their topic difference is less or equal than Λ, that is $\Delta_{s_t} = |\alpha_{s_t} - \alpha_{s_{t-1}}| \leq \Lambda$. The topic-difference parameter Λ is a convenient way to model how topic specific is the cartel. In a cartel that is not topic-specific, the parameter $\Lambda = 180°$, while in a narrowly topic specific cartel the topic difference is small.

The payoff from joining the cartel are the similar to the payoff without cartels, but instead of choice whether to engage, now the engagement is defined by the cartel agreement. When deciding whether to join the cartel, players take an expectation over other cartel members' types. A player with type (α_{s_t}, R_{s_t}), who joins the cartel, gets the expected payoff:

$$\begin{aligned} u^{\text{cartel}}(R_{s_t}) = \mathbb{E}\left[\mathbf{1}_{\Delta_{s_t} \leq \Lambda} \left(\gamma R_{s_t} \cos(\Delta_{s_t}) - R_{s_t} C(\Delta_{s_t})\right)\right] \\ + \mathbb{E}\left[\mathbf{1}_{\Delta_{s_{t+1}} \leq \Lambda} (1-\gamma) R_{s_{t+1}} \cos(\Delta_{s_{t+1}})\right], \end{aligned} \tag{3}$$

where Δ_{s_t} and $\Delta_{s_{t+1}}$ are the topic differences with previous and next member of the cartel respectively, and the expectations over $\Delta_{s_t}, \Delta_{s_{t+1}}$, and $R_{s_{t+1}}$ are taken over the distribution of cartel members.

Equilibria. We focus on symmetric equilibria, where players join the cartel independently of topic α_t. Therefore, the distributions of Δ_{s_t} and $\Delta_{s_{t+1}}$ are still uniform. Let us first focus on the case where $\Lambda \leq 90°$, so that the cost function $C(\Delta_{s_t}) = \sin(\Delta_{s_t})$. Then

the cartel benefit from equation (3) is

$$\begin{aligned} u^{\text{cartel}}(R_{s_t}) &= R_{s_t} 2 \int_0^{\Lambda} \left[\gamma \cos(\Delta_{s_t}) - \sin(\Delta_{s_t})\right] d\Delta_t \\ &\quad + (1-\gamma)\mathbb{E}R_{s_{t+1}} 2 \int_0^{\Lambda} \cos(\Delta_{s_{t+1}}) d\Delta_{s_{t+1}} \\ &= \frac{4\lambda(\lambda-\gamma)}{\lambda^2+1}\left(\frac{1-\gamma}{\lambda-\gamma}\mathbb{E}R_{s_{t+1}} - R_{s_t}\right), \end{aligned} \tag{4}$$

where we simplified the expressions by using a monotonic transformation $\lambda = \tan\left(\frac{\Lambda}{2}\right)$.[14]

Using this expression, we can study the entry to the cartel and formalize it with proposition 2 below. There are three cases depending on the engagement requirement Λ. If the engagement requirement is low, then all players join the cartel. It is easy to see this when $\Lambda \leq \tan^{-1}(\gamma)$ as then the direct benefits exceed the costs. But even if the engagement requirement is slightly larger, benefits the player expects from the cartel are larger than the costs of fulfilling the engagement requirement. If engagement requirement is moderate, only players with smaller reach join the cartel. This is because the benefit of the engagement from the cartel depends on the average reach of a cartel member, $\mathbb{E}R_{s_{t+1}}$, but the cost depends on the player's own reach R_{s_t}. Hence, the first players to stay out of the cartel are with the highest reach. Therefore the equilibrium is described by a threshold $\overline{R}$, so that only players with reach $R_{s_t} \leq \overline{R}$ join the cartel. Finally, if the engagement requirement is $\Lambda = 90°$, nobody joins the cartel. Furthermore, if $\Lambda > 90°$, then equation (4) is an upper bound for the cartel payoff, and it is strictly negative, so in this case also nobody joins the cartel.

Proposition 2. *Depending on cartel agreement, we can have three possible types of equilibria in the entry game to the cartel:*

1. *If $\lambda \leq \gamma$, all players join the cartel.*
2. *If $\gamma < \lambda < 1$, all players with $R_t \leq \overline{R} = \frac{2-\gamma-\lambda}{\lambda-\gamma}$ join the cartel.*
3. *If $\lambda \geq 1$, nobody joins the cartel.*

Welfare. The proposition implies that only cartels with engagement requirement less than 90° are sustainable. Within this range, all cartels are welfare-improving. In appendix C, we derive welfare-maximizing cartels and show that the engagement requirement is always weakly less than 45°. In other words, based on this model of influencer

[14] Note that $\frac{1+\cos(\Lambda)}{\sin(\Lambda)} = \tan\left(\frac{\Lambda}{2}\right) = \lambda$ and $\sin(\Lambda) = \sin(2\tan^{-1}(\lambda)) = \frac{2\lambda}{\lambda^2+1}$.

cartels, we would only expect to see cartels with some restrictions on topics.[15]

3.3 Advertising Market

We model the advertising market as a competitive market, where a continuum of advertisers each has an ideal target topic. Each player t is matched with an advertiser with the same topic $\alpha = \alpha_t$.[16] We study how the existence and type of cartel affects the price that advertisers pay to influencers. The key aspect of the analysis is that the advertiser cannot observe the quality of engagement and hence the price that the advertiser offers reflects the expected quality of such engagement instead of its true value. In other words, the advertiser *pays for the quantity* of engagement.

The realized value of engagement from the follower $t+1$ to the advertiser is

$$\underbrace{a_{t+1}(1-\gamma)R_{t+1}}_{\text{quantity of engagement}} \times \underbrace{\cos(\Delta_{t+1})}_{\text{match quality}} \times \underbrace{v}_{\text{marginal value}}, \tag{5}$$

where v is the value of a marginal unit of engagement. This expression captures the idea that advertisers can measure the quantity of engagement quite accurately (number of views, clicks, likes, comments), but it is much harder to determine whether the engagement comes from the target audience. The product of the last two terms, $\cos(\Delta_{t+1})v$, is the unit value of engagement to the advertiser and its expectation determines the price of engagement.

As we assume that the advertising market is competitive, the price of engagement is equal to the expected unit value of engagement to the advertiser,

$$p^{\text{engagement}} = \mathbb{E}\left[\cos(\Delta_{t+1})\right] v, \tag{6}$$

where the expectation is taken over the distribution of the influencers who engage, conditional on their equilibrium behavior.[17]

Benchmark: only natural engagement. Before studying the impact of cartels, let us consider the case when all engagement comes from natural equilibrium behavior, as discussed in section 3.1. Adding the advertising value to the influencer's payoff function,

[15] In appendix C we also show that having a minimal reach requirement could be beneficial for the cartel. This can explain why some cartels have a minimum reach requirement in practice.

[16] Advertiser's topic α captures the audience the advertiser tries to reach rather than the characteristic of the product. For example, an advertiser in some tourist location may sometimes want to reach people who consider between this and other locations rather than people who visit this location regularly.

[17] Our results remain unchanged if players are able to capture a constant fraction of the value the advertiser gets from the engagement, for example via Nash bargaining.

we get

$$u_t^{ad}(a_t, a_{t+1}) = u_t(a_t, a_{t+1}) + a_{t+1}(1-\gamma)R_{t+1}p^{\text{natural}}, \tag{7}$$

where $u_t(a_t, a_{t+1})$ is the payoff defined by (2) and price of engagement is the price of natural engagement $p^{\text{natural}} = v\mathbb{E}[\cos(\Delta_{s_{t+1}})|\text{Natural}]$.

Notice that the new term $a_{t+1}(1-\gamma)R_{t+1}p^{\text{natural}}$ is independent of player t's action a_t, therefore equilibrium behavior is unchanged. By proposition 1, player t engages if and only if $\Delta_t \leq \tan^{-1}(\gamma)$. Hence, the price of natural engagement is

$$p^{\text{natural}} = v\mathbb{E}\left[\cos(\Delta_{s_{t+1}})\middle|\, \Delta_{s_{t+1}} \leq \tan^{-1}(\gamma)\right] = v\frac{\gamma}{\tan^{-1}(\gamma)\sqrt{\gamma^2+1}} \in (0.9v, v). \tag{8}$$

Cartels with advertising market. We assume that the advertiser is unable to distinguish cartel engagement from natural engagement. In particular, we assume that with probability $1-\varepsilon$, the engagement is natural, i.e., comes from equilibrium behavior discussed above, and with the remaining probability $\varepsilon \in (0,1)$ the engagement comes from a cartel. The price of engagement is therefore

$$p^{\text{engagement}} = (1-\varepsilon)p^{\text{natural}} + \varepsilon p^{\text{cartel}}, \tag{9}$$

where p^{natural} is the price of natural engagement from (8) and $p^{\text{cartel}} = v\mathbb{E}[\cos(\Delta_{s_{t+1}})|\text{Cartel}]$ is the price of engagement coming from cartels.

To determine the equilibrium price of engagement, we need to study how the advertising market affects the engagement within a cartel. The payoff function of a player joining the cartel with the added value from advertising is

$$\begin{aligned} u^{\text{cartel+ad}}(R_{s_t}) &= u^{\text{cartel}}(R_{s_t}) + \mathbb{E}\left[\mathbf{1}_{\Delta_{s_{t+1}} \leq \Lambda}(1-\gamma)R_{s_{t+1}}p^{\text{engagement}}\right] \\ &= u^{\text{cartel}}(R_{s_t}) + \frac{\Lambda}{180^\circ}(1-\gamma)\mathbb{E}[R_{s_{t+1}}|\text{Cartel}]p^{\text{engagement}}, \end{aligned} \tag{10}$$

where $u^{\text{cartel}}(R_{s_t})$ is defined by (3) and $\mathbb{E}[R_{s_{t+1}}|\text{Cartel}]$ is the expected reach of a cartel member.

For clarity, let us focus on the case when the advertising market incentives are large, i.e., the marginal value of engagement, v is large enough. Then also $p^{\text{engagement}}$ is large, because by (8) and (9), $p^{\text{engagement}} \geq (1-\varepsilon)0.9v$. For any fixed reach R_{s_t}, the first term $u^{\text{cartel}}(R_{s_t})$ in equation (10) is bounded, so with sufficiently large $p^{\text{engagement}}$, the second part of the expression dominates. This means that for any $\Lambda > 0$ there is $v > 0$, such that the payoff from joining the cartel is positive for players with reach below some threshold $\overline{R} > 1$. This implies the following result.

Proposition 3. *With advertising market, for all $\Lambda > 0$ and any $\overline{R} > 1$, there exists $v > 0$, such that all players with $R_t \leq \overline{R}$ join the cartel.*

We can conclude that if the incentives from the advertising market are large, then a cartel with $\Lambda = 180°$ is sustainable. In fact, it is even in some sense desirable for cartel members, as joining such a cartel brings more engagement than a cartel that limits engagement to a narrower topic. Thus, we would expect to see such general cartels in practice. Indeed, most of the cartels in our sample are non-specific cartels that require engagement regardless of the topic and do not put any restrictions on the topics of the posts.

Welfare impact of general cartels. Such general cartels reduce the welfare of consumers and influencers outside cartels, and can also be undesirable for the advertisers and some influencers within the cartel. In particular, as the following corollary 1 states, they unambiguously reduce the welfare of consumers and influencers outside the cartel. Such cartels provide no social value in expectation, but create substantial cost due to the attention cost for the audience. They always hurt influencers who do not belong to the cartel, because these influencers get a lower price of engagement from their advertisers.

The effect on other parties is subtler. The advertising market is competitive, so that their expected value is always zero. General cartels drive down the price of engagement, so that the advertisers who happen to be matched with an influencer involved in natural engagement, actually benefit from the cartel by paying a lower price, whereas the advertisers who are matched with cartel members, pay for worthless engagement. The expectation over the two possibilities is zero, just the outcomes are more uncertain, and the uncertainty itself could be undesirable for advertisers.

Finally, let's consider the members of the cartel. They receive a positive expected benefit from belonging to the cartel; otherwise, they would not have joined. However, when the share of engagement coming from cartels becomes large enough (with $\varepsilon > 1/2$), even members of general cartels would prefer that these cartels would have a stricter engagement requirement $\Lambda < 180°$. This is because the reduction in the quantity of engagement is offset by the increase in the price.

Corollary 1. *With sufficiently large v, general cartels with $\Lambda = 180°$ are sustainable.*

1. *General cartels strictly reduce consumer welfare.*
2. *General cartels strictly reduce the welfare of influencers outside the cartel.*
3. *General cartels create uncertainty for the advertisers.*

4. *If $\varepsilon > 1/2$, then all members of general cartels would prefer that these cartels would have slightly lower Λ.*

4 Data and measures of engagement quality

4.1 Data sources

We combine data from two sources: first, the detailed cartel communications from Telegram, and second, Instagram posts and engagement data. A detailed description of our data collection is in online appendix D.

Telegram cartel history. From Telegram, we collected the communication history of nine cartels: six general interest cartels and three topic-specific cartels: fitness & health, fashion & beauty, travel & food. The Telegram cartel interaction history consists of three pieces of information: Telegram username, Instagram post shortcode, and time. According to the cartel rules, this information determines which cartel member has to comment and like which Instagram post. This is because one has to comment and like five posts by other users directly preceding one's own. This implies that we observe, instead of having to infer, which posts are included in the cartel. Similarly, we observe, instead of having to infer, which engagement originates from the cartel according to the cartel rules. The Telegram cartels include 220,893 unique Instagram posts that we were able to map to 21,068 Instagram users.

Instagram data. Our goal is to compare natural engagement to that obtained via cartels. In engagement, we focus on comments instead of likes or views, because information on who views the post is not available and data on who likes the post is more difficult to collect than comments. We already know which cartel members have to comment according to the cartel rules. But we also want to obtain information on natural engagement.

We define natural engagement as comments from users who don't belong to any of the cartels in our data. To obtain information on natural engagement, we focus on each cartel member's first post in any of the cartels. For each cartel member's first post in cartels, we collected information on who commented on the post. Then we used a random number generator and picked a random non-cartel user who had commented on the post. The randomly chosen commenting Instagram users who are not cartel members form our control group. Since these are from the earliest post in the cartel, they are less likely to be indirectly affected by the cartel activity. We collected the content of all public

Instagram posts for all cartel members and for these randomly picked non-cartel users. We could not collect information on non-cartel commenting users when the first post itself had been deleted or made private, when the first post had no non-cartel commenters, or when the randomly picked commenting user account was private. We also didn't collect information on non-cartel commenting users, when they had less than 10 posts. We also excluded about 5% of the non-cartel commenting users who had an associated post with a cartel member.[18] Online appendix presents details of the sample reduction. We were able to collect natural engagement for 10,683 cartel members. But some of these randomly picked non-cartel users were the same across posts, hence, this corresponds to 9,729 unique non-cartel Instagram users. Online appendix shows that the cartel members for whom we were or were not able to collect commenting non-cartel users are similar.

4.2 Measuring Engagement Quality

Our goal is to compare engagement that originates from cartels to that of natural engagement. Motivated by our model, we consider engagement to be of high quality if it comes from Instagram users who post on similar topics. Therefore, we measure the similarity of the posts of commenting users to those of the post author and the topic of the post itself. To calculate similarity, we use text and/or photos in Instagram posts and three alternative methods.

Sentence embeddings and cosine similarity. First, we use a large language model Language-agnostic BERT Sentence Embedding (LaBSE) to construct embeddings of the text in Instagram posts (Feng et al., 2022). An embedding represents text as a numerical vector in a multidimensional vector space. The vector representation of text is useful allowing easy comparison of texts via cosine similarity. Cosine similarity is a standard measure of text similarity. It is defined as the cosine of the angle between two vectors, providing a similarity score between -1 and 1, where close to 1 means that the texts (vectors) are highly similar. LaBSE is based on one of the first large language models: Bidirectional Encoder Representations from Transformers (BERT) developed by researchers in Google (Devlin et al., 2019). While BERT was originally implemented in the English language, LaBSE extends it to 109+ languages. The multilingual effectiveness is useful for us because our sample is multilingual. The LaBSE model transforms each post into a vector of length 768. It does so using a large neural network with approximately 470 million parameters. This enables the model to capture a large range of semantic features in multiple languages.

[18]The association can happen as Instagram allows post to be associated with multiple users (this is different from tagging a user) or it could happen when user changes usernames.

In the main analysis, we focus on 100 posts per user closest in the symmetric time window to the first post for the cartel member and to the post they commented on for the non-cartel users. Results are qualitatively similar when using a random sample or all posts from 2017 till 2020 (presented in supplementary material). In our main analysis, we create an embedding of each post using hashtags in the post. We focus on hashtags because typically, in Instagram posts hashtags informatively capture the essence of the post. Supplementary material presents results where the embeddings are created using the whole text of the post. To create the input for the embedding we first pre-process the text: (i) transform to lower case; (ii) replace all characters that are not letters, numbers, underscores, or hashtags with a space (these are the only characters allowed in an Instagram hashtag); (iii) add a space before each hashtag; (iv) keep only words that start with a hashtag; (v) keep only the first 30 hashtags in each post because Instagram allows only up to 30 hashtags per post; (vi) drop all hashtags that have only a single character because these tend to be uninformative; (vii) drop all hashtags that don't include any letters because these tend to be uninformative. Before creating an embedding we replace hashtag and underscore symbols with a space. After obtaining embeddings of posts, we generate a single measure for each user, by taking the average of the post embeddings for each user. Supplementary material presents results where instead of post embeddings, we first combine posts for each month, and obtain one embedding per month, and then take the average over the months for each user. Using the average embeddings we calculate the cosine similarity of user pairs. How to interpret the cosine similarity of average embeddings of users? Since cosine similarity is a linear transformation, the cosine similarity of average embeddings is essentially equivalent to calculating cosine similarity separately of all the posts of the two users and then taking the average of the cosine similarities of all these post pairs.

Photo and text embeddings and cosine similarity. We also construct embeddings of photos and text. As the above LaBSE model is able encode only text, we have to use a different model for processing images alongside text. We use the Contrastive Language Image Pre-training (CLIP) model, developed by OpenAI (Radford et al., 2021). CLIP maps the contents of images and text into a shared embedding space. Because CLIP generates embeddings for images and text that are directly comparable, it allows us to calculate similarity by combining both forms of information. The CLIP model transforms the text and photos into a vector of length 512. It does so using a neural network with approximately 86 million parameters. The advantage of the CLIP model is that it allows to combine photos and text. On the other hand, the LaBSE model allows to more precisely capture text.

In the analysis, we use the photos and text of the first post in the cartel for the cartel member, and for the non-cartel member, the closest post in the symmetric time window to the post they commented on. To create the text input for the CLIP embedding we first pre-process the text and here we keep the whole text not only the hashtags: (i) transform to lower case; (ii) replace question marks, exclamation marks, and new line breaks with a full stop; (iii) replace all characters that are not letters, numbers, full stops, underscores, hashtags, at symbols, or apostrophes with a space; (iv) drop groups of characters that don't include any letters of numbers; (v) add a space before each hashtag; (vi) drop posts that are shorter than 3 characters. Then, first, we generate embeddings separately of the text and photos of each post, and second, we take the average of the of text and photo embeddings of each post.

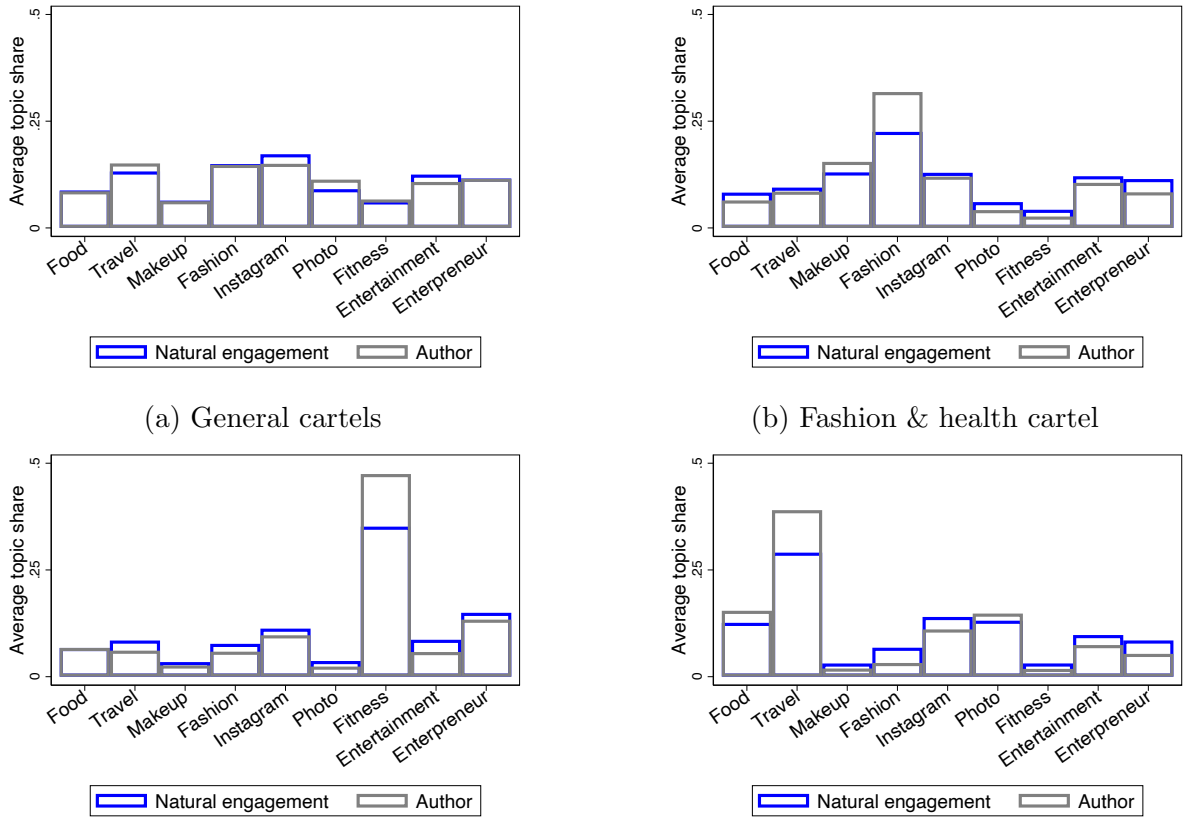
Determining topics using Latent Dirichlet Allocation. The models generating embeddings allow to measure similarity, but are somewhat black boxes. To shed some light on the comparison of users topics, we use Latent Dirichlet Allocation (LDA). The LDA algorithm estimates a probability distribution of topics for each user based on his posts, specifically, the hashtags used in the posts; and a probability distribution over the hashtags for each topic. In the main analysis, we use the same sample of posts and the same pre-processing of text as for the LaBSE model used with text embeddings. To improve learning from the underlying content, we reduce the set of hashtags. Specifically, we exclude hashtags that less than 100 users use. We also exclude users with less than 10 unique tags, because there is not enough inormation to learn their topics. We fix the number of topics to nine based on the content and the coherence score (figure E.1). Based on the most representative hashtags in each topic, that is, the hashtags with the highest probability (table E.1), we assign each topic a label. The labels are: food, travel, makeup, fashion, Instagram, photo, fitness, entertainment, enterpreneur.

The distribution of topics in the cartels is as expected (figure 1). In the fashion & health cartel, users are talking more about fashion; in the fitness & health cartel, about fitness; in the travel & food cartel, about travel and food. While in the general cartels, the topic distribution is rather uniform, with slightly more concentrated on Instagram, travel, and fashion topics.

4.3 Sample

The main analysis focuses on about eight thousand cartel members and the engagement they receive from cartel and non-cartel users. This sample includes only the cartel members for whom we were able to collect users who engaged with their content and is smaller because not all users members have enough posts with text to calculate the embeddings

and similarity. Users included and excluded from the main sample are similar in terms of their LDA topics (figure E.2).



(a) General cartels

(b) Fashion & health cartel

(c) Fitness & health cartel

(d) Travel & food cartel

Figure 1: LDA topic distributions: post authors (cartel members) versus natural engagement (non-cartel members)

5 Empirical Results

Empirical strategy. To answer the question whether engagement from cartels is of lower quality compared to natural engagement, we estimate a panel data fixed effects regression where the outcome variable is the cosine similarity of an influencer and his commenter. An observation is an influencer and his commenter pair. For each influencer, we focus on the first post in the cartel. Thus, for each influencer we have only one post. But we have several commenters for each influencer, some originating from the cartel and others what we call natural. Hence we have several observations for each influencer.

For the first post in cartel of influencer i, the similarity to its commenter j is:

$$\begin{aligned} Similarity_{ij} &= \beta_{Gen} GeneralCartelCommenter_j + \beta_{Top} TopicCartelCommenter_j \\ &+ \beta_{Ran} RandomNoiseCommenter_j + InstagramPostFE_i + \varepsilon_{ij}, \quad (11) \end{aligned}$$

where $Similarity_{ij}$ refers to the cosine similarity between influencer i and commenter j; $GeneralCartelCommenter_j$ indicates that a general cartel member j is required to comment; $TopicCartelCommenter_j$ is an indicates that a topic cartel member j is required to comment; $RandomNoiseCommenter_j$ indicates a random Instagram user not in the cartel who didn't actually comment. $InstagramPostFE_i$ is the fixed effect for each instagram post. Since we only have one post per influencer, this is equivalent to the influencer fixed effects. The base category is natural engagement, that is, a commenter who is not in the cartel.

We estimate the regressions using two alternative outcome variables: first, cosine similarity calculated based on the text (hashtags) of 100 posts (columns 1–3 of table 1), and second, cosine similarity based on the text and photo of a single post (columns 4–6 of table 1). We look at three different samples: first, posts included only in general cartels (columns 1 and 4), second, posts included in topic cartels (columns 2 and 5), and third, posts included in both general and topic cartels (columns 3 and 6).

Quality of engagement measured by the cosine similarity of users. We find that in general cartels (columns 1 and 4 in Table 1), influencer's similarity with commenting cartel members is significantly lower compared to the non-cartel commenters (base category). Furthermore, similarity to general cartel members is almost as bad as random noise. In contrast, in topic cartels (columns 2 and 5), similarity with commenting cartel members is only slightly worse than non-cartel commenters. Similar results hold for posts which are in both general and topic cartels (columns 3 and 6).

Distribution of LDA topics. To further study the quality of engagement, we present the distribution of topics that characterizes the engagement originating from general versus topic cartels. We focus on the influencers whose main topic corresponds to one of the topics in the topic cartels. We then compare the topics of their commenters from general versus topic cartels. Figure 2 presents the average distribution of topics of influencers and their commenters. It shows that an influencer whose main topic is, for example, fitness, from topic cartel receives engagement from users who also mostly post about fitness, while that is not the case for general cartels. The same pattern that general cartels compared to topic cartels generate engagement less similar to the influencer, holds for other topics.

Table 1: Estimates from panel data fixed effects regressions measuring influencer's similarity with commenters from cartels (or random users) versus non-cartel. Dependent variable: cosine similarity of influencer and commenter (or random user).

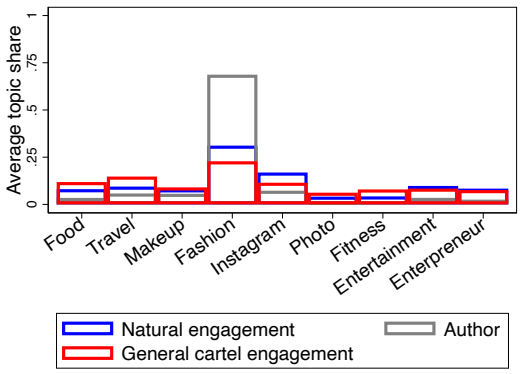
	(1)	(2)	(3)	(4)	(5)	(6)
	Dependent variable: Cosine similarity					
	Posts in general or topic cartels					
	General	Topic	Both	General	Topic	Both
	Similarity of users Text embeddings			Similarity of 1st posts Photo+text embeddings		
General cartel commenter	-0.058***		-0.057***	-0.046***		-0.040***
	(0.003)		(0.008)	(0.001)		(0.004)
Topic cartel commenter		-0.023***	-0.008		-0.027***	-0.025***
		(0.003)	(0.008)		(0.002)	(0.004)
Random noise commenter	-0.071***	-0.078***	-0.059***	-0.053***	-0.049***	-0.045***
	(0.003)	(0.003)	(0.008)	(0.001)	(0.002)	(0.004)
Wald test, $\beta_{Gen} = \beta_{Top}$			0.000			0.000
Base (non-cartel) mean	0.515	0.521	0.519	0.512	0.506	0.511
Post fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Posts	4729	3246	487	4729	3246	487
Observations	44528	30248	6601	44528	30248	6601

Notes: Each column presents estimates from a separate panel data fixed effects regression. Unit of observation is an influencer-commenter pair. Outcome variable is the cosine similarity of an influencer and his commenter or a random user. Each regression includes influencer (post) fixed effects. In all the regressions, the base category is influencer's similarity to a non-cartel commenter; and *Base cat. (non-cartel) dep. v. mean* presents their average cosine similarity. *General cartel commenter* is an indicator variable whether the commenter with whom the influencer's cosine similarity is calculated, is in the general cartel, and *Topic cartel commenter* whether he is in the topic cartel. *Random noise commenter* indicate that the influencer's similarity is calculated with a random user not in the cartel. Standard erros in parenthesis are clustered at the influencer level.

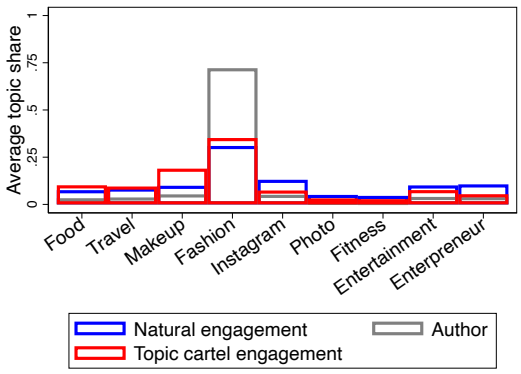
6 Policy Implications

Our empirical and theoretical results suggest two main policy implications. Our theory shows that cartels that require engagement with only closely related influencers are welfare improving, whereas cartels that require engagement regardless of the topic match are welfare reducing. Our empirical results show that general cartels generate low-quality engagement. This engagement is about as good as counterfactual engagement, where comments would come from random Instagram users. On the other hand, the topic-specific cartel generates engagement, which is at least as high quality as natural engagement.

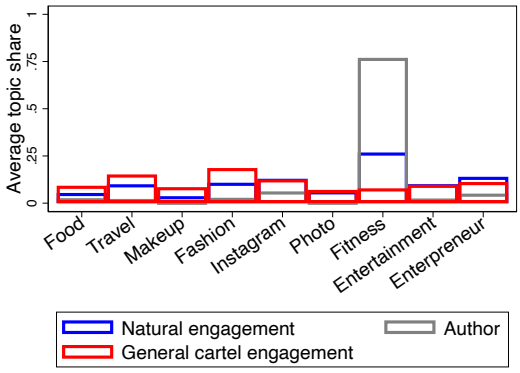
Our results, therefore, suggest that the highest priority for the regulator should be addressing general cartels. Our theory suggests that the engagement must come from influencers that are "close enough" in the topic. In practice, this means that cartels focusing on sufficiently specific topics could be welfare-improving. Our empirical approach allows measuring the similarity of influencers. For example, we find that the engagement originating from the "fitness and health" cartel is not worse than natural and hence, the



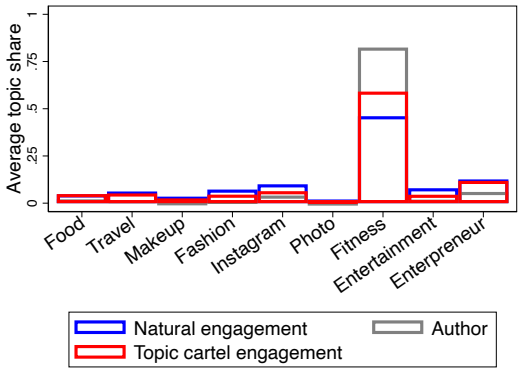
(a) Fashion in general cartels



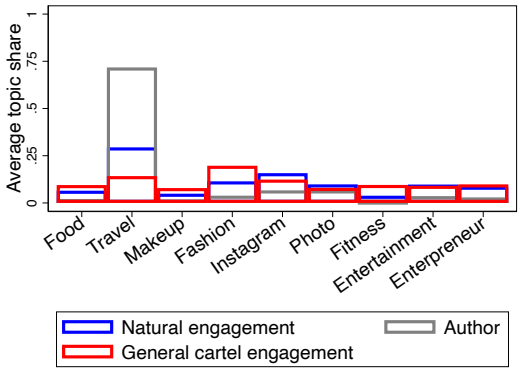
(b) Fashion in fashion & health cartel



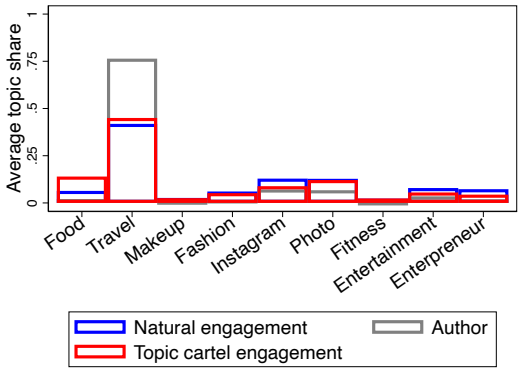
(c) Fitness in general cartels



(d) Fitness in fitness & health cartel



(e) Travel in general cartels



(f) Travel in travel & food cartel

Figure 2: LDA topic distributions: cartel-originating versus natural engagement for general versus topic cartels

additional engagement could be welfare improving.

The second implication of our theory is that monetary payments for engagement quantity may lead to large distortions. This was the only case where cartel members may choose and even prefer general cartels, which require engagements regardless of the topic match. The reason is simple: if only the quantity of engagement matters and the market pays well for it, it would be optimal for the players to create lots of engagement, even if this

is socially highly undesirable. Such a scenario, i.e., paying for the quantity of engagement, is common in practice, and our results suggest that this practice should be discontinued. In most situations, it should be possible to switch to a different compensation scheme, which combines lump-sum payments with payments for results (such as added sales). Alternatively, advertisers or the platforms could also use our methodology to evaluate the match quality. For example, instead of paying for the number of comments, they could weigh each comment by the match quality. Both suggested changes would reduce the appeal to generate fake engagement.

7 Conclusions

We documented and studied influencer cartels, a collusive behavior in the growing industry of influencer marketing, which has so far stayed under the radar of regulators. Our empirical results show that the engagement from general cartels is of significantly lower quality than the natural engagement, whereas the engagement from topic-specific cartels can be closer to natural engagement. Our theoretical model highlights the trade-offs and provides welfare implications. The key distortion is the free-rider problem, and commitment through cartels could potentially help to mitigate this problem. But cartels also create new distortions by over-engagement and exclusion of high-reach influencers. This problem of fake-engagement is especially serious when the advertising market offers large monetary rewards for engagement quantity.

References

ANA (2020): "The State of Influence: Challenges and Opportunities in Influencer Marketing," Tech. rep., Association of National Advertisers, association of National Advertisers.

Aridor, G., R. Jiménez Durán, R. Levy, and L. Song (2024): "The Economics of Social Media," *Manuscript*.

Ash, E. and S. Hansen (2023): "Text Algorithms in Economics," *Annual Review of Economics*, 15, 659–688.

Asker, J. (2010): "A Study of the Internal Organization of a Bidding Cartel," *American Economic Review*, 100, 724–762.

Berman, R. and X. Zheng (2020): "Marketing with Shallow and Prudent Influencers," *manuscript*.

Blei, D. M., A. Y. Ng, and M. I. Jordan (2003): "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, 3, 993–1022.

Burns, S. (2020): "Thinking About Influencer Marketingö Here's What To Look For," *Forbes*.

Chen, Y., R. Farzan, R. Kraut, I. YeckehZaare, and A. F. Zhang (2023): "Motivating Experts to Contribute to Digital Public Goods: A Personalized Field Experiment on Wikipedia," *Management Science*.

Clark, R. and J.-F. Houde (2013): "Collusion with Asymmetric Retailers: Evidence from a Gasoline Price-Fixing Case," *American Economic Journal: Microeconomics*, 5, 97–123.

Deltas, G., A. Salvo, and H. Vasconcelos (2012): "Consumer-Surplus-Enhancing Collusion and Trade," *RAND Journal of Economics*, 43, 315–328.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova (2019): "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," ArXiv:1810.04805 [cs].

Ershov, D. and M. Mitchell (2020): "The Effects of Influencer Advertising Disclosure Regulations: Evidence From Instagram," *manuscript*.

Fainmesser, I. P. and A. Galeotti (2021): "The Market for Online Influence," *American Economic Journal: Microeconomics*, 13, 332–72.

Feng, F., Y. Yang, D. Cer, N. Arivazhagan, and W. Wang (2022): "Language-agnostic BERT Sentence Embedding," ArXiv:2007.01852 [cs].

Fershtman, C. and A. Pakes (2000): "A Dynamic Oligopoly with Collusion and Price Wars," *RAND Journal of Economics*, 31, 207–236.

Filippas, A., J. J. Horton, and E. Lipnowski (2023): "The Production and Consumption of Social Media," *Manuscript*.

Franck, G. (1999): "Scientific Communication–A Vanity Fairö," *Science*, 286, 53–55.

Gabaix, X. (2016): "Power Laws in Economics: An Introduction," *Journal of Economic Perspectives*, 30, 185–206.

Genesove, D. and W. P. Mullin (2001): "Rules, Communication, and Collusion: Narrative Evidence from the Sugar Institute Case," *American Economic Review*, 91, 379–398.

Gentzkow, M., B. Kelly, and M. Taddy (2019): "Text as Data," *Journal of Economic Literature*, 57, 535–574.

Glazer, J., H. Herrera, and M. Perry (2021): "Fake Reviews," *The Economic Journal*, 131, 1772–1787.

Grossman, J. M., T. S. Bodenheimer, and K. McKenzie (2006): "Hospital-Physician Portals: The Role of Competition in Driving Clinical Data Exchange," *Health Affairs*, 25, 1629.

Hansen, S., M. McMahon, and A. Prat (2018): "Transparency and Deliberation Within the FOMC: A Computational Linguistics Approach," *Quarterly Journal of Economics*, 133, 801–870.

Harrington, J. E. (2006): *How Do Cartels Operateö*, Now Publishers Inc.

He, S., B. Hollenbeck, and D. Proserpio (2022): "The Market for Fake Reviews," *Marketing Science*, 41, 896–921.

Hinnosaar, M., T. Hinnosaar, M. Kummer, and O. Slivko (2022): "Externalities in Knowledge Production: Evidence from a Randomized Field Experiment," *Experimental Economics*, 25, 706–733.

Hyytinen, A., F. Steen, and O. Toivanen (2018): "Cartels Uncovered," *American Economic Journal: Microeconomics*, 10, 190–222.

——— (2019): "An Anatomy of Cartel Contracts," *Economic Journal*, 129, 2155–2191.

Igami, M. and T. Sugaya (2022): "Measuring the Incentive to Collude: The Vitamin Cartels, 1990–99," *Review of Economic Studies*, 89, 1460–1494.

Kawai, K., J. Nakabayashi, and J. M. Ortner (2021): "The Value of Privacy in Cartels: An Analysis of the Inner Workings of a Bidding Ring," Tech. Rep. w28539, National Bureau of Economic Research.

Lerner, J., M. Strojwas, and J. Tirole (2007): "The Design of Patent Pools: The Determinants of Licensing Rules," *RAND Journal of Economics*, 38, 610–625.

Lerner, J. and J. Tirole (2004): "Efficient Patent Pools," *American Economic Review*, 94, 691–711.

——— (2015): "Standard-Essential Patents," *Journal of Political Economy*, 123, 547–586.

Luca, M. and G. Zervas (2016): "Fake It Till You Make It: Reputation, Competition, and Yelp Review Fraud," *Management Science*, 62, 3412–3427.

Marshall, R. C. and L. M. Marx (2012): *The Economics of Collusion: Cartels and Bidding Rings*, MIT Press.

Mayzlin, D., Y. Dover, and J. Chevalier (2014): "Promotional Reviews: An Empirical Investigation of Online Review Manipulation," *American Economic Review*, 104, 2421–2455.

Miller, A. R. and C. Tucker (2009): "Privacy Protection and Technology Diffusion: The Case of Electronic Medical Records," *Management Science*, 55, 1077–1093.

Mitchell, M. (2021): "Free Ad(vice): Internet Influencers and Disclosure Regulation," *RAND Journal of Economics*, 52, 3–21.

Moser, P. (2013): "Patents and Innovation: Evidence from Economic History," *Journal of Economic Perspectives*, 27, 23–44.

Pei, A. and D. Mayzlin (2022): "Influencing Social Media Influencers Through Affiliation," *Marketing Science*, 41, 593–615.

Pesendorfer, M. (2000): "A Study of Collusion in First-Price Auctions," *Review of Economic Studies*, 67, 381–411.

Porter, R. H. (1983): "A Study of Cartel Stability: The Joint Executive Committee, 1880-1886," *Bell Journal of Economics*, 14, 301–314.

Porter, R. H. and J. D. Zona (1993): "Detection of Bid Rigging in Procurement Auctions," *Journal of Political Economy*, 101, 518–538.

Radford, A., J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever (2021): "Learning Transferable Visual Models From Natural Language Supervision," .

Rhodes, A. and C. M. Wilson (2018): "False advertising," *The RAND Journal of Economics*, 49, 348–369.

Röller, L.-H. and F. Steen (2006): "On the Workings of a Cartel: Evidence from the Norwegian Cement Industry," *American Economic Review*, 96, 321–338.

Salop, S. C. (1979): "Monopolistic Competition with Outside Goods," *Bell Journal of Economics*, 10, 141–156.

Smirnov, A. and E. Starkov (2022): "Bad News Turned Good: Reversal under Censorship," *American Economic Journal: Microeconomics*, 14, 506–560.

Szydlowski, M. (2023): "Deprioritizing Content," Place: Rochester, NY Type: SSRN Scholarly Paper.

Van Noorden, R. (2013): "Brazilian Citation Scheme Outed," *Nature*, 500, 510–511.

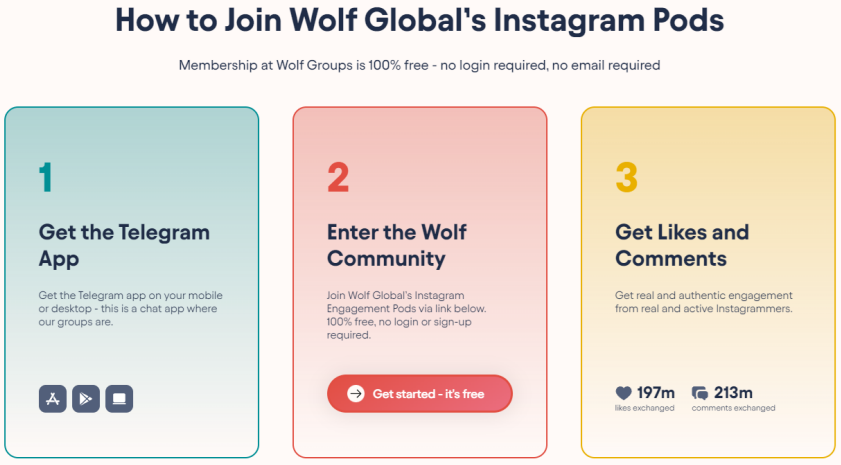
Weerasinghe, J., B. Flanigan, A. Stein, D. McCoy, and R. Greenstadt (2020): "The Pod People: Understanding Manipulation of Social Media Popularity via Reciprocity Abuse," in *Proceedings of The Web Conference 2020*, Taipei, Taiwan: Association for Computing Machinery, WWW '20, 1874–1884.

Wilhite, A. W. and E. A. Fong (2012): "Coercive Citation in Academic Publishing," *Science*, 335, 542–543.

Zinman, J. and E. Zitzewitz (2016): "Wintertime for Deceptive Advertisingö," *American Economic Journal: Applied Economics*, 8, 177–192.

A Online Appendix: Screenshots of Wolf Global Instagram Engagement Pods

(a) Main page



(b) How it works?

Figure A.1: Main page of `https://www.wolfglobal.org/`

Notes: Screenshots of `https://www.wolfglobal.org/`, taken on March 4, 2024.

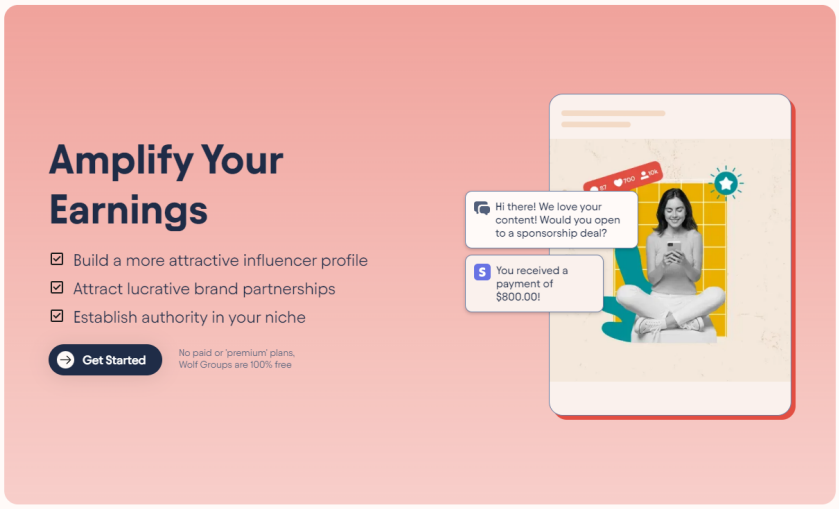


Figure A.2: Wolf Global Instagram Engagement Pods description of how influencers can amplify their earnings

Notes: Screenshot of `https://www.wolfglobal.org/`, taken on March 4, 2024.

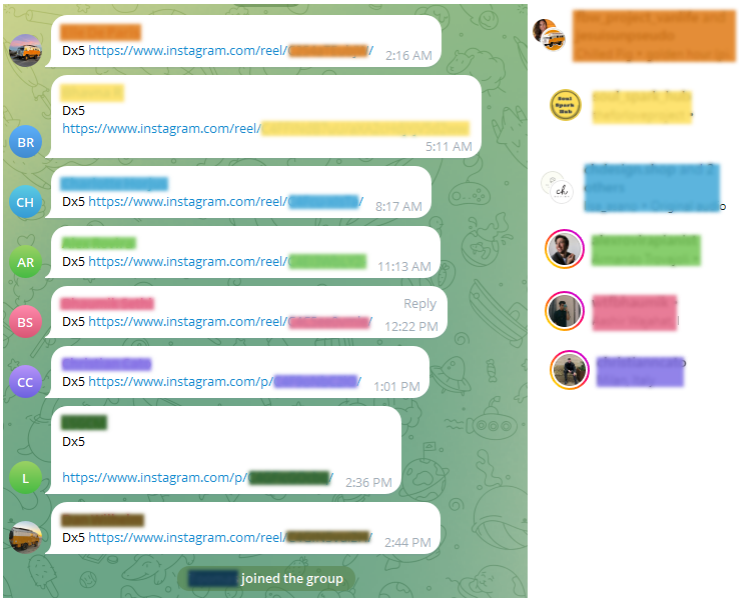


Figure A.3: Wolf Onyx Comments on Telegram app, mapped to Instagram users

Notes: Screenshot of Telegram Wolf Onyx Comments, taken on March 4, 2024.

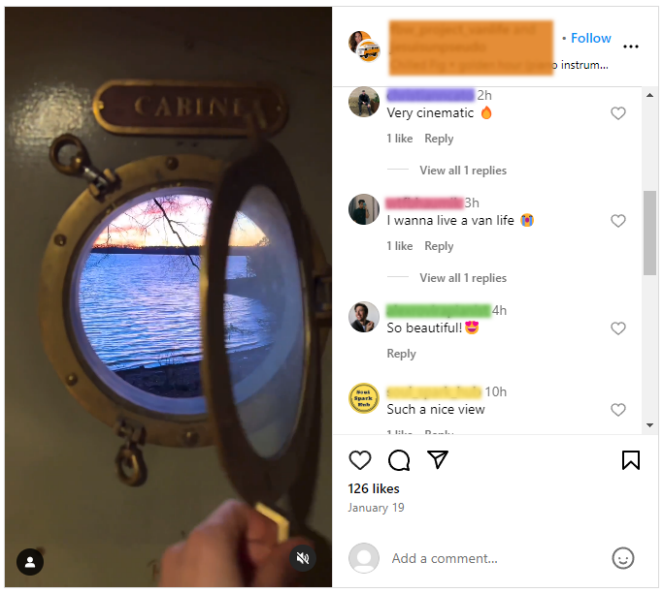


Figure A.4: Instagram comments coming from Wolf Onyx Comments

Notes: Screenshot of Instagram, taken on March 4, 2024.

B Online Appendix: Proofs

B.1 Proof of Proposition 1

Proof In equilibrium, player t chooses $a_t = 1$ if and only if the benefits outweigh the costs, $\gamma R_t \cos(\Delta_t) \geq R_t C(\Delta_t)$. As costs are always non-positive, it requires that $\Delta_t \leq 90°$ and thus $\gamma \cos(\Delta_t) \geq \sin(\Delta_t)$, which is equivalent to $\Delta_t \leq \tan^{-1}(\gamma)$.

On the other hand, engagement is socially optimal whenever total benefits outweigh the costs, $\gamma R_t \cos(\Delta_t) + (1-\gamma) R_t \cos(\Delta_t) \geq C(\Delta_t)$ or equivalently $\Delta_t \leq \tan^{-1}(1) = 45°$. As $\gamma < 1$, there exist socially optimal engagements that don't occur in equilibrium.

Finally, any engagement that is socially optimal, but not individually optimal, has strictly lower quality than compared to those that are individually optimal. That is, for any $\Delta_t \leq \tan^{-1}(\gamma)$, we have $\cos(\Delta_t) \geq \cos(\Delta_t')$ for any $\Delta_t' \in (\tan^{-1}(\gamma), 45°]$.

The proof can be summarized by the following figure B.1 □

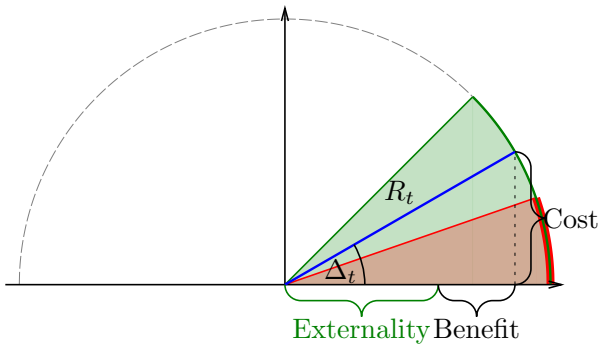


Figure B.1: Benefits and costs of an engagement

Notes: the angle in blue depicts a particular example, where the direct benefit is lower than the cost and therefore it is not optimal for the player. It would be optimal if the topics would be more similar, so that the angle would be within the red region. On the other hand, this engagement is socially optimal because the social benefit is greater than the cost.

B.2 Proof of Proposition 2

Proof Consider first the case when $\lambda < \gamma$. Then equation (4) is a product of two strictly negative values and therefore strictly positive, so that all players join the cartel. If $\lambda = \gamma$, then the expression simplifies to $\frac{4\gamma(1-\gamma)}{\gamma^2+1}\mathbb{E}R_{s_{t+1}} > 0$.

Next, suppose that $\gamma < \lambda < 1$. Then $u^{\text{cartel}}(R_{s_t})$ is strictly decreasing function of R_{s_t}, so the equilibrium must be characterized by a (possibly infinite) threshold $\overline{R}$, so that players join the cartel if and only if $R_{s_t} \leq \overline{R}$, which happens with probability $1 - \frac{1}{\overline{R}^2}$. Therefore, the expected reach of the following cartel member is $\mathbb{E}[R_{s_{t+1}} | R_{s_{t+1}} \leq \overline{R}] =$

$\frac{2}{1+\overline{R}^{-1}}$. This allows us to determine the marginal player's reach $\overline{R}$ as

$$u^{\text{cartel}}(\overline{R}) = \frac{4\lambda(\lambda-\gamma)}{\lambda^2+1}\left(\frac{1-\gamma}{\lambda-\gamma}\frac{2}{1+\overline{R}^{-1}} - \overline{R}\right) = 0 \quad \Longleftrightarrow \quad \overline{R} = \frac{2-\gamma-\lambda}{\lambda-\gamma}.$$

Finally, suppose that $\lambda = 1$, so that $\Lambda = 90°$. Then equation (4) simplifies to $2(1-\gamma)\left(\mathbb{E}R_{s_{t+1}} - R_{s_t}\right)$, which is positive only if the player's own reach R_{s_t} is smaller than the average reach. This means that only players with the lowest reach $R_{s_t} = 1$ would be willing to join, but we assume that the probability of such an event is zero.

To conclude the proof, observe that if $\lambda > 1$ (that is $\Lambda > 90°$), then the expression in equation (4) is strictly negative. Also, remember that equation (4) was computed with cost equal to $\sin(\Delta_{s_t})$. However, if $\Delta_{s_t} > 90°$ the cost is $C(\Delta_{s_t}) = 1 > \sin(\Delta_{s_t})$. Therefore the cartel payoff is strictly lower than the expression in equation (4). This implies that the cartel payoff would always be strictly negative and nobody would join the cartel. □

B.3 Proof of Proposition 3

Proof Fix $\Lambda > 0$ and $\overline{R} > 1$. The first term $u^{\text{cartel}}(\overline{R})$ in (10) is strictly negative, but bounded for player with reach $\overline{R}$. Therefore there exists value $v > 0$ such that the second term in (10) dominates it and the cartel payoff $u^{\text{cartel+ad}}(\overline{R})$ for player with reach $\overline{R}$. As $u^{\text{cartel}}(R_{s_t})$ is decreasing in R_{s_t}, we must have that $u^{\text{cartel+ad}}(R_{s_t}) > 0$ for all $R_{s_t} \leq \overline{R}$. □

B.4 Proof of Corollary 1

Proof We prove each part separately:

1. The realized welfare generated by engagement was defined by (1). When taking expectation over this expression over $\Delta_{s_{t+1}}$ that is distributed uniformly in $[0°, 180°]$, the benefit term becomes zero, whereas the expected cost is strictly positive, therefore such engagement is always reducing consumer welfare.

   Even when replacing general cartels with cartels that have lower Λ, this change would clearly be welfare-improving for consumers.

2. General cartels give influencers outside the cartels price of engagement equal to $p^{\text{engagement}} = (1-\varepsilon)p^{\text{natural}}$. Without the cartels, the price of engagement would be $p^{\text{natural}} > p^{\text{engagement}}$, while all other parts of the payoff would remain the same.

   Again, just reducing the engagement requirement Λ would increase the price of engagement, therefore increasing the welfare of influencers that do not belong to cartels.

3. The cartels cannot affect advertisers' profits as they are equal to zero due to the assumption that the market is perfectly competitive. Cartels drive down the price of engagement, so that advertisers who happen to be matched with an influencer who does not belong to a cartel, get a positive expected profit, as they pay less than the true value of engagement in expectation. However, advertisers who are matched with a general cartel members, pay for worthless engagement. The expectation over the two possibilities is zero, just the outcomes are more uncertain.

   As in earlier cases, replacing $\Lambda = 180^\circ$ with a slightly lower value lowers this uncertainty.

4. Consider a general cartel with engagement requirement equal to 180° and suppose v is large enough so that the cartel has members. It is straigtforward to check that there exists a marginal type, whose reach is $\overline{R}$, whose payoff of joining the cartel is zero.

   Consider a cartel with Λ that is marginally smaller than 180°. We claim that if $\varepsilon > 1/2$ then the payoff of $\overline{R}$ is strictly larger with Λ than with the general cartel. To see this observe that under such scenario, the expected reach of the follower, $\mathbb{E}[R_{s_{t+1}}|\text{Cartel}]$, is strictly increased. This is because the former marginal player $\overline{R}$ now gets strictly positive value and the payoff function is decreasing in reach, so the new marginal player has higher reach.

   Now, the payoff of $\overline{R}$ from joining the cartel, $u^{\text{cartel+ad}}(\overline{R})$, as defined by equation (10), has two elements. It is easy to see that, $u^{\text{cartel}}(\overline{R})$, is strictly increased. Moreover, the payoff from advertising is

$$\frac{\Lambda}{180^\circ}(1-\gamma)\mathbb{E}[R_{s_{t+1}}|\text{Cartel}]p^{\text{engagement}}, \tag{12}$$

   where the term $(1-\gamma)\mathbb{E}[R_{s_{t+1}}|\text{Cartel}]$ is again increased by the same argument.

   What remains to show is that the remaining term $\frac{\Lambda}{180^\circ}p^{\text{engagement}}$ is increased. By equation (9),

$$\frac{\Lambda}{180^\circ}p^{\text{engagement}} = v\frac{\Lambda}{180^\circ}\left[(1-\varepsilon)\frac{\sin(\Lambda^{eq})}{\Lambda^{eq}} + \varepsilon\frac{\sin(\Lambda)}{\Lambda}\right]$$

where $\Lambda^{eq} = \tan^{-1}(\gamma)$.[19] If we differentiate $\frac{\Lambda}{180^\circ} p^{\text{engagement}}$ with respect to Λ, we get

$$v\frac{1}{180^\circ}\left[(1-\varepsilon)\frac{\sin(\Lambda^{eq})}{\Lambda^{eq}} + \varepsilon\frac{\sin(\Lambda)}{\Lambda}\right] + v\frac{\Lambda}{180^\circ}\left[\varepsilon\frac{\cos(\Lambda)}{\Lambda} - \varepsilon\frac{\sin(\Lambda)}{\Lambda^2}\right].$$

To analyze a marginal decrease of Λ from 180°, we can take $\Lambda = 180^\circ$ and get

$$\frac{v}{180^\circ}\left(\frac{(1-\varepsilon)\sin(\Lambda^{eq})}{\Lambda^{eq}} - \varepsilon\right).$$

This expression is negative if and only if

$$\varepsilon > \frac{\sin(\Lambda^{eq})}{\Lambda^{eq} + \sin(\Lambda^{eq})} = \frac{\gamma}{\gamma + \tan^{-1}(\gamma)\sqrt{\gamma^2+1}} < \frac{1}{2}.$$

Therefore, indeed, if $\varepsilon > 1/2$, the payoff of $\overline{R}$ is strictly higher under Λ than under 180°.

Finally, note that the same argument applies for any member with $R_{s_t} \leq \overline{R}$. This is because $u^{\text{cartel}}(R_{s_t})$ is decreasing in Λ near 180° for the same reason as above, and the second term, i.e., the payoff from advertising market is independent of R_{s_t} and therefore decreasing as well.

□

C Additional Theoretical Results

C.1 Welfare from Cartels

Using the equilibrium description from proposition 2, we can now study the welfare implications of the cartel. As with individual payoffs, we normalize social welfare without any engagements to zero. Then the social welfare generated by the cartel, which we denote again by W, is proportional to the average payoff of all players in the model.[20] It is useful to compute also another measure V^{cartel}, which denotes the average payoff of cartel members. Both measures depend on the engagement requirement Λ, and it is

[19] Note that

$$p^{\text{cartel}} = v\mathbb{E}[\cos(\Delta_{s_{t+1}})|\Delta_{s_{t+1}} \leq \Lambda] = v2\int_0^\Lambda \cos(\Delta_{s_{t+1}})\frac{d\Delta_{s_{t+1}}}{2\Lambda} = v\frac{\sin(\Lambda)}{\Lambda}.$$

[20] Remember that the players extract constant fraction (β, normalized to 1) of social welfare as their own payoffs.

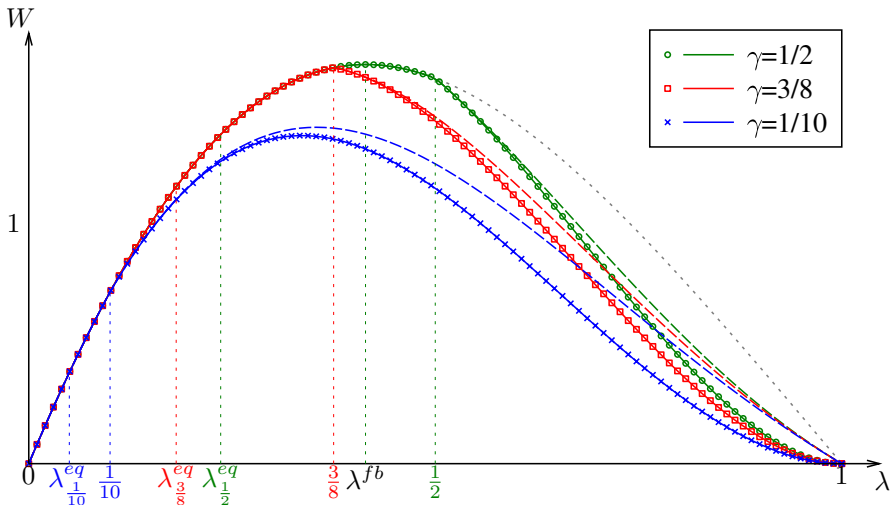


Figure C.1: Welfare as a function of engagement requirement λ for different free-riding parameters γ. λ^{fb} denotes the first-best engagement requirement, λ^{eq}_{γ} the equilibrium engagement threshold. Corresponding dashed lines indicate mean payoffs for cartel members (V^{cartel}).

convenient to express these in terms of the transformed version $\lambda = \tan\left(\frac{\Lambda}{2}\right)$. Formally,

$$V^{\text{cartel}}(\lambda) = \mathbb{E}_{R_{s_t}}\left[u^{\text{cartel}}(R_{s_t}) | u^{\text{cartel}}(R_{s_t}) \geq 0\right], \tag{13}$$

$$W(\lambda) = \mathbb{E}_{R_{s_t}}\left[\max\left\{0, u^{\text{cartel}}(R_{s_t})\right\}\right] = Pr(u^{\text{cartel}}(R_{s_t}) \geq 0) V^{\text{cartel}}(\lambda). \tag{14}$$

Using proposition 2, we can directly compute the welfare. We postpone the explicit calculations to the end of the subsection and use figure C.1 to discuss the results. Suppose that all players would belong to the cartel. Then the welfare would initially increase with the engagement requirement λ, as the social benefits exceed the social cost. The social welfare reaches the peak at the first-best level $\lambda^{fb} = \sqrt{2} - 1$ (corresponding to $\Lambda = 45°$) and then starts to decline, going back to zero at $\lambda = 1$ (corresponding to $\Lambda = 90°$). At this level, the average social cost is exactly equal to the average social benefit so that the welfare generated by engagement would be zero. The hypothetical welfare from a cartel that everyone joins it is an upper bound for welfare generated by the cartel and it is depicted by the dashed gray line on the figure. When $\lambda \leq \gamma$, the cartel achieves this upper bound, but if the engagement requirement is larger, some players choose not to join the cartel, and therefore the welfare is lower than the upper bound.

There are three qualitatively different possibilities for the free-riding parameter γ. First, high level, for example $\gamma = \frac{1}{2}$ (the green line with circle markers), the cartel can achieve the first-best outcomes by requiring first-best engagement λ^{fb}. Naturally, above this level, the welfare starts to decrease as costs exceed the benefits. At $\lambda = \gamma$ there is a kink due to a second distortion—above this engagement requirement, players with

the highest reach choose not to participate. At a moderate level, for example $\gamma = \frac{3}{8}$ (the red line with square markers), the first-best outcome is not achievable by the cartel because at λ^{fb}, players with the highest reach would not participate in the cartel. The welfare-maximizing engagement is $\lambda = \gamma = \frac{3}{8}$, this is the highest engagement where all players join the cartel. Finally, at low level, for example $\gamma = \frac{1}{10}$ (the blue line with cross markers), the optimal cartel is such that not all players join the cartel. It balances the trade-off between requiring more engagement and excluding fewer high-reach players. Figure C.1 also shows the mean payoffs to cartel members, V^{cartel}, which coincides with W when $\lambda \leq \gamma$ as all players join the cartel, but is strictly higher when λ is higher, as it does not account for the fact that the cartel only includes a fraction of influencers.

These results are formally characterized by the following corollary, where γ^{inc} is defined as

$$\gamma^{inc} = \frac{1}{3}\left(-2 - \frac{11}{\sqrt[3]{64+9\sqrt{67}}} + \sqrt[3]{64+9\sqrt{67}}\right) \approx 0.3444. \tag{15}$$

Corollary 2. *Depending on γ, we have one of three cases:*

1. *If $\gamma \geq \lambda^{fb}$, then first-best outcomes are achieved by a cartel with $\lambda = \lambda^{fb}$. Both $V^{cartel}(\lambda)$ and $W(\lambda)$ are strictly increasing in λ for $\lambda < \lambda^{fb}$ and strictly decreasing for $\lambda > \lambda^{fb}$.*

2. *If $\gamma^{inc} \geq \gamma < \lambda^{fb}$, then first-best outcomes are not achievable by a cartel and the welfare maximizing engagement is $\lambda = \gamma$, the highest λ where all players join the cartel. Again, both $V^{cartel}(\lambda)$ and $W(\lambda)$ are strictly increasing in λ for $\lambda < \gamma$ and strictly decreasing for $\lambda > \gamma$.*

3. *If $\gamma < \gamma^{inc}$, then the first-best outcomes are not achievable by a cartel. Welfare-maximizing $\lambda^* \in (\gamma^{inc}, 1)$ involves some players staying out of the cartel.*

Proof By proposition 2, when $\lambda \leq \gamma$, all players join the cartel and therefore $\mathbb{E}R_{s_t} = \mathbb{E}R_{s_{t+1}} = 2$, so that

$$W(\lambda) = V^{\text{cartel}}(\lambda) = \frac{4\lambda(1-\lambda)}{\lambda^2+1}\mathbb{E}[R] = \frac{8\lambda(1-\lambda)}{\lambda^2+1}. \tag{16}$$

This expression is strictly increasing for $\lambda \in [0, \lambda^{fb})$ and strictly decreasing for $\lambda \in (\lambda^{fb}, 1]$.

When $\gamma < \lambda < 1$, some players with highest reach choose not to join the cartel. By proposition 2, then the expected reach of a cartel member is $\mathbb{E}[R_{s_t} | R_{s_t} \leq \overline{R}] = \frac{2}{1+\overline{R}^{-1}} =$

$\frac{2-\gamma-\lambda}{1-\gamma}$. Therefore, the expressions become

$$V^{\text{cartel}}(\lambda) = \frac{4\lambda(1-\lambda)}{\lambda^2+1}\frac{2-\gamma-\lambda}{1-\gamma}, \tag{17}$$

$$W(\lambda) = Pr(R_{s_t} \leq \overline{R})V^{\text{cartel}}(\lambda) = \frac{16\lambda(1-\lambda)^2}{(\lambda^2+1)(2-\gamma-\lambda)}. \tag{18}$$

The derivative of $W(\lambda)$ is

$$W'(\lambda) = \frac{16(1-\lambda)\left(\gamma\lambda^3+\gamma\lambda^2+3\gamma\lambda-\gamma-6\lambda+2\right)}{\left(\lambda^2+1\right)^2(2-\gamma-\lambda)^2.}$$

For brevity, let us denote

$$w(\lambda) = \gamma\lambda^3+\gamma\lambda^2+3\gamma\lambda-\gamma-6\lambda+2.$$

Then $\operatorname{sgn} W'(\lambda) = \operatorname{sgn} w(\lambda)$. The function $w(\lambda)$ is a continuous, $w(0) = 2-\gamma < 0$ and $w(1) = -4(1-\gamma) < 0$, so $w(\lambda)$ has a root in $(0,1)$. Let us denote it by λ^*. Moreover, as $w(\lambda)$ is a polynomial, with leading coefficient $\gamma > 0$, $w(\lambda) > 0$ for sufficiently large λ and $w(\lambda) < 0$ for sufficiently small $\lambda < 0$. Therefore it must have one root in $(1,\infty)$ and one root in $(-\infty, 0)$. As it is a third-order polynomial, it has at most three roots. We have therefore determined that λ^* is its only root in $(0,1)$.

These arguments establish that $W'(\lambda^*) = 0$, $W'(\lambda) > 0$ for all $\lambda < \lambda^*$, and $W'(\lambda) < 0$ for all $\lambda > \lambda^*$. Therefore $W(\lambda)$ is maximized at λ^*. If we set $\gamma = \lambda$, we get a polynomial $w(\lambda) = \lambda^4+\lambda^3+3\lambda^2-7\lambda+2$. In this case, we can directly check the roots and see that it again has a unique root in $(0,1)$, which is γ^{inc} defined by equation (15). The combination of these observations proves all claims for $V^{\text{cartel}}(\lambda)$.

The proof for $V^{\text{cartel}}(\lambda)$ is analogous, with the exception that its derivative with respect to λ has slightly higher root $\lambda^{**} > \lambda^*$. Notice that if we set $\gamma = \lambda$ to this expression, we get the same polynomial as before and its root is again γ^{inc}. This is not surprising, because at the limit $\lambda = \gamma = \gamma^{inc}$ all players participate the cartel and therefore $V^{\text{cartel}}(\lambda)$ coincides with $W(\lambda)$. □

C.2 Entry requirements to the cartel.

Our model can also shed some light on the reasons why influencer cartels in practice often impose entry requirements. A typical requirement is to have at least some minimum number of followers, ranging from 1,000 to 100,000 in our sample.

We saw that the cost of joining the cartel depends on player's own reach, while the benefit depends on the average reach of a cartel member. By imposing a minimum

entry requirement for reach, the cartel can increase the average reach, making the cartel more appealing for players with higher reach. The combination of these two effects raises the average reach and benefits all members. Therefore we would expect the entry requirement to raise the average benefits for the cartel member, $V^{\text{cartel}}(\lambda)$. On the other hand, excluding players with low reach means that fewer players are eligible to join the cartel, which may reduce the social welfare, $W(\lambda)$. The following proposition confirms this intuition.

Proposition 4. *Suppose that in addition to engagement requirement $\Lambda > 0$, the cartel imposes an entry requirement $\underline{R} > 1$, so that only players with $R_t \geq \underline{R}$ are eligible to join. The mean payoff of a cartel member, $V^{cartel}(\lambda)$, is proportional to $\underline{R}$ and the mean payoff of a player, $W(\lambda)$, is proportional to $\underline{R}^{-1}$.*

The cartel may therefore choose to restrict the eligibility as such a restriction would raise the cartel member's welfare. On the other hand, the cartel organizer must be wary of the downside—eligibility restriction reduces the number of cartel members and this effect is large enough to reduce the overall welfare. If there is a single cartel, it depends on the cartel organizer's objective whether the restriction is beneficial. However, it is easy to imagine an extension where multiple cartels can be arranged: some that focus on smaller players who will then engage more actively, and others that limit access to large players and require less engagement. As we see in the data, this is what happens in practice.

Proof If $\lambda \leq \gamma$, then by the same arguments as above, all eligible players join the cartel, and therefore the expected reach of cartel members is $\mathbb{E}(R_{s_t}|R_{s_t} \geq \underline{R}) = 2\underline{R}$. The mean payoff for cartel members is

$$V^{\text{cartel}}(\lambda) = \frac{8\lambda(1-\lambda)}{\lambda^2+1}\underline{R}.$$

This is the same expression as above, in equation (16), but multiplied with $\underline{R}$. The difference is that now players with $R_t < \underline{R}$ cannot join. Their probability that player is eligible is $Pr(R_t \geq \underline{R}) = \underline{R}^{-2}$. Therefore the social welfare is

$$W(\lambda) = \frac{8\lambda(1-\lambda)}{\lambda^2+1}\underline{R}^{-1}.$$

Again, the same expression as equation (16), but now multiplied with $\underline{R}^{-1}$.

Suppose now that $\gamma < \lambda < 1$. By the same arguments as before, only players with a reach below marginal value $\overline{R}$ will join the cartel. Therefore average reach of a cartel

member is now

$$\mathbb{E}\left[R_{s_t}|\,\underline{R} \leq R_{s_t} \leq \overline{R}\right] = \frac{\int_{\underline{R}}^{\overline{R}} R_{s_t} 2R_{s_t}^{-3} dR_{s_t}}{\int_{\underline{R}}^{\overline{R}} 2R_{s_t}^{-3} dR_{s_t}} = \frac{2}{\underline{R}^{-1} + \overline{R}^{-1}}.$$

Using this value, we can now compute the marginal type using $u^{\text{cartel}}(\overline{R}) = 0$ and get $\overline{R} = \frac{2-\gamma-\lambda}{\lambda-\gamma}\underline{R}$. Therefore the expected reach of a cartel member is in equilibrium $\mathbb{E}(R_{s_t}|\underline{R} \leq R_{s_t} \leq \overline{R}) = \frac{2-\gamma-\lambda}{1-\gamma}\underline{R}$. Inserting this to the expected payoff expression gives the expected payoff for a cartel member,

$$V^{\text{cartel}}(\lambda) = \frac{4\delta\lambda(1-\lambda)}{\lambda^2+1}\frac{2-\gamma-\lambda}{1-\gamma}\underline{R}.$$

Again, this expression is identical with the unconditional payoff expression, just scaled with $\underline{R}$. Finally, the probability that a player is eligible and chooses to join the cartel is

$$Pr(\underline{R} \leq R_{s_t} \leq \overline{R}) = \underline{R}^{-2} - \overline{R}^{-2} = \frac{4(1-\gamma)(1-\lambda)}{(2-\gamma-\lambda)^2}\underline{R}^{-2}.$$

Therefore the social welfare is

$$W(\lambda) = \frac{16\delta\lambda(1-\lambda)^2}{(\lambda^2+1)(2-\gamma-\lambda)}\underline{R}^{-1}.$$

In each case, our findings are the same. Increasing $\underline{R}$ increases mean cartel member's payoff, $V^{\text{cartel}}(\lambda)$ from the cartel linearly. However, it reduces cartel membership quadratically and therefore reduces the overall average payoff $W(\lambda)$ linearly. □

D Online Appendix: Data Collection

D.1 Telegram cartel history

We collected Telegram cartel interaction history for 9 cartels: 6 general cartels (1K, 5K, 10K, 30K, 50K, 100K) and 3 topic cartels (fashion & beauty, health & fitness, travel & food). The 9 cartels were formed the earliest in August 2017 (10K and 50K) and the latest in February 2018 (5K). We downloaded the data in June 2020. In June 2020 all cartels had new posts.

The Telegram cartel interaction history consists of three pieces of information: Telegram username, Instagram post shortcode, and time. The interaction history tells us which Telegram user, added when, and which Instagram post to the cartel. According to the cartel rules, this information allows to determine which cartel member has to comment and like which post. This is because one has to comment and like five posts by other users directly preceding one's own.

D.2 Mapping Telegram posts to Instagram users

The Telegram cartels included 220,893 unique Instagram posts that we were able to map to 21,068 Instagram users. Specifically, the Telegram cartels included 316,462 unique Instagram posts altogether. Some posts are posted multiple times and/or multiple cartels, the 316,462 unique Instagram posts were posted in total 527,498 times. The cartel interaction files don't include the Instagram username of the author of the Instagram post. We mapped the Instagram posts included in cartels to Instagram users using the following interactive procedure. For the first Instagram post in the cartel of each Telegram user, we searched for the post on Instagram to learn the Instagram username of the post's author. Then we obtained from CrowdTangle the full list of all the Instagram posts of that Instagram username and matched those to the posts in the cartel. We checked the remaining unmatched posts in the cartel one by one until we either found a match for it on Instagram or determined that the post had been deleted on Instagram or made private. In this way, we were able to determine the Instagram usernames of 70% of the posts in the cartels, altogether 21,068 Instagram usernames.

Of the 21,068 Instagram users, 22% of users had posted in both topic and general cartels. Altogether, 11,158 users had posted in topic cartels and 14,566 users in general. This includes 4,656 users who had posted in both. Hence, the total number of users equals 11,158 + 14,566 - 4,656 = 21,068.

From CrowdTangle, for all the 21,068 Instagram users, we obtained the history of all their Instagram posts. This data included the time of the post, the text of the post

including the tags, and the number of comments and likes. The data was downloaded from August 19 to September 16, 2021.

D.3 Instagram comments

For each Instagram user, for their first post in cartels (no matter which cartel), we collected information on who commented on the post. Our goal was to learn who engaged with the post and compare natural engagement to that obtained via cartels. Therefore, we did not collect the comment itself, but only the username that posted the comment. We focused only on the comments (commenters) on the post itself, not comments on a comment.

We focused on each user's first post in pods, to minimize the possibility that involvement in pods had affected pod members' natural engagement. However, when the first post did not have enough information for the analysis, that is, when for the first post none of the pod members who were required to comment existed anymore, then we focused on the second (if the second post existed) and so on. Note we did not require that the pod members actually commented, only that the users still existed. For 18 users no post existed that satisfied the requirement reducing the sample to 21,050 users. Among the remaining 21,050 users, for 99.8% (20,999), it was of their actual first posts. In the robustness analysis, we restrict attention to the sample of the actual first posts. To simplify the exposition, in going forward, we call all the first posts satisfying the requirement, simply the first posts.

We used Apify to collect the comments. It allows access to the comments that are available without logging in to Instagram and provides only up to 50 comments for each post. The comments were downloaded in January 2024.

We were able to collect comments only for 16,630 posts, which is 79.0% of the total 21,050 first posts. We could not collect comments for all the first posts, because these posts either did not have any comments but mostly because we attempted to collect these comments more than two years after collecting posts itself, and in two years these users or their posts were either deleted or made private. Of those posts we were able to collect comments on, some did not have any non-pod commenters and we were left with 16,386 posts. Hence, we were able to find non-pod comments for 77.8% of the total 21,050 first posts. For both topic and non-topic cartels the percentage of first posts for which we got non-pod comments was similar, 78%.

D.4 Random non-cartel commenter on each cartel member's first cartel post

For each pod member's first post in pods, we used a random number generator and picked a random non-pod user who had commented on the post. We picked these random non-pod users for 16,386 posts. Some of the randomly picked non-pod commenting users were the same across posts. Hence, we were left with 14,490 unique non-pod commenting users.

For these 14,490 non-pod users, we collected their information about their number of public posts. We collected this information using Apify in January 2024. Of these users, 24 didn't exist anymore. So that were were left with 14,466 non-pod users. Of those, 3,049 (21.0%) were private. So that were were left with 11,417 public non-pod users. We then limited the sample to public non-pod users who had at least 10 posts and this restriction reduced the sample to 10,394 non-pod users.

For these random non-pod users, we obtained the history of all their Instagram posts from CrowdTangle. We did this to calculate its similarity to the author of the post and the users from the pod who posted a comment and were required to do that according to the cartel rules. For these non-pod users, the data was downloaded from CrowdTangle in January 2024. We were able to get the history only for 10,280 (99%). For the remaining 114 usernames either they had changed the username, made the account private or deleted it. Furthermore, we learned that 551 (5%) of the 10,280 non-pod users were associated with pod members as they had posted at least one post associated with the pod member. The association can happen as Instagram allows post to be associated with multiple users (this is different from tagging a user) or it could happen when user changes usernames. We excluded those 551 non-pod users from our sample, while keeping the corresponding pod members. That reduced the sample of non-pod users to 9,729. These 9,729 non-pod users mapped to 10,683 first posts because, as said above, some of these users were commenting on multiple posts.

E Online Appendix: Additional Tables and Figures for Empirical Analysis

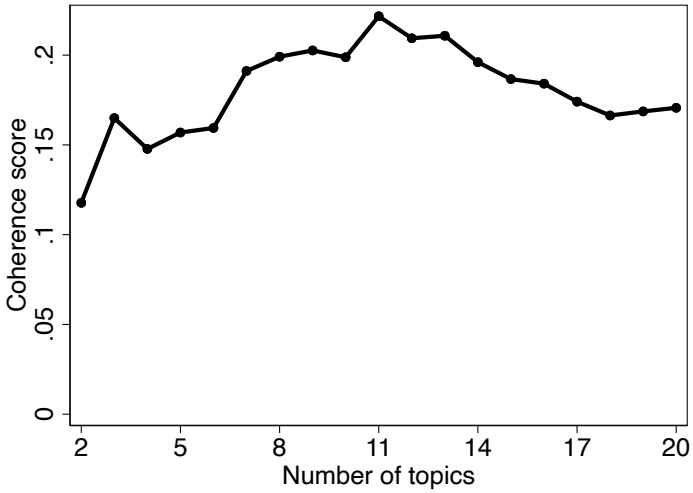
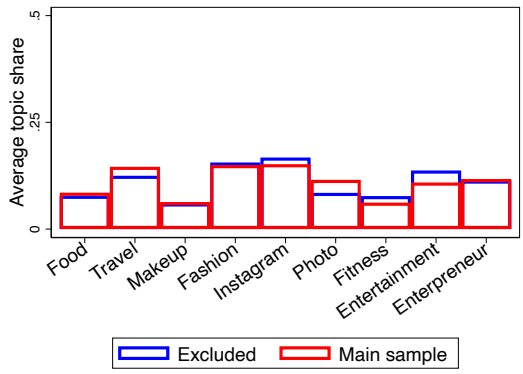


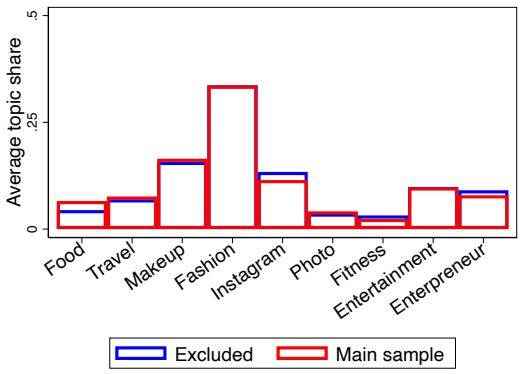
Figure E.1: LDA coherence scores by the number of topics

Table E.1: Most informative hashtags for each LDA topic

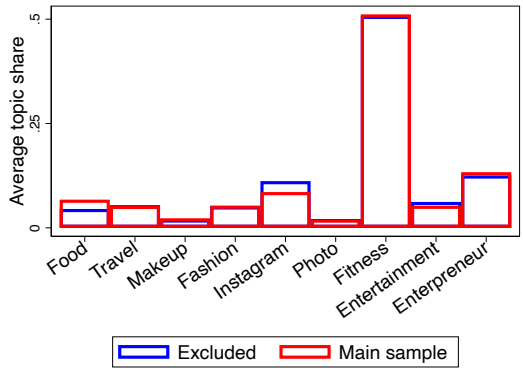
Topic number	Topic label	Hashtags
Topic 1	Food	#foodie #foodporn #food #instafood #liketkit #foodphotography #foodblogger #foodstagram #yummy #foodgasm
Topic 2	Travel	#travel #travelgram #travelblogger #travelphotography #wanderlust #instatravel #traveling #travelling #nature #traveltheworld
Topic 3	Makeup	#makeup #beauty #skincare #ad #makeupartist #mua #beautyblogger #hair #wakeupandmakeup #hudabeauty
Topic 4	Fashion	#fashion #ootd #fashionblogger #style #streetstyle #fashionista #blogger #model #styleblogger #outfitoftheday
Topic 5	Instagram	#instagood #photooftheday #photography #picoftheday #beautiful #instadaily #nature #instagram #happy #follow
Topic 6	Photo	#agameoftones #moodygrams #artofvisuals #beautifuldestinations #art #exploretocreate #photography #visualambassadors #justgoshoot #streetphotography
Topic 7	Fitness	#fitness #workout #gym #motivation #fitnessmotivation #fitfam #fit #bodybuilding #health #training
Topic 8	Entertainment	#music #nyc #losangeles #wedding #hiphop #artist #dance #newyork #dj #art
Topic 9	Enterpreneur	#motivation #entrepreneur #success #business #inspiration #quotes #motivationalquotes #mindset #entrepreneurship #wine



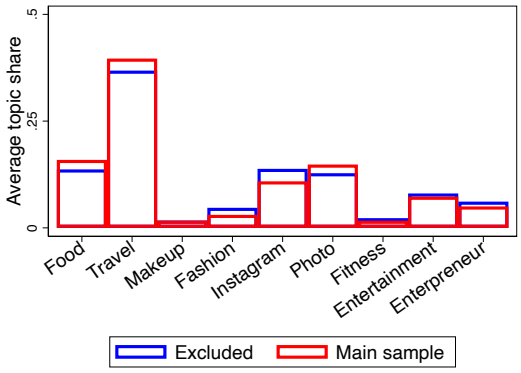
(a) General cartels



(b) Fashion & beauty cartel



(c) Fitness & health cartel



(d) Travel & food cartel

Figure E.2: Average topic distribution in the main sample versus excluded cartel members