Influencer Cartels*

Marit Hinnosaar† Toomas Hinnosaar‡

June 17, 2022§

First draft: February 12, 2021

Abstract

Influencer marketing is a large and growing but mostly unregulated industry. The majority of influencers are not paid based on their marketing campaigns' success. Instead, their prices are based on engagement (number of likes and comments). This gives incentives for fraudulent behavior—for inflating engagement. We study influencer cartels, where groups of influencers collude to increase engagement to improve their market outcomes. Our theoretical model shows that such cartels mitigate the free-rider problem and may increase or decrease welfare, depending on the quality of induced engagement. We use a novel dataset of Instagram influencer cartels and confirm that the cartels increase engagement as intended. Importantly, we show that engagement from non-specific cartels is of lower quality, whereas engagement from topic-specific cartels may be as good as natural engagement. Therefore topic-specific cartels may sometimes be welfare-improving, whereas typical non-specific cartels hurt everyone.

JEL: L41, C72, L86, M31, D26

Keywords: collusion, influencers, cartels, free-riding, commitment, cosine similarity, LDA, marketing

*We would like to thank seminar participants at University of East Anglia, University of Leicester, University of Nottingham, Collegio Carlo Alberto, Mannheim Virtual IO Seminar, 12th Paris Conference on Digital Economics, IIOC, Munich Summer Institute, Baltic Economics Conference, Australasian Meeting of the Econometric Society, Econometric Society European Meeting, EARIE, and NBER Summer Institute on IT and Digitization.

†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 a value about 8 billion dollars in 2019.[1]

The majority of influencers are not paid based on their marketing campaigns' success, instead, their prices are based on reach and engagement (number of followers, likes, and comments). This gives incentives for fraudulent behavior—for inflating their influence. An estimated 15% of influencer marketing spending was misused due to exaggerated influence.[2] There are many ways to exaggerate influence. In this paper, we are studying one of these—influencer cartels.

In an influencer cartel, a group of influencers collude to inflate their engagement in order to increase their prices. In traditional industries, a cartel is a formal (often secret) agreement to manipulate the market for members' benefit, typically involving 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 chat rooms, where members submit links to their content for additional engagement. In return, they agree to engage with other members' content. An algorithm enforces the cartel rules.

In this paper, we build a theoretical model and analyze a novel dataset of influencer cartels. Our goal is to answer three questions. First, what are the welfare implications of influencer cartels? Second, how do influencer cartels work in practice? Third, how to regulate influencer cartels?

To study the welfare implications of the influencer cartels, we build a theoretical model of influencer engagement. In this market, the key distortion is the free-rider problem, as 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

[1]Source: Audrey Schomer, Dec 17, 2019, "Influencer Marketing: State of the social media influencer market in 2020", Business Insider. `https://www.businessinsider.com/influencer-marketing-report`.

[2]Source: Daniel Carnahan, Nov 15, 2019, "Facebook has released Instagram content moderation data for the first time", Business Insider. `https://www.businessinsider.com/facebook-shares-instagram-content-moderation-data-for-first-time-2019-11`.

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, therefore, 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 use data from two sources: influencer cartels and 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: one topic-specific (fitness and health) and six others with unrestricted topics. Altogether, the cartels include almost 20,000 members. A typical cartel member has about 10,000 followers on Instagram.

We use natural language processing and machine learning 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 Instagrammer who engages. To quantify the similarity of Instagrammers, first, we use latent Dirichlet allocation (LDA) to map each Instagrammer's content to a probability distribution over topics. This allows us to compare the topics of the influencer and the Instagrammer who engages with the influencer's content. Second, we calculate a pairwise cosine similarity score for each influencer and the engaging Instagrammer pair. The cosine similarity score gives us a summary measure of the similarity of their content.

Using this data, we confirm that the cartels increase user engagement on Instagram as intended. We show that the engagement that originates from non-specific cartels is of lower quality in terms of the content match. But engagement originating from a topic-specific cartel is almost as good as natural.

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.

The paper contributes to three fields. First, it adds to a small but growing literature 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, as well as the benefits of mandatory advertising disclosure (Fainmesser and Galeotti, 2021; Pei and Mayzlin, 2022; Mitchell, 2021; Berman and Zheng, 2020). In contrast to these papers, our focus is on collusion between the influencers.

Second, the paper adds to the empirical literature on the operation of cartels.[3] 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 enforced by an automated bot.

Third, 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

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

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 economics literature that uses text as data.[4] In particular, we are 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. (2019) and Hinnosaar et al. (2021).

Trade-offs similar to our model arise in other settings, including patent pools, record sharing, and citation cartels. 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. Finally, in recent decades, there has been a growing concern about citation cartels (Franck, 1999).[5] Researchers who cite other works create a positive externality. By agreeing to cite more within a certain group, both individual researchers and journals could boost their observable impact. Within a certain limit, this may be helpful for the readers, but not if done excessively. Van Noorden (2013) and Wilhite and Fong (2012) have studied citation cartels. In contrast to the settings above, in influencer cartels, the collusion and outcomes are directly observable.

The rest of the paper is organized as follows. In the next section, we provide some institutional details. 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. All proofs are in appendix A.

[4]For a recent survey of the uses of text as data in economics, see Gentzkow et al. (2019).

[5]For example, Thomson Reuters regularly excludes journals from the Impact Factor listings due to anomalous citation patterns: `http://help.prod-incites.com/incitesLiveJCR/JCRGroup/titleSuppressions`.

2 Institutional Background

Instagram is a social network for sharing photos and videos. Instagram users engage with other users' content by liking and commenting their posts and can follow other users to see more of their content. As of 2019, Instagram had about one billion active users.[6] In the US, 56% of internet users aged 16–64 use Instagram.[7] As of February 2021, it is the 19th most popular website in the US according to the Alexa ranking.[8] Instagram is owned by Facebook. Instagram earns its revenue from advertising. In 2019, Instagram generated about 20 billion USD in advertising revenue.[9] The ads from which Instagram earns revenue are not influencers' posts, instead these are ads generated by businesses. As of 2017, one million firms were advertising on Instagram.[10]

In influencer marketing, firms pay influencers for product placement and product endorsement. In 2020, 65% of the member firms of the Association of National Advertisers in the US used influencer marketing (ANA, 2020), and the majority expected to use it more in the future.

Who can be an influencer? Any person with a large enough audience to have some influence over other consumers' choices. The largest influencers are athletes, musicians, and actors with hundreds of millions of followers, 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). The firms selected influencers mainly based on brand alignment and relevance (97%), content quality (95%), and engagement rate (95%).

The majority of influencers are not paid based on the actual success of the current marketing campaign.[11] As of 2020, only 19% of the firms using influencer marketing were tracking the sales induced by influencers (ANA, 2020). Instead, they are paid before the start of the campaign, based on their characteristics. Which characteristics? Initially, Instagram influencers were paid largely for the number of followers. This led to influencers

[6]Source: Emily S. Rueb, June 4, 2019, "Your Instagram Feed Is About to Have More Ads From Influencers", New York Times. `http://nyti.ms/2ZjBi2L`.

[7]Source: Mansoor Iqbal, Jan 28, 2021, "Instagram Revenue and Usage Statistics (2021)", Business of Apps. `https://www.businessofapps.com/data/instagram-statistics/`.

[8]Source: Feb 12, 2021, "Top Sites in United States", Alexa. `https://www.alexa.com/topsites/countries/US`.

[9]Source: Ellen Simon, Feb 7, 2021, "How Instagram Makes Money", Investopedia. `https://www.investopedia.com/articles/personal-finance/030915/how-instagram-makes-money.asp`.

[10]Source: Ken Yeung, Mar 22, 2017, "Instagram now has 1 million advertisers, will launch business booking tool this year", VentureBeat. `https://venturebeat.com/2017/03/22/instagram-now-has-1-million-advertisers-will-launch-business-booking-tool-this-year/`.

[11]The influencers with a large following are sometimes paid after the marketing campaign ends based on the actual success of the campaign. For example, their posts include links to online stores or coupons, which allow stores to track the number of product sales originating from the influencer. But most smaller influencers are paid before the start of the campaign.

getting fake followers. The industry then moved to detect fake followers and measure and compensate engagement—likes and comments. Nowadays, a combination of factors determines influencers' prices, including the number of followers, and importantly, the engagement rate on previous posts. This still creates conditions for allowing fraudulent behavior. It led influencers to use automatic bots that generate likes and comments. But automatic bots are relatively easy to detect and 60% of advertisers report that they vet influencers for fraud (ANA, 2020). These changes have motivated Instagram user cartels, where the engagement is generated by humans and is, therefore, more difficult to separate from the natural engagement.

Instagram influencer cartels. Instagram influencer cartels (called *pods*) are groups of influencers who agree to coordinate with each other to increase the engagement of their posts. The increased engagement brings a direct benefit with advertisers. Furthermore, the Instagram algorithm gives higher exposure to posts with higher engagement, which leads to even more engagement. Instagram considers the groups as violating Instagram's policies.[12]

The groups coordinate their work online using either a group chat or a discussion board.[13] Typically, a member adds a link of his Instagram post to the forum, and other group members have to comment and like that post on Instagram. In more secretive groups, a member posts only a code word to indicate that other members should go to his Instagram profile and like and comment on the latest post. Details of the rules are group-specific. For example, how many previous posts a member should engage with before adding a link to his own post. Or how quickly one should like and comment. Some groups have specific entry requirements, for example, regarding the minimum number of followers or topic of the Instagram content. The rules are enforced by automatic bots. Typically, the Instagram influencer cartels use other platforms such as Telegram or Reddit to coordinate their work. But other more secretive groups work on Facebook and Instagram itself.

Since the cartels' activity of artificially increasing engagement is fraudulent, the groups are secret, and there is not much aggregate information about these. The only scientific study of influencer cartels that we know of is by computer scientists Weerasinghe et al. (2020) who studied the characteristics of over 400 Instagram cartels in Telegram. On average, a cartel in their sample had about 900 members, but larger cartels had over

[12]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/`.

[13]For more details, see for example: Apr 9, 2019 "Instagram Pods: What Joining One Could Do For Your Brand", Influencer Marketing Hub. `https://influencermarketinghub.com/instagram-pods/`

10,000 members. Over 70% of cartels in their sample use the rule requiring that before a cartel member can add a link to his own Instagram post, he must first engage with the previous N Instagram posts added to the cartel, where a typical N equals 5 or 10, but ranges from 2 to 87. In their sample, the majority of cartels did not have any entry requirements regarding the number of followers nor topics and were easily discoverable using search engines. They also compared the engagement on cartel members posts over time before and after joining cartels, and found that after joining cartels, the number of likes and comments is 2 and 5 times larger, respectively. These numbers suggest that joining a cartel coincides with engagement growth. However, it doesn't measure the causal effect of cartels, because this doesn't take into account that over time engagement tends to increase for all users and most importantly, joining a cartel is a choice.

All cartels in our sample operate through Telegram chatrooms, where members submit their Instagram posts. The main requirement is 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. The rules are strictly enforced by an algorithm (a bot) that deletes submissions that are in an incorrect format. If the person submitting did not engage first, the bot removes the submission and gives the violating member a warning. Multiple warnings lead to a ban. 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 engagement without collusion and the advertising market. We then add collusion and, finally, the advertising market.

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).[14] The first parameter, angle $\alpha_t \in [0°, 360°]$, captures the topic of t's content. It can be thought of as a position in the Salop (1979) circular horizontal differentiation model. The

[14]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.

second parameter is the player's reach R_t, it measures his size of the audience (number of followers and typical search traffic).

In this paper, we focus on engagement. Each player has a piece of content. Player t chooses between two actions $a_t \in \{0, 1\}$: to engage with the previous player's content $a_t = 1$ or not to engage $a_t = 0$. In practice, engagement means commenting or liking other influencers' posts. We normalize all payoffs without engagement to zero.

Player t's choice to engage creates a social benefit and a social cost. 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.[15] The social benefit captures player t's provision of information and entertainment to his audience. It is therefore proportional to the size of the audience R_t and increasing in the similarity of topics, which is captured by the term $\cos(\Delta_t)$. 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 $\Delta \leq 90^\circ$ and 1 otherwise. The social cost represents the cost of attention of the audience. It increases in reach R_t and the distance between topics Δ_t. This functional form assumes that paying attention to something the reader regularly follows is quite costless, whereas consuming unrelated information is much more costly.[16] The difference between costs and benefits, $R_t(\cos(\Delta_t) - C(\Delta_t))$ therefore describes the total social value of engagement by player t. We differentiate the costs and benefits to model the positive externality.

We assume that player t who chooses to engage, internalizes fraction $\beta \in (0, 1)$ of the social costs, but only fraction $\beta\gamma < \beta$ of the social benefits. These assumptions encapsulate the long-term relationships with the audience. If player t provides high value to its audience, his followers continue to follow his content and take his product recommendations. Similarly, player t gets blamed for the additional attention costs he created, decreasing his payoff. The remaining fraction $\beta(1-\gamma)$ of the social benefit raises the payoff of player $t-1$, whose content received the additional attention. As factor β multiplies all influencers' payoffs related to engagement, we can, without loss in generality, normalize β to one.

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

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

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

[16] 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.

On the other hand, we define the social welfare as the average of all social benefits and costs. That is, $W(\{a_t\}_t)$, which is the average of individual terms generated by each action a_t:

$$W_t(a_t) = a_t \underbrace{R_t \cos(\Delta_t)}_{\text{Benefit}} - a_t \underbrace{R_t C(\Delta_t)}_{\text{Cost}}. \tag{2}$$

We assume that players' actions are not observable to the following players.[17] 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. There is a common knowledge that the each α_s is independent draw from uniform distribution in $[0°, 360°]$ and each R_s is distributed identically and independently in $[1, \infty)$ with power law distribution with mean 2. That is, the probability density function is $f(R_s) = 2R_s^{-3}$.[18]

We consider non-cooperative equilibria, i.e., 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. Figure 1 illustrates the costs and benefits of engagement, and non-cooperative equilibrium behavior and socially optimal outcomes. It depicts a particular example, where the (internalized) benefit for the influencer is lower than the cost and therefore it is not optimal for t to create engagement. It would be optimal if the topics would be more similar and therefore Δ_t smaller (angles within the thick red arc). On the other hand, this engagement would be socially optimal as as the social benefit (which is proportional to internalized benefit plus the externality) is greater than the cost (socially optimal engagement region is marked by thin green arc). It is the standard free-riding problem. Player t's action creates a positive externality for $t-1$, which t does not internalize. Therefore there is too little engagement in non-cooperative equilibrium.

One way to see the intuition of the formal results is to consider different values of γ. If $\gamma \to 0$, then this is a game with pure externality. Player t bears all the costs of $a_t = 1$ and gets none of the benefits. Therefore in equilibrium, we would expect no engagement. On the other hand, if $\gamma \to 1$, then player t internalizes all benefits of engagement, so we expect the non-cooperative equilibrium to coincide with the social optimum. As $\gamma \in (0, 1)$, player t internalizes the externality only partially, and therefore in equilibrium, there is too little engagement compared to the social optimum. The following proposition formalizes this intuition.

[17] This assumption allows us to exclude folk-theorem-type of equilibria, where players cooperate whenever there has been cooperation in the past.

[18] 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.

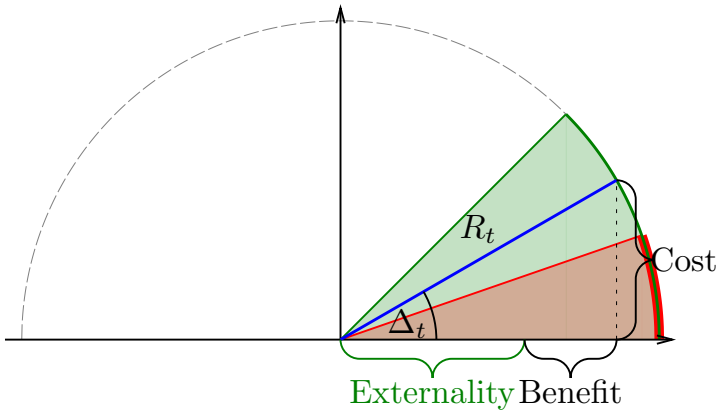


Figure 1: Benefits and costs of an additional engagement

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 social optimum and non-cooperative equilibrium behavior shows that there is room for improvement from cooperation. If players could compensate their followers for engaging more, they would be happy to do so within the limits of socially optimal engagement. Alternatively, if all players could pre-commit to socially optimal behavior, they would be happy to do so.

3.2 Influencer Cartels

We model a cartel as an entry game to a cartel agreement with parameter $\Lambda \geq 0$. After learning their own types (α_t, R_t), but before learning other players' types, players simultaneously choose whether to enter into the cartel. A player who does not join the cartel gets outside option, which we normalize to 0. Players who join, form their a subsequence of players $(\dots, s_{-1}, s_0, s_1, s_2, \dots)$, where s_t is the t'th member of the cartel.

A cartel agreement is defined by a parameter Λ and requires that each cartel member s_t must engage with the content of previous member s_{t-1} of the cartel whenever $\Delta_{s_t} = |\alpha_{s_t} - \alpha_{s_{t-1}}| \leq \Lambda$. We assume that the cartel is able to enforce this rule, but joining the cartel is voluntary so that each player t who joins the cartel must get at least the outside option of zero in terms of expected payoff. We focus on symmetric equilibria in the entry game, where players join the cartel independently of topic α_t. To simplify the expressions, it is useful to use a monotonic transformation $\lambda = \tan\left(\frac{\Lambda}{2}\right)$.

A player with type (α_{s_t}, R_{s_t}), who joins the cartel, gets payoff:

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

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. The interpretation of the payoff function is the same as before, but now the engagements are determined by the cartel agreement.

The cartel agreement parameter Λ captures the breadth of the cartel. In one extreme, if $\Lambda = 180°$, then it is a non-specific cartel that requires engagement regardless of the topic. Allowing $0° < \Lambda < 180°$ is a convenient way to model topic-specific cartel. The smaller is Λ, the more specific the cartel is, requiring engagement only in closely related topics. If $\Lambda \in (\tan^{-1}(\gamma), 1]$, i.e., if the engagement requirement is higher than the equilibrium engagement but lower than the socially optimal engagement, then the first term in equation (3) (i.e., the difference between direct benefits and costs of the required engagement) is negative, but the second term (additional engagement from the cartel) is positive. The first term can be interpreted as the cost and the second as the benefit from joining the cartel. The cost depends on the player's own reach R_{s_t}, whereas the benefit depends on the reach of the follower. An influencer is willing to join this cartel as long as the expected benefit of joining is greater than the expected cost.

Equilibria. As we are focusing on symmetric equilibria, the conditional distributions of Δ_{s_t} and $\Delta_{s_{t+1}}$ are uniform. Let us suppose for a moment that $\Lambda \leq 90°$, so that the cost function $C(\Delta_{s_t}) = \sin(\Delta_{s_t})$. Then the cartel benefit from equation (3) is[19]

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

Using this expression, we can now study the entry to the cartel and formalize this in proposition 2 below. There are three cases. If the engagement requirement Λ is low, then all players join the cartel. It is easy to see this when $\Lambda \leq \tan^{-1}(\gamma)$ as then even the direct benefits exceed the costs. But even if the engagement requirement is larger, benefits

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

the player expects from the cartel are larger than the costs of fulfilling the engagement requirement. The second region is when the entry requirement is moderate. In this case, some players join the cartel, and some do not. As the benefit of the engagement coming 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}, 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 Λ is sufficiently large, nobody joins the cartel. Moreover, if $\Lambda > 90°$, then equation (4) is just an upper bound for the cartel payoff, and it is strictly negative, so in this case, 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 implications. Using the equilibrium description, 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 mean payoff of all players in the model.[20] It is useful to compute also another measure V^{cartel}, which denotes the mean payoff of cartel members. Both measures depend on the engagement requirement Λ, and it is convenient to express these in terms of the transformed version $\lambda = \tan\left(\frac{\Lambda}{2}\right)$. Formally,

$$V^{\text{cartel}}(\lambda) = \mathbb{E}_{R_{s_t}}\left[u^{\text{cartel}}(R_{s_t}) | u^{\text{cartel}}(R_{s_t}) \geq 0\right], \tag{5}$$

$$W(\lambda) = \mathbb{E}_{R_{s_t}}\left[\max\left\{0, u^{\text{cartel}}(R_{s_t})\right\}\right] = Pr(u^{\text{cartel}}(R_{s_t}) \geq 0) V^{\text{cartel}}(\lambda). \tag{6}$$

Using proposition 2, we can directly compute the welfare. We postpone the explicit calculations to appendix A.3 and use figure 2 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

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

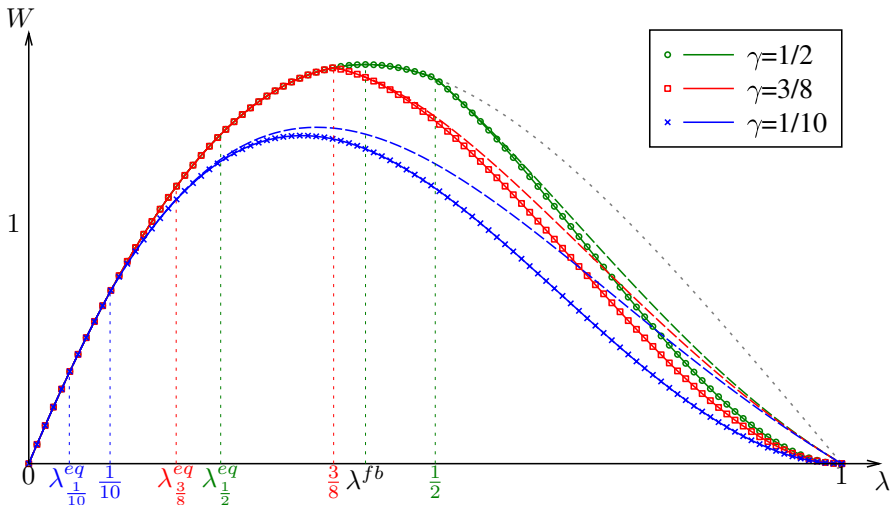


Figure 2: Welfare as a function of engagement requirement λ for different free-riding parameters γ. λ^{fb} denotes the first-best engagement requirement, λ^{eq}_{γ} the equilibrium engagement threshold. Corresponding dashed lines indicate mean payoffs for cartel members (V^{cartel}).

generated by engagement would be zero. This frontier is the upper bound for welfare generated by the cartel and is depicted by the dashed gray line on the figure. When $\lambda \leq \gamma$, the cartel can achieve this level, but as the engagement requirement is larger, some players choose not to join the cartel, and therefore the welfare is lower.

There are three qualitatively different possibilities for the free-riding parameter γ. First, high level $\gamma = \frac{1}{2}$ (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 $\gamma = \frac{3}{8}$ (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 $\lambda = \gamma = \frac{3}{8}$, i.e., the highest engagement where all players join the cartel. Finally, at low $\gamma = \frac{1}{10}$ (the blue line with cross markers), the optimal engagement would be interior. It balances the trade-off between requiring more engagement and excluding fewer high-reach players. Figure 2 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

$$\gamma^{inc} = \frac{1}{3}\left(-2 - \frac{11}{\sqrt[3]{64+9\sqrt{67}}} + \sqrt[3]{64+9\sqrt{67}}\right) \approx 0.3444. \tag{7}$$

Corollary 1. *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.*

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 to 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 3. *Suppose that in addition to engagement requirement $\Lambda > 0$, the cartel imposes an entry requirement $\underline{R} \geq 1$, 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.

3.3 Advertising Market

Each player t is matched with an advertiser with type $\alpha = \alpha_t$. The realized value of engagement from the follower $t+1$ to the advertiser is

$$\underbrace{a_{t+1}(1-\gamma)R_{t+1}}_{\text{quantity of engagement}} \times \underbrace{\cos(\Delta_{t+1})}_{\text{match quality}} \times \underbrace{v}_{\text{marginal value}} . \tag{8}$$

We assume that the advertising market is competitive, so that the player is able to capture this value as an addition to the payoff.[21]

We consider two main scenarios. The first scenario is *paying for the value* of engagement, which we model by assuming that the product $a_{t+1}(1-\gamma)R_{t+1}\cos(\Delta_{t+1})$ is contractible. Therefore, player receives $\max\{0, a_{t+1}(1-\gamma)R_{t+1}\cos(\Delta_{t+1})v\}$ in addition to other costs and benefits from engagements. This scenario captures situations where players are compensated for the value added, such as a percentage of added sales. The second scenario is *paying for the quantity* of engagement, which we model by assuming that only the quantity of engagement $a_{t+1}(1-\gamma)R_{t+1}$ is contractible. Then the advertiser needs to take an expectation of $\cos(\Delta_{t+1})$, given the available information, and the player gets $a_{t+1}(1-\gamma)R_{t+1}\mathbb{E}\cos(\Delta_{t+1})v$ as an additional payment.

We assume that the advertising market 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, and with the remaining probability ε it comes from a cartel. In the case of natural engagement, the equilibrium is unchanged, as the additional part of the payoff function does not depend on the player's own action a_t. Therefore in equilibrium player $t+1$ engages if and only if $\Delta_{t+1} \leq \tan^{-1}(\gamma)$. When the engagement comes from a cartel, the two scenarios lead to different conclusions, so we need to consider them separately.

Finally, we also assume that there is a continuum of cartels and each individual cartel is small. One cartel alone cannot change the advertising market's beliefs about the quality of engagement and, therefore, cannot affect engagement price.

[21] 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.

Paying for the value of engagement. The payoff function of a player joining the cartel is now

$$u^{\text{cartel+ad}}(R_{s_t}) = u^{\text{cartel}}(R_{s_t}) + \max\{0, a_{t+1}(1-\gamma)R_{t+1}\cos(\Delta_{t+1})v\}, \tag{9}$$

where $u^{\text{cartel}}(R_{s_t})$ is defined by equation (3). If $\lambda \leq 1$, then the part of the payoff function that is multiplied by the expected reach of the follower is multiplied by $(1+v)$ and therefore, the new expression for the cartel payoff is

$$u^{\text{cartel+ad}}(R_{s_t}) = \frac{4\lambda(\lambda-\gamma)}{\lambda^2+1}\left[(1+v)\frac{1-\gamma}{\lambda-\gamma}\mathbb{E}[R_{s_{t+1}}] - R_{s_t}\right].$$

This expression leads to the same qualitative conclusions as equation (4), with one difference. Without advertising, the critical value where no player joins the cartel was $\lambda = 1$. Now, at $\lambda = 1$, the payoff function is $u^{\text{cartel+ad}}(R_{s_t}) = 2(1-\gamma)\left[(1+v)\mathbb{E}[R_{s_{t+1}}] - R_{s_t}\right]$, which is strictly positive, at least for some players with low reach R_{s_t}. Therefore the marginal level of engagement, which we denote by $\overline{\lambda}$ is now greater than 1.

Proposition 4. *If the advertising market pays for the value of engagement, 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 $\gamma < \lambda < \overline{\lambda}$, all players with $R_t \leq \overline{R}$ join the cartel.*
3. *If $\lambda \geq \overline{\lambda}$, nobody joins the cartel,*

where $\overline{\lambda} > 1$ and $\overline{R} > 1$.

The main observation from the previous proposition is that the equilibrium behavior remains unchanged qualitatively, but cooperation through the cartel is now easier to sustain. In other words, players are willing to join cartels with bigger engagement requirements because they benefit more from engagement. If v is sufficiently large, even non-specific cartel with $\Lambda = 180°$ would attract some members.

What kind of engagement requirement would we expect if the cartel rules are fixed by a third party who optimizes a function depending on players' payoffs? Notice that we can think both welfare functions $V^{\text{cartel}}(\lambda)$ and $W(\lambda)$ as a sum of corresponding measure without advertising plus advertising payoffs. We already saw above that both measures without advertising are strictly decreasing above the first-best engagement level $\lambda^{fb} < 1$ and zero for $\lambda \geq 1$. The advertising payoff is zero beyond $\lambda = 1$. Therefore the optimal

cartel requires engagement level λ, which is higher than in corollary 1, but strictly less than 1. Therefore we can state the following corollary.

Corollary 2. *If advertising market pays for the value of engagement, there exists $\lambda^* < 1$ such that both $V^{cartel}(\lambda)$ and $W(\lambda)$ are strictly decreasing for all $\lambda > \lambda^*$.*

As an implication of this observation, although we do not take an explicit stand on the cartel organizer's incentives, we would expect that the cartel would never set $\lambda > 1$. There is no value-added in creating counter-productive engagements. Therefore, we can conclude that while the advertising market that pays for value makes distortionary cartels easier to sustain, these distortions are limited and qualitatively similar to the case without advertising market.

Paying for the quantity of engagement. The payoff function of a player joining the cartel is now

$$u^{\text{cartel+ad}}(R_{s_t}) = u^{\text{cartel}}(R_{s_t}) + a_{t+1}(1-\gamma)R_{t+1}p^{\varepsilon}, \tag{10}$$

where $p^{\varepsilon} = v\mathbb{E}\cos(\Delta_{s_{t+1}})$ is the price of engagement.

We now need to discuss the process of determining the beliefs about $\cos(\Delta_{s_{t+1}})$ and therefore the price of engagement p^{ε}. As the advertising market cannot distinguish the cartel engagement from natural engagement, the market price of engagement is

$$p^{\varepsilon} = v\mathbb{E}[\cos(\Delta_{s_{t+1}})] = (1-\varepsilon)p^{\text{natural}} + \varepsilon p^{\text{cartel}}, \tag{11}$$

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

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

Now, let us focus on cartel behavior. As each cartel is small, it takes p^{ε} as given. The expected value of joining the cartel is therefore

$$\begin{aligned} u^{\text{cartel+ad}}(R_{s_t}) &= u^{\text{cartel}}(R_{s_t}) + \mathbb{E}\left[\mathbf{1}_{\Delta_{s_{t+1}} \leq \Lambda}(1-\gamma)R_{s_{t+1}}p^{\varepsilon}\right] \\ &= u^{\text{cartel}}(R_{s_t}) + \frac{\Lambda}{180^\circ}(1-\gamma)\mathbb{E}[R_{s_{t+1}}]p^{\varepsilon}. \end{aligned} \tag{12}$$

As in the previous case, monetary incentives from the advertising market make the cartel more appealing. But there is a crucial difference—the payoff coming from the advertising market is strictly increasing in the engagement requirement Λ. This is because the market only rewards engagement quantity, not adjusting it to the quality.

For clarity, let us focus on the case when the advertising market incentives are large, i.e., the marginal value of engagement, v is big enough. As $p^{\varepsilon} \geq (1-\varepsilon)0.9v$, this also means that p^{ε} is large. As the first term in equation (12) is bounded, with sufficiently large v, the second part of the expression dominates. This means that for any $\Lambda > 0$ the expression is positive for all players, so all players join the cartel and $\mathbb{E}[R_{s_t}] = 2$. Moreover, the payoff for each cartel member is strictly increasing in Λ. These observations are formalized as the following two results.

Proposition 5. *If advertising market pays for quantity of engagement, for all $\Lambda > 0$ there exists $\overline{v} > 0$ such that for all $v \geq \overline{v}$, all players join the cartel.*

Corollary 3. *If advertising market pays for quantity of engagement, there exists $\overline{v} > 0$ such that for all $v \geq \overline{v}$, payoffs of all players are strictly increasing in Λ. Therefore both $V^{cartel}(\lambda)$ and $W(\lambda)$ are strictly increasing in λ for all $\lambda > 0$ and maximized when $\Lambda = 180°$.*

We can conclude from these results that if the advertising market pays for quantity of engagement and the incentives from the advertising market are large, we would expect each cartel to set $\Lambda = 180°$, i.e., require engagement regardless of the topic match. We call such cartels non-specific cartels. As in this case $\mathbb{E}[\cos(\Delta_{s_{t+1}})] = 0$, we would expect that $p^{\text{cartel}} = 0$ and therefore

$$p^{\varepsilon} = (1-\varepsilon)p^{\text{natural}}.$$

If the share of engagement coming from the cartels is small ($\varepsilon \to 0$), the existence of cartels only impacts only a small fraction of advertisers who are paying for the value they do not get. But if the fraction of cartels becomes significant, it also affects all players who get a lower price for engagement than they would otherwise. This outcome is a coordination failure. All participants would prefer the engagement requirement to be lower. However, this would require collusion between many cartel organizers, which would presumably be easily detected and punished by regulators.

4 Data

Our dataset combines data from two sources. First, the cartel data includes full history of interactions within cartels. Second, the outcome data includes posts and engagement measures from Instagram.

4.1 Cartel Data

We collected data from Telegram of seven Instagram influencer cartels. One was topic-specific (fitness and health), and the remaining six did not limit its members to any topics. These cartels include 180,280 Instagram posts altogether. We mapped these posts to 18,452 Instagram users. In the following empirical analysis, we call these users cartel members and influencers.

By construction of the dataset, we observe which of the cartel members' Instagram posts were submitted to the cartel asking for additional engagement and which were not. The cartel rules require that before adding a post to the cartel one has to comment and like previous five posts by other members. Therefore, according to the cartel rules, we also observe which engagement originated from the cartel.

4.2 Outcome Data

For a sample of Instagram users in the cartels, we collected information about their Instagram public posts. For the posts, we observe hashtags, tagged users, and other tags, and also the number of comments and likes.

Table 1 presents summary statistics of the users in the cartels. The median user has about 10,000 followers and about 300 posts. On their latest regular (not posted to the cartel) post, it has about 200 likes and about 30 comments. While on their latest post posted to the cartel, it has more than twice as many likes and more than three times as many comments. What share of the cartel members generate sponsored posts and earn from advertisement? While we cannot provide a complete answer to the question, we can analyze tags that they use. In the U.S., influencers are required to disclose that sponsored posts. They can do it using tags, such as #ad, #advertisement, or #sponsored, or simply say it in the text of the post. In our sample, 22% of cartel members use such tags. The percentage is higher for those with more followers (51% with at least 100,000 followers). We don't know whether the remaining cartel members don't disclose sponsored posts, disclose it in another way instead of using these tags, or don't advertise.

Our goal is to compare engagement originating from the cartel to natural engagement, that is, engagement that does not come from the cartel members. To do that, for a sample of posts in cartels, we collected information about their public comments. Specifically, we observe who commented on the post. Then we also collected information about a random sample of Instagram users who comment but are not in the cartel. Specifically, for each cartel member, we find the earliest Instagram post that was added to Telegram that has comments by the other cartel members that were supposed to comment according to the cartel rules. For that post, we randomly pick commenting Instagram users that were not

Table 1: Summary statistics of Instagram users in cartels

	By the number of followers				Total	
	<10K	10K-25K	25K-100K	100K+		
	Mean	Mean	Mean	Mean	Mean	Median
# followers	3700	15515	49111	299019	37270	9952
# posts	323	640	936	1770	621	303
# posts in cartels	5	9	16	26	10	3
# likes on cartel post	445	928	1739	6516	1266	520
# likes on non-cartel post	230	500	1017	4015	729	220
# comments on cartel post	35	59	80	167	59	33
# comments on non-cartel post	16	29	64	113	35	10
User has indicated ads (0/1)	0.082	0.270	0.436	0.508	0.224	0.000
Number of users	9251	4391	3413	1397	18452	18452

Notes: Unit of observation is a user. *# likes (or comments) on (non-)cartel post* measures the number of likes (or comments) of the user's latest post in cartel (or not in cartel). *User has indicated ads(0/1)* is an indicator variable that equals one if the user has ever used a hashtag indicating that the post is an advertisement or sponsored (#ad, #advertisement, #sponsored).

members of any of the cartels. 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. For these users, we also collected information about their Instagram public posts.

4.3 Measuring Engagement Quality

We measure engagement quality by the closeness of the match between the topics of the user and the one who comments his posts. Our goal is to evaluate whether the engagement comes from a user with similar interests based on the similarity of the users' content. To do that we first assign a distribution of topics to each user based on his posts, specifically, the tags used in the posts. In the second step, we calculate a pairwise cosine similarity score for each influencer and the engaging Instagrammer pair. The cosine similarity score gives us a summary measure of the similarity of their content.

4.3.1 Latent Dirichlet Allocation and Cosine Similarity

Figure 3 illustrates how the quality measures are calculated. The first step is to use the Latent Dirichlet Allocation model (LDA). In particular, for each Instagram user in our sample, we collect all tags and hashtags the user has used. In particular, we cluster these tags into 12 topics, and the LDA model maps topics to probability distributions over tags and users to probability distributions over topics. This allows us to map each user to a point on n-dimensional simplex (where n is three on the figure and 12 in our actual analysis) or equivalently probability distribution over topics. For example, on the figure,

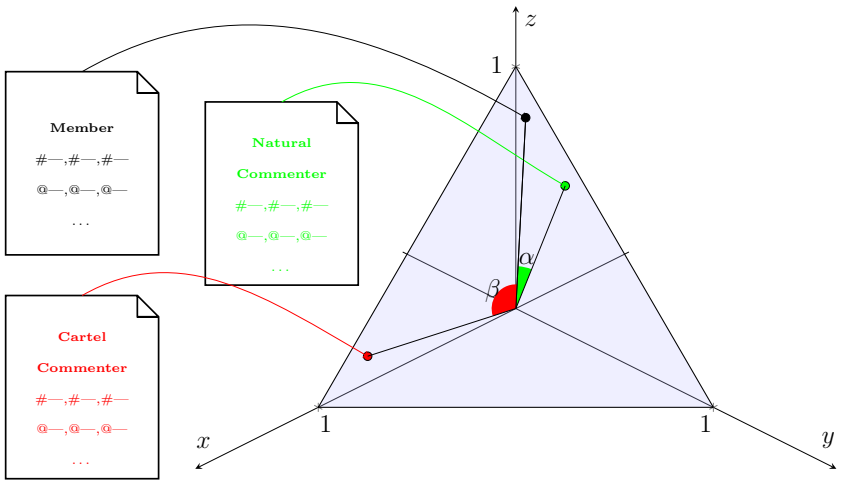


Figure 3: Quality measures of engagement. We first use the Latent Dirichlet Allocation model on tags and hashtags of each Instagram user to map the user to a probability distribution over topics (for simplicity, only three: x, y, z on the figure). Then we compute the Cosine Similarity of a cartel member and a set of natural and cartel commenters. $\cos(\alpha)$ is a quality measure for natural engagement and $\cos(\beta)$ quality measure for cartel engagement.

the cartel member happens to write mostly about topics z and slightly more on topic y than topic x. On the other hand, the representative cartel commentator often writes about x and rarely about y. Using this step alone allows us to make some comparisons. We can look at the topics users write on most often. On the figure, both the cartel member and the natural commenter write most frequently on topic z, whereas the cartel commenter writes mostly about x. Therefore we can conclude that the first two users are more similar to each other than the cartel commenter is to them.

To further formalize this idea and make the analysis single-dimensional, we calculate cosine similarity. In particular, we treat the points on the simplex as vectors from the origin and then compute the cosines of the angles between these vectors. On figure 3, $\cos(\alpha)$ captures the similarity of topics between the cartel member and the natural commenter. As the angle α is relatively small, $\cos(\alpha)$ is quite close to 1, which is interpreted as high similarity and therefore high (match) quality. On the other hand, the angle β between the topics that the member writes and the cartel commenter writes is quite close to 90°, which means $\cos(\beta)$ is close to 0. Therefore we would conclude that these users are not similar and the match quality is low.

4.3.2 Pre-Processing Tags

Before applying the LDA model, we pre-process the data as is standard in the literature. The goal is to reduce the set of tags, in order to improve learning from the underlying content.

In the first step, we shorten the hashtags. A common strategy in the literature is stemming, which is shrinking the words to their root form. However, in our case, hashtags typically combine several words, and therefore the standard stemming algorithms are not appropriate. Instead, we recursively shorten the hashtags deleting one-by-one characters from the end, until there are at least 100 users that have used the hashtag in our sample.

Then we exclude the tags that less than 100 users use. We also exclude the users who don't have enough tags because there wouldn't be enough information to determine their topics' distribution credibly. Specifically, we require that each user has at least 10 unique tags.

Finally, we use term frequency inverse document frequency (tf-idf) to limit the set of tags further. The tf-idf measures informativeness and punishes tags that are used either too seldom or too often. We rank the remaining tags using tf-idf and keep for each user only the 100 highest ranked tags.

4.3.3 Estimated LDA Topics

LDA algorithm groups the tags into 12 topics, estimating a probability distribution over the tags for each topic, and for each user a probability distribution over the topics. Based on the most representative tags in each topic, we assign labels to the topics. Figure 4 presents the list of 12 topics in the first column and for each topic, the four most representative tags. There are a few rather related topics, such as fashion, men's fashion, and fashion brands. At the same time, other LDA topics combine into one sub-topics such as Instagram and photos, and make-up and motherhood.

5 Empirical Results

5.1 Quality of engagement

Distribution of LDA topics. To analyze the quality of engagement, we start by looking at the distribution of topics that characterizes the cartel-originating versus natural (non-cartel) engagement. To do that, we assign each user a single main topic—the one with the highest estimated probability according to the LDA model. We then look at the distribution of main topics of influencers' commenters. We do that separately for commenters coming from the cartel and those who are not part of the cartels.

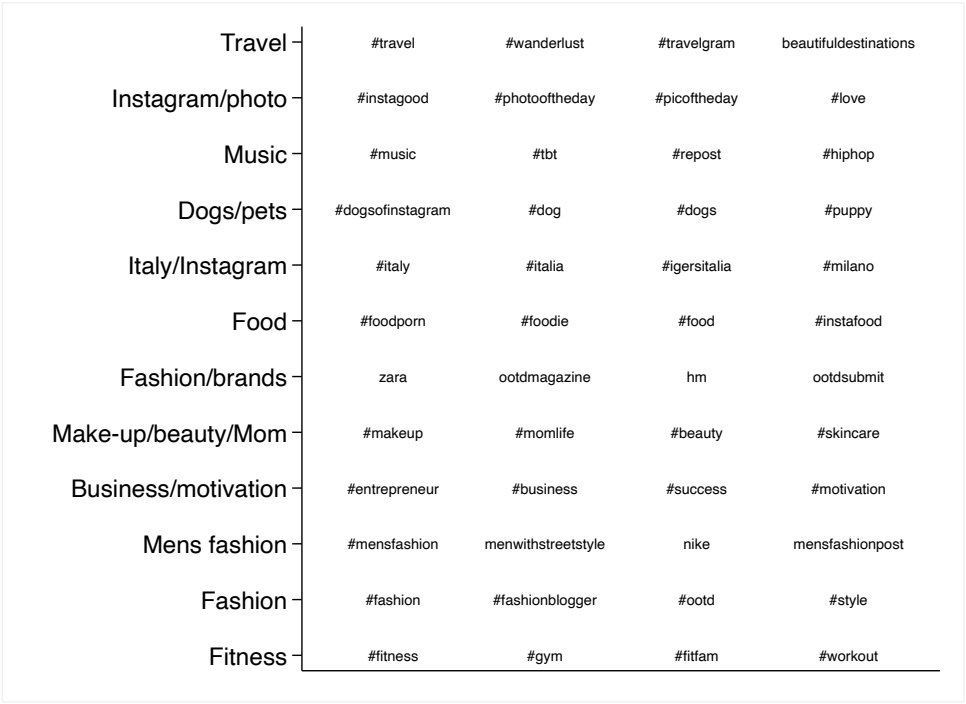
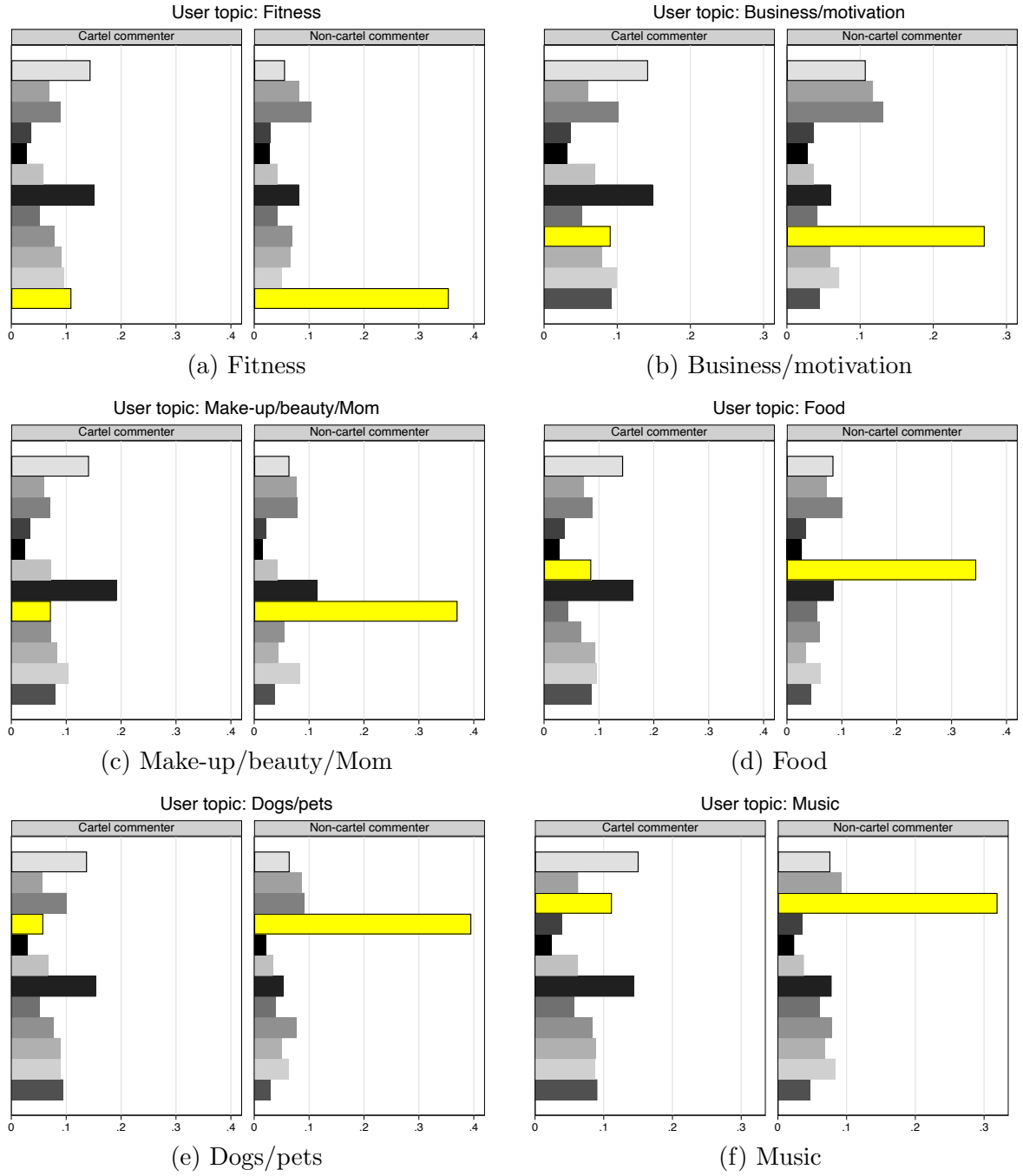


Figure 4: LDA topics and 4 representative tags for each topic

Notes: LDA algorithm groups the tags (hashtags, tagged users, and other tags) into 12 topics listed in the first column. In each row in columns 2-5 are the 4 most representative tags corresponding to the topics.

Consider influencers whose main topic is fitness. The left graph on Figure 5a presents the topic distribution of engagement originating from cartels for the influencers who post mainly about fitness. The yellow bar measures the fraction of commenters whose main topic is also fitness. The other bars colored in various shades of grey represent the remaining 11 topics (as presented on Figure 4). We see that the cartel members who engage with fitness influencers are themselves writing about other topics more or less as much as about fitness. The right graph on Figure 5a presents the topic distribution of natural engagement (non-cartel commenters). Again, the yellow bar measures the fraction of commenters whose main topic is fitness. For natural engagement, we see that a large share of those engaging with fitness influencers, write mostly about fitness themselves.

Figure 5 presents analogous topic distributions of cartel-originating versus natural engagement for other influencers, those whose main topic is business/motivation, make-up/beauty/Mom, food, dogs/pets, and music. The graphs show that a large share of the natural engagement comes from the users who themselves post mostly about the same topic. But topics of users who comment because of the cartel are more or less uniformly distributed, being largely unrelated to the main topic of the influencer. Figure B.1 in Online Appendix B presents analogous figures for other topics.



(a) Fitness (b) Business/motivation

(c) Make-up/beauty/Mom (d) Food

(e) Dogs/pets (f) Music

Figure 5: LDA topic match: cartel vs. natural engagement in non-specific cartels

Notes: Each of the 12 figures presents the distribution of commenters' topics of influencers in non-specific cartels. Each user is characterized by a single main topic—the one with the highest estimated probability according to the LDA model. The sample on Figure 5a is restricted to commenters who comment on the influencer whose main topic is fitness. The figure presents the distributions of main topics separately for commenters coming from the cartel (left figure) and those who are not part of the cartels (right). The x-axes measures the fraction of commenters with a given topic. The yellow bar measures the fraction of commenters whose main topic is also fitness. The grey bars measure the fraction of commenters with the remaining 11 topics as presented on Figure 4. The remaining graphs present the distribution of commenters commenting on influencers whose main topics is business/motivation, make-up/beauty/Mom, food, dogs/pets, or music.

For the fitness and health topic cartel the picture looks different. In the topic-specific cartel (figure Figure 6), we see that the distribution of engagement originating from the cartel is an even closer match to the influencer than the natural engagement.

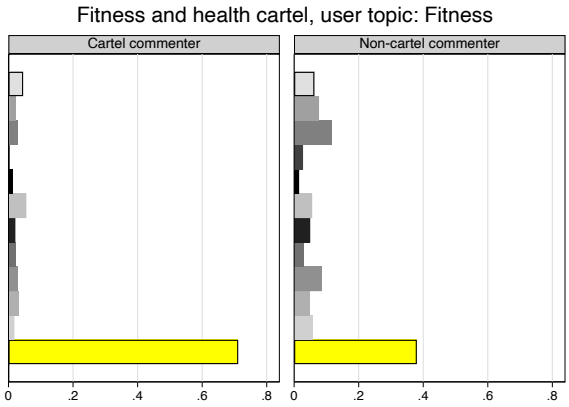


Figure 6: LDA topic match: cartel-originating versus natural engagement for the fitness and health topic cartel

Notes: Each figure presents the distribution of commenters' topics of influencers in the fitness and health topic cartel. Each user is characterized by a single main topic—the one with the highest estimated probability according to the LDA model. The sample is restricted to commenters who comment on the influencer whose main topic is fitness. The figure presents the distributions of main topics separately for commenters coming from the cartel (left figure) and those who are not part of the cartels (right). The x-axes measures the fraction of commenters with a given topic. The yellow bar measures the fraction of commenters whose main topic is also fitness. The grey bars measure the fraction of commenters with the remaining 11 topics as presented on Figure 4.

Cosine similarity of users. To answer the question whether engagement from cartels is of lower quality, we estimate a panel data fixed effects regression where the outcome variable is the cosine similarity of an influencer and his commenter. In the analysis, 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 and we estimate a panel data regression with influencer fixed effects. Our goal is to compare whether cartel commenters are less similar to the influencer than the natural commenters, who are the base group in the regressions. In columns 1–2 in Table 2, the coefficient of interest is on the variable indicating that the commenter is from the cartel. The sample in column 1 consists of influencers who are themselves from non-specific cartels, while in column 2, the influencers are from the fitness and health topic cartel.

The estimates in column 1 show that in non-specific cartels, influencer's similarity with commenting cartel members is significantly lower compared to the non-cartel commenters

(base category). In contrast, estimates in column 2 show that in the fitness and health topic cartel, similarity with commenting cartel members is even slightly higher compared to the non-cartel commenters. Overall these results confirm what we already saw from the LDA distributions on Figures 5 and 6, that the topic cartel delivers engagement with a better topic match.

Table 2: 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)
	Non-specific cartels	Topic cartel	Counterfactuals in non-specific cartels Similarity with random users			
Cartel commenter	-0.203***	0.061***				
	(0.003)	(0.011)				
Random non-cartel user			-0.220***			
			(0.005)			
Random cartel member				-0.219***		
				(0.005)		
Random cartel member from 6 topics					-0.140***	
					(0.005)	
Random cartel member from 2 topics						0.090***
						(0.004)
Base group average	0.426	0.399	0.426	0.426	0.426	0.426
Influencers	10764	1990	10764	10764	10764	10764
Observations	62514	6451	29439	30553	31193	48658

Notes: Each column presents estimates from a separate panel data fixed effects regression. Unit of observation is an influencer-user pair. Outcome variable is the cosine similarity of an influencer and his commenter or a random user. Each regression includes influencer fixed effects. In all the regressions, the base category with whom the influencer's similarity is calculated, is the non-cartel commenter; and *Base group average* presents their average cosine similarity. *Cartel commenter* is an indicator variable whether the commenter with whom the influencer's cosine similarity is calculated, is in the cartel. *Random non-cartel user* and *Random cartel member* indicate that the influencer's similarity is calculated with a random user not in the cartel or in the cartel, respectively. To calculate the similarity with *Random cartel member from 6 topics*, we split influencers by their main topic into two groups (six topics each) and then for each influencer pick random users with the main topic from the same set. Analogously, for *Random cartel member from 2 topics*, we divide the influencers into six groups (two topics each) and pick random users with the main topic from the same set. Standard erros in parenthesis are clustered on influencers.

How much worse is the cartel-originating engagement compared to the natural engagement in the non-specific cartels? To answer the question we run counterfactual analysis measuring the similarity of influencers and random users. The random users give us a benchmark estimate for the lowest quality engagement. Columns 3–6 re-estimate the regression in column 1, but instead of using influencer's similarity with cartel commenters, they use similarity with random users. In all regressions the base category is natural en-

gagement (non-cartel commenters). Columns 3–4 show that influencer's similarity with random users (not from cartel and cartel, respectively) is significantly lower compared to natural engagement. The magnitude of the difference is similar to that in column 1. The estimates imply that the engagement originating from non-specific cartels is about as bad as engagement from random users. Columns 5–6 provide counterfactual exercises to complement estimates in column 2. Specifically, in column 5 we split influencers by their main topic into two groups (six topics each) and then for each influencer pick random users with the main topic from the same set. Column 5 shows that hypothetical cartels restricting topics to six out of twelve would improve engagement quality but would still be significantly lower than natural engagement. In column 6, we divide the influencers into six groups (two topics each). Our estimates in column 6 replicate the results in column 2, showing that hypothetical cartels limited to only two topics would not be worse than natural engagement.

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 non-specific 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 non-specific 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 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 non-specific 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 cartels exist and work as intended, bringing additional engagement to cartel members. However, the engagement from non-specific cartels is of significantly lower quality than the natural engagement, whereas the engagement from topic-specific cartels can be as good as 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.

Our analysis focuses on short-term effects. There is evidence by Weerasinghe et al. (2020) that cartel members' following grows faster than the average influencer's following, conditional on having the same number of followers initially. This suggests that cartels may also have long-term effects, but this observation may also be explained by unobservable differences (perhaps influencers joining a cartel are also otherwise more active in their growth efforts). More research is needed to understand the implications of the dynamic effects.

References

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

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 Marketing? Here's What To Look For," *Forbes*.

Chen, Y., R. Farzan, R. E. Kraut, I. YeckehZaare, and A. F. Zhang (2019): "Motivating Contributions to Public Information Goods : A Personalized Field Experiment on Wikipedia," *manuscript*.

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.

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.

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

Franck, G. (1999): "Scientific Communication–A Vanity Fair?" *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.

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 Operate?*, Now Publishers Inc.

Hinnosaar, M., T. Hinnosaar, M. Kummer, and O. Slivko (2021): "Externalities in Knowledge Production: Evidence from a Randomized Field Experiment," *Experimental Economics*.

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.

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

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.

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.

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.

A Online Appendix: Proofs

A.1 Proof of Proposition 1

Proof In non-cooperative equilibrium, player t chooses $a_t = 1$ if and only if the benefits outweigh the costs, $V_t \geq C_t$. As costs are always non-positive, it requires that $\cos(\Delta_t) \geq 0$ or equivalently $\Delta_t \leq \frac{\pi}{2}$ and, moreover, $\gamma \cos(\Delta_t) \geq \sin(\Delta_t)$, which is equivalent to $\Delta_t \leq \tan^{-1}(\gamma)$.

On the other hand, engagement is socially optimal whenever total benefits outweigh the costs, $U_{t-1} + V_t \geq C_t$, which is equivalent to $\Delta_t \leq \tan^{-1}(1) = 45$. Clearly, as $\gamma < 1$, there are strictly more engagements that are socially optimal than taking place in equilibrium. Moreover, for any additional engagement in this form, i.e., Δ_t such that $\Delta_t \in (\tan^{-1}(\gamma), 45°]$, we have a property that $\cos(\Delta_t) < \cos(\Delta'_t)$ for any $\Delta'_t \leq \tan^{-1}(\gamma)$. That is, the added engagements are of strictly lower quality. □

A.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 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 $\lambda > 1$, then the cartel payoff is strictly lower than equation (4) and the expression is strictly negative, so nobody would join the cartel in this region. □

A.3 Proof of Corollary 1

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

$$W(\lambda) = V^{\text{cartel}}(\lambda) = \frac{4\lambda(1-\lambda)}{\lambda^2+1}\mathbb{E}[R] = \frac{8\lambda(1-\lambda)}{\lambda^2+1}. \tag{13}$$

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

$$V^{\text{cartel}}(\lambda) = \frac{4\lambda(1-\lambda)}{\lambda^2+1}\frac{2-\gamma-\lambda}{1-\gamma}, \tag{14}$$

$$W(\lambda) = Pr(R_{s_t} \leq \overline{R})V^{\text{cartel}}(\lambda) = \frac{16\lambda(1-\lambda)^2}{(\lambda^2+1)(2-\gamma-\lambda)}. \tag{15}$$

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 (7). 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 $W(\lambda)$. □

A.4 Proof of Proposition 3

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 (13), 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 (13), 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. □

A.5 Proof of Proposition 4

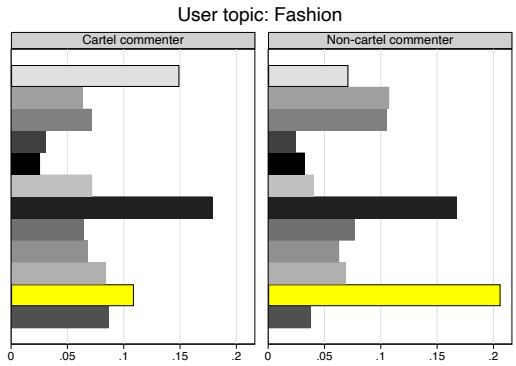
Proof Cases when $\lambda < 1$ are analogous to proposition 2 and we already argued that at $\lambda = 1$ some players join the cartel. Consider the case when $\lambda > 1$ and suppose that some players join the cartel. Then there must again exist a marginal reach $\overline{R}$ such that only players with $R_{s_t} \leq \overline{R}$ join the cartel. Therefore the expected reach of the follower is $\mathbb{E}[R_{s_t}|R_{s_t} \leq \overline{R}] = \frac{2}{1+\overline{R}^{-1}}$. The expected payoff from the cartel to the marginal type is now

$$u^{\text{cartel+ad}}(\overline{R}) = u^{\text{cartel}}(\overline{R}) + (1-\gamma)\frac{2}{1+\overline{R}^{-1}}v2\underbrace{\int_{0^\circ}^{90^\circ}\cos(\Delta_{s_{t+1}})d\Delta_{s_{t+1}}}_{=1},$$

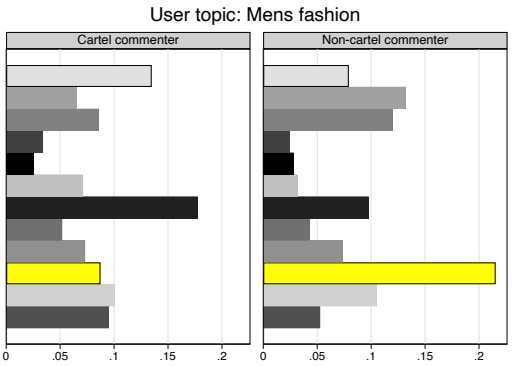
because the advertising revenue is paid only for engagement with positive $\cos(\Delta_{s_{t+1}})$. Notice that $u^{\text{cartel}}(\overline{R}) < 0$, but the second term is positive. For each v, this equation defines a marginal value $\overline{R}$. Therefore $\overline{R} > 1$.

Finally, note that in the limit where $\lambda \to \infty$ or equivalently, $\Lambda = 90^\circ$, the first part of the payoff $u^{\text{cartel}}(R_{s_t})$ is strictly negative, but bounded. Therefore, if v is large enough, the second part of the payoff, coming from the advertising revenue, is sufficient compensation so that some players still join the cartel. Therefore we know that $\overline{\lambda} > 1$, but not necessarily that it is finite. □

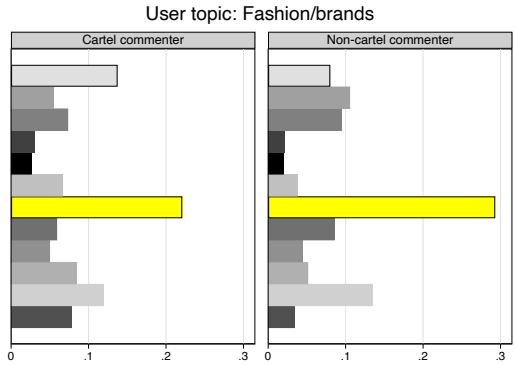
B Online Appendix: Additional Figures



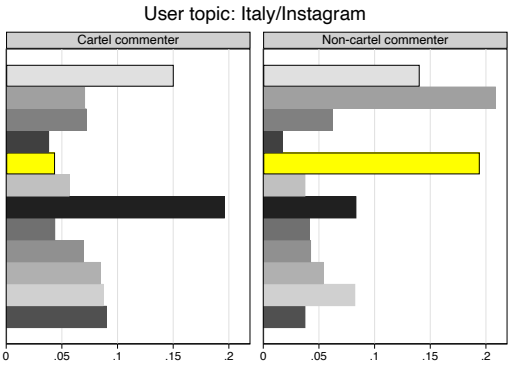
(a) Fashion



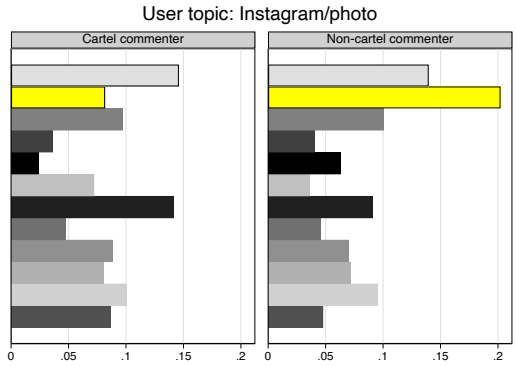
(b) Men's fashion



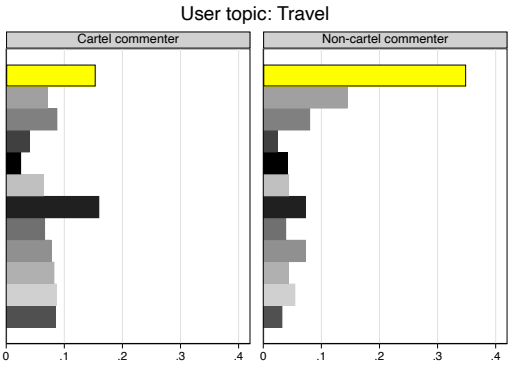
(c) Fashion/brands



(d) Italy/Instagram



(e) Instagram/photo



(f) Travel

Figure B.1: LDA topic match: cartel-originating versus natural engagement for non-specific cartels