IRSEM
INSTITUT DE RECHERCHE STRATÉGIQUE
DE L'ÉCOLE MILITAIRE

October 2022

THE BUSINESS MODEL OF CONTENT-SHARING PLATFORMS AND THE SUPPLY OF CONTENT MODERATION

IMPLICATIONS FOR COMBATING INFORMATION MANIPULATIONS

Olivier Chatain

Professor, Strategy and Business Policy Department, HEC Paris

MINISTÈRE
DES ARMÉES
Liberté
Égalité
Fraternité

REPORT – NO. 99

IRSEM
INSTITUT DE RECHERCHE STRATÉGIQUE DE L'ÉCOLE MILITAIRE

THE BUSINESS MODEL OF CONTENT-SHARING PLATFORMS AND THE SUPPLY OF CONTENT MODERATION

IMPLICATIONS FOR COMBATING INFORMATION MANIPULATIONS

Olivier Chatain

Professor, Strategy and Business Policy Department, HEC Paris

To quote this publication
Olivier Chatain, *The Business Model of Content-Sharing Platforms and the Supply of Content Moderation: Implications for Combating Information Manipulations*, Report 99, IRSEM, October 2022.

Dépôt légal
ISSN : 2268-3194
ISBN : 978-2-11-167762-3

RECENTLY PUBLISHED

98. *Comprendre le Moyen-Orient par la donnée – Technologies numériques et acquisition de la connaissance dans la région Afrique du Nord / Moyen-Orient*
COL Olivier PASSOT

97. *La Russie au Mali : une présence bicéphale*
Maxime AUDINET et Emmanuel DREYFUS

96. *La singularité du métier militaire : persistances et nouveautés – Pourquoi défendre un modèle de singularité ?*
Clément SORBETS

95. *L'armée, les Français et la crise sanitaire : une enquête inédite*
Anne MUXEL, Florian OPILLARD et Angélique PALLE

94. *L'extrémisme islamiste au nord du Mozambique : terrorisme et insécurité à Cabo Delgado*
Régio CONRADO

93. *La latence nucléaire du Japon : un levier diplomatique à double usage ?*
Timothée ALBESSARD

92. *Le régime milicien iranien en Irak – Les milices chiites pro-iraniennes à la conquête de l'État*
Arthur QUESNAY

91. *Facing a pandemic: African armies and the fight against COVID-19*
Anne-Laure MAHÉ and Nina WILÉN (eds)

90. *L'intervention française au Sahel et l'évolution de la doctrine de contre-insurrection*
Michael SHURKIN

89. *Observatoire de la génération Z*
Anne MUXEL

TEAM

Director
Jean-Baptiste JEANGÈNE VILMER

Deputy Director
Marjorie VANBAELINGHEM

Scientific Director
Jean-Vincent HOLEINDRE

General Secretary
Caroline VERSTAPPEN

Editor
Chantal DUKERS

Find IRSEM on social medias:

@ https//www.irsem.fr

@IRSEM1

DISCLAIMER: One of IRSEM's missions is to contribute to public debate on issues relating to defence and security. The views expressed in IRSEM's publications are the authors' alone and are in no way representative of an official Ministry of the Armed Forces stance.

ABOUT IRSEM

Founded in 2009, the Institute for Strategic Research (IRSEM) is a research institute attached to the Ministry of the Armed Forces' General Directorate for International Relations and Strategy (DGRIS). The institute employs a staff of forty-five civilian and military personnel, and its primary aim is to further French research on defense and security stakes.
The research team is divided into six departments:
• The 'Transatlantic Studies' department analyses strategic and geopolitical developments in North America, Europe, Russia and the Eurasian areas which include Eastern Europe (Moldova, Ukraine, Belarus), the South Caucasus (Armenia, Georgia, Azerbaijan) and the five Central Asian countries. The department's research team analyzes competition for power in that region, the evolving role of NATO, maritime safety, and strategies of influence.
• The 'Africa – Asia – Middle East' department analyses strategic and geopolitical developments in those regions through the following themes: political authoritarianism and economic liberalization in emerging countries; the role of the army and the security apparatus in the way states and societies function; strategic and regional security challenges; ideologies, nationalisms and the redefining of regional interstate balances.
• The 'Weaponry and Defense Economics' department's team focuses on economic issues related to defense. More broadly, it includes strategic issues resulting from technological developments, problems of access to natural resources and those related the environment. The department's research is based on an interdisciplinary approach, both qualitative and quantitative, which mobilizes various scientific fields: defense economics, history of technologies, and geography.
• The 'Defense and Society' department is at the crossroad of issues specific to military circles and of the social evolutions they face. The following aspects are put forward in particular: the link between civilian society and the armed forces, sociology of military personnel, integration of women in armed conflicts, relations between political power and the Army as an institution, renewal in the forms of commitment, socialization and integration of the youth, rise of

radicalisms. Beyond its research activities the Defense and Society department also promotes defense issues within civilian society, towards all its constituents, including those in the academia.

• The 'Strategies, Norms and Doctrines' department is dedicated to the study of contemporary armed conflicts, particularly in their political, military, legal and philosophical dimensions. The main threads of research developed in its publications and the events it arranges relate to international law, in particular from a technological standpoint (cyber, artificial intelligence, robotics), deterrence doctrines, arms control, including nuclear disarmament and the fight against such proliferation. The transformations of international relations and in their stakes in terms of power and security, as well as the philosophy of war and peace are also part of its field of study.

• The 'Intelligence, Anticipation and Hybrid Threats' department conducts research on the "knowledge and anticipation" strategic function put forward by the Defense White Paper since 2008. This programme therefore aims at contributing to a more subtle understanding of intelligence in its broadest sense (i.e. as information, process, activity and organization); secondly, it aims at contributing to the consolidation of analytical approaches, particularly in the field of anticipation; finally, it works on the different dimensions of so-called "hybrid" warfare, particularly on information manipulation. The field also contributes to strengthening the hybrid nature of the IRSEM by publishing notes which are halfway between academic research and open source intelligence analysis.

BIOGRAPHY

Olivier Chatain is a Professor at HEC Paris in the Strategy and Business Policy Department. His research interests include business strategy, the role of firms in the public sphere, and how states and non-state actors exploit privately managed infrastructures to further their interests. He holds a PhD from INSEAD and served on the faculty of the Wharton School at the University of Pennsylvania before joining HEC Paris. He was an invited researcher at IRSEM during the academic year 2021-2022.

Contact: chatain@hec.fr

CONTENTS

ABSTRACT

New forms of conflictuality below the threshold of violence often unfolds in spaces that are created and administered by private organizations, yet the roles played by these organizations in shaping the context in which conflicts happen, and their motivations, is rarely explored in security studies. This note explores the role played by content-sharing digital platforms in shaping the environment conducive to information manipulations. The note clarifies the economic incentives and constraints under which platforms operate. These incentives and constraints shape the essential design choices made by platforms, especially regarding the potency of network effects. This makes content-sharing platforms attractive targets for information manipulators who adapt their tactics to this new domain, but also affects the platforms' ability and incentives to conduct effective content moderation to counter manipulations. Using this conceptual toolbox, the note makes a preliminary assessment of the potential impact of the forthcoming Digital Service Act prepared by the European Union on platforms' efforts to moderate content, and the possible responses of malicious actors.

Acknowledgements: I thank Maxime Audinet for his discussion of an earlier version of the paper as well as seminar participants at IRSEM and at the 2022 IRSEM-RAS conference for their feedback.
I also thank the various interviewees and platforms insiders who have anonymously shared their insights. The content of this report exclusively reflects my personal views.

INTRODUCTION

This report studies how the business logic of digital content-sharing platforms – firms intermediating between users to allow users to communicate and share content by digital means – plays a role in shaping the terrain on which information manipulations unfold. Relying on the business and economic literature on platforms, it seeks to clarify the economic incentives and constraints under which they operate. These incentives and constraints drive the essential design choices these platforms make, especially regarding network effects, which make them attractive targets for information manipulations, as well as their ability and incentives to supply content moderation, which in turn limits the success of these manipulations.

Putting to the fore, and breaking it down in details, the business logic that private firms follow, is rarely done in the field of security studies. However, conflicts between states below the threshold of armed conflict (e.g., cyber conflicts, economic coercion, information manipulation) are overwhelmingly involving infrastructures, supply chains, production capabilities and communication networks that are managed by private, for-profit, entities. These entities have shaped according to their own logic what has become arenas of conflict and continue to do so even as those conflicts unfold. Understanding their logic, beyond remarking that they are profit-seeking, requires grappling with concepts developed in the adjacent disciplines of management and economics.

Content-sharing platforms are first and foremost businesses and, as such, face specific incentives and exploit specific economic and managerial mechanisms to ensure their growth, profitability, and survival. Conceptually, this note is thus an invitation to look at platforms not only as legal entities, vessels for free speech, practitioners of lobbying, carriers of public diplomacy. But also to open the business economics toolbox and analyze them in terms of concepts such as network effects, economies of scale, competitive strategy, barriers to entry, organizational capabilities, and managerial incentives. All these concepts imbue

the incentives and constraints that drive the decision-making of platforms. Using them as an analytical lens can ultimately contribute to explaining why information manipulations on platforms take the shape that they do.

The report does not address the broader issue of how platforms influence the information ecosystem of democracies, which would involves understanding the actions of many more actors.[1] Instead, the scope of the report is restrained to an examination of how the business model of digital content-sharing platforms influences their content moderation efforts, and of how this bears on instigators of information manipulations, i.e., entities that organize covert efforts to cause political harm thanks to the dissemination of manipulated information.

This report will first outline the essential characteristics of the business model of content-sharing platforms by situating at the root of their efforts the fostering and harnessing of network effects with the goal of making user-generated content attractive and easy to find. A side effect of these efforts is to make content-sharing platforms particularly vulnerable to information manipulations, which seek to spread politically harmful content. Some form of content moderation is needed to combat disinformation, a prominent form of manipulation. In a second part, this report will detail the challenges faced and the systems put into place by platforms to moderate content in general and assess their effectiveness and scalability for dealing with disinformation and information manipulations. Finally, and relying on the same conceptual framework, the report will assess the potential impact of the forthcoming Digital Service Act prepared by the European Union on competition among platforms and their efforts to moderate content. This analysis can help build scenarios of how the content-sharing platform landscape may evolve and thus how the environment of information manipulations may itself change.

1. In that vein, see Yochai Benkler, Rob Faris, and Hal Roberts, *Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics* (New York, NY: Oxford University Press, 2018).

I. WHAT DO PLATFORM-BASED BUSINESSES CARE ABOUT?

To understand the frame of mind of the top management of content-sharing platforms', we need to explain how they perceive the link between their business decisions and the survival and profitability of their platform in a competitive business environment. A detailed understanding of this is important for scholars interested in disinformation and information manipulations because this can help them evaluate the claims made by platforms, but also those made by them most strident critics. While the former can be prone to slanting facts in a self-serving manner, some of the latter may rely on oversimplifications that can undermine their argument.

Research in management and economics suggests that two key economic concepts that undergird the business model of digital content-sharing platforms are network effects,[1] and (almost) zero marginal costs.[2] The former explains the drive for keeping and growing the user base in a competitive environment, causing races to acquire as many users as possible, while the latter is a fundamental feature of the digital economy which enables unprecedented economies of scale and as a result advantages larger platforms. These two mechanisms can parsimoniously explain the key features of the business model of digital platform strategy.[3]

1. Michael L. Katz and Carl Shapiro, "Network Externalities, Competition, and Compatibility," *The American Economic Review* 75, no. 3 (1985): 424–40.

2. Carl Shapiro and Hal R. Varian, *Information Rules: A Strategic Guide to the Network Economy* (Boston, Mass: Harvard Business School Press, 1999).

3. Michael A. Cusumano, Annabelle Gawer, and David B. Yoffie, *The Business of Platforms: Strategy in the Age of Digital Competition, Innovation, and Power*, First edition (New York, NY: Harper Business, an imprint of HarperCollinsPublishers, 2019).

THE PRIMACY OF NETWORK EFFECTS AND THEIR MANAGEMENT

A platform exhibits network effects when the value of its service for a user is increasing in the number of users simultaneously using the platform. The simplest example of such effect arises in a communication network, such as a telephone network. The value of being connected to the network is lowest when there is only one other user, however it dramatically, and exponentially, increases as the number of other users increases. That the value of the connection increases as the network grows is an example of a *positive* network effect. Network effects can also be negative in cases of congestion – too many users can degrade the experience, as in the case of road networks.

It is often useful to categorize the users of a platform into different *sides*. On some platforms (e.g., WhatsApp), all users play a similar role in their provision of content and are in effect on the same side. On other platforms (e.g., YouTube), there is a sharper distinction between the users who consume content and the users who create content, which places them on different "sides" of the platform. If viewers leave comments on videos and enjoy other people's comment this would create *same-side* network effect, i.e., there is an effect of the number of participants in one side on the utility of the participants of the same side. Network effect can be *cross-side*, i.e., the more there are users on one side, the more users on another side value participating to the platform. For instance, the more creators post content on YouTube, the more the viewers value being connected to YouTube (cross-side network effects from creators to viewers). At the same time, the more viewers are connected to YouTube, the more creators value posting their creations on YouTube as the audience is larger (cross-side network effects from viewers to creators). These reciprocal cross-side network effects reinforce each other, in a positive feedback loop. While there are network effects between consumers and creators of content, there is also a third side that YouTube connects to its platforms: the advertisers who bring revenues, although the inconvenience of watching ads may degrade the experience of consumers. The most committed users can avoid ads by paying

a subscription. This example also illustrates the general principle that users on sides that are essential to network effects are usually paying little, or are even being paid to participate, while users with higher willingness to pay (committed viewers, advertisers) are asked to pay to participate to the network and effectively subsidize the participation of others.[4]

Network effects matter for business strategy and market structure because they create opportunities for *market tipping*, a situation that arises when a firm becomes dominant on a market in a self-reinforcing process. The positive feedback loops or "snowball effects" due to network effects can make a market converge to the dominance of one or a very limited number of platforms. Managers of platforms, and their financial backers, are explicitly playing for this sort of dynamics when they frontload investments and spend significant financial resources for user acquisitions upfront. In markets with network effects, there can be even more competition *for* the market as *in* the market. The nature of this competition is to lock in users and manage expectations to make who wins a self-fulfilling prophecy.[5] Interestingly, there are examples that the phase of user acquisition, which is entirely based on the exploitation of network effect, can take a very long time, sometimes years, while the firm involved have no clear idea yet of how to make money with the asset that their user base has become. Twitter is a case in point. As a company Twitter has

4. Mark Armstrong, "Competition in Two-Sided Markets," *The RAND Journal of Economics* 37, no. 3 (2006): 668–91; Bernard Caillaud and Bruno Jullien, "Chicken & Egg: Competition among Intermediation Service Providers," *The RAND Journal of Economics* 34, no. 2 (2003): 309–28, https://doi.org/10.2307/1593720; Jean-Charles Rochet and Jean Tirole, "Platform Competition in Two-Sided Markets," *Journal of the European Economic Association* 1, no. 4 (June 1, 2003): 990–1029, https://doi.org/10.1162/154247603322493212; Geoffrey G. Parker and Marshall W. Van Alstyne, "Two-Sided Network Effects: A Theory of Information Product Design," *Management Science* 51, no. 10 (2005): 1494–1504.

5. Kevin J. Boudreau, "Promoting Platform Takeoff and Self-Fulfilling Expectations: Field Experimental Evidence," *Management Science*, June 8, 2021, https://doi.org/10.1287/mnsc.2021.3999.

created massive network effects, occupies a unique position in the information ecosystem, but still struggles to be consistently profitable.

Another fundamental characteristic of digital platforms influences their behavior and strategy. In the digital economy, information is extremely cheap to reproduce and transmit and, as a result, adding another user to a network is very cheap. The marginal costs of operation – the additional costs related to producing one more unit of output – are very low, especially compared to what is prevalent in the non-digital world. Contrast for instance the cost of delivering a newspaper's worth of information via a high-speed network and the cost of printing and delivering the paper version. While marginal costs are very low, most of the costs and investment of content-sharing platforms are fixed, especially IT development. A piece of computer code must only be written once but can be reused endlessly. Of course, this is an oversimplification since maintaining a code base and hosting servers is costly. But the key point is that the balance of fixed vs. marginal costs is heavily biased towards fixed costs in a manner that sets these businesses apart from traditional firms. This makes *economies of scale*, the reduction of unit costs achieved thanks to higher volumes, a fundamental driver of profitability for these firms, and as a result, realizing these economies is central to what digital platforms do.

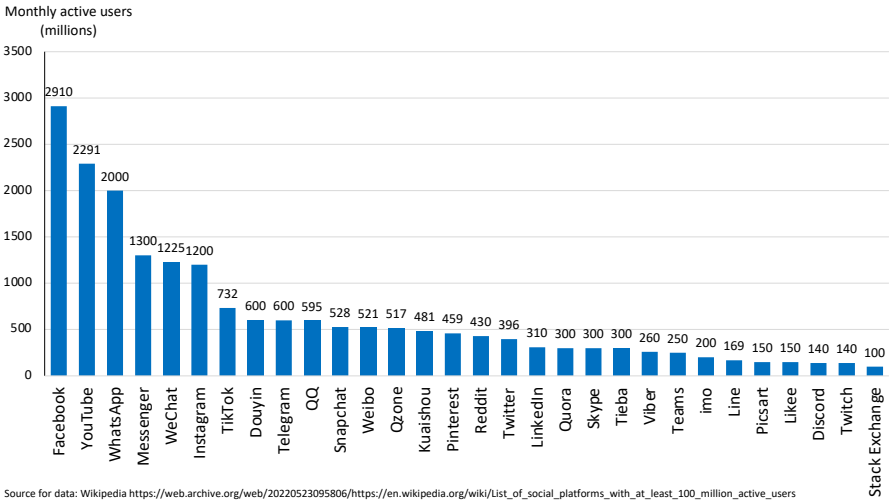
Putting together network effect, low marginal costs, and economies of scale creates an incentive for growth and user acquisition and retention that is only exacerbated by competition between platforms. Such competition creates the condition for winner-takes-all situation which result in a very unequal distribution of market shares and profitability. Indeed, the largest content-sharing firms such as those owned by Meta (Facebook, Instagram, WhatsApp) and Google (YouTube) are very profitable and have achieved unprecedented user bases, above a billion users. By contrast, the second tier of platforms, including Twitter (396 million monthly active users in 2021), are much less impressive in this respect. Sheer size is thus highly desirable to sustain a platform business because of network effects

and economies of scale. Absent this, a content-sharing platform needs to find a niche of users willing to pay even a small amount to be a member.

Figure 1 shows the distribution of monthly users in 2021 among large digital content-sharing platforms (more than 100 million users).

Figure 1

Worldwide monthly active users of large content-sharing platforms (2021)



Source for data: Wikipedia https://web.archive.org/web/20220523095806/https://en.wikipedia.org/wiki/List_of_social_platforms_with_at_least_100_million_active_users

PLATFORM DESIGN AND NETWORK EFFECTS

The plasticity of digital platforms, thanks to the use of code, offers many opportunities for making users enjoy more the content posted by others as well as making their own content more interesting to others. Software also determines how content is made available and to whom. These together contribute to increasing network effects and content-sharing platform design decisions typically seek to add features to improve the appeal of the platform in the face of the offering available at competing platforms.

Improving the supply of user-generated content: Designing for virality

What constitutes "design" for a digital platform should be understood in a very broad way. It comprises not only the elements of the user interface, but more fundamentally, the decisions taken by the platform to regulate and frame interactions between users, and their implementation as code, or elements of their terms of services. By virtue of their reliance on software and code, and the ability to change it at will, digital platforms have a unique ability to influence how their users contribute and benefit from their activities on platform.[6] This, combined with the computational and media capabilities of the digital devices in the hands of users, makes platform design decisions key to their ability to foster and maintain network effects. The details of these decisions have implications on discourse and influence in the public sphere.[7] It is worth emphasizing that even though "virality" may be seen pejoratively, it is also exactly what users are looking for in social interactions wherever they take place: interesting bits to hear and see and to repeat, whether around the office watercooler or on a digital interface. Understanding the motivations behind the design decisions made by platforms matters to improve our understanding of how disinformation may spread.

The first, and easily taken-for-granted, set of rules implemented in code, is how users can interact with each other and create new connections. Platform run the gamut from allowing anyone to participate to emphasizing exclusivity and requiring special invitations. Similarly, some platforms allow to interact with any users, while other require existing connections,

6. Joel R. Reidenberg, "Lex Informatica: The Formulation of Information Policy Rules through Technology," *Texas Law Review* 76 (1998 1997): 553; Lawrence Lessig, *Code and Other Laws of Cyberspace* (New York: Basic Books, 1999).

7. Olivier Chatain and Madhulika Kaul, "No Easy Way Out? Platform-Mediated Political Externalities and Platform Strategy," HEC Paris Working Paper, 2022.

sometimes mediated by common acquaintances. These decisions are calculated, in conjunctions with other features, to create a balance between making it easy for users to join, but also to foster a feeling of belonging and community. In this respect, platforms that are superficially similar can make very different decisions. For instance, a chat group in WhatsApp (allowing two-way communication) is limited to 256 members, while it can comprise 200,000 members on Telegram.[8] Moreover, a Telegram channel (one-way communication) can have an unlimited number of members. This could reflect different objectives regarding, e.g., the vulnerability to spam, the intent to cultivate a feeling of small community. It is notable that in May 2022 WhatsApp announced it would double the group limit to 512 members, as well as add other features such as emoji reactions and the support for larger files (from 100 MB to 2GB),[9] possibly in a bid to keep up with Telegram's growth.

Content-sharing platforms decide on what media can be distributed, including text, audio, pictures, and video, as well as the format of it (length, availability). Moreover, platforms often provide software to enhance the media. For instance, Instagram provided distinctive filters to enhance the poor quality of the photos taken by early smart phones, in an example of making a technical limitation an opportunity for product differentiation. Sometimes limiting what can be done on the platform turns out to contribute to making the platform attractive to users. Twitter's early 140-character limit was due to limitation of the SMS short messaging protocol but serendipitously turned out to lower barriers to post and enable the development of a distinctive discursive style. Later, Snap pioneered the ability to send time-limited messages to create emotional closeness. TikTok's short video format

8. Telegram, "Telegram FAQ," April 11, 2022, https://web.archive.org/web/20220411101553/https://telegram.org/faq; "WhatsApp Help Center – How to Create a Group," January 26, 2022, https://web.archive.org/web/20220126061618/https://faq.whatsapp.com/kaios/chats/how-to-create-a-group/?lang=en.

9. "WhatsApp Blog," accessed May 11, 2022, https://web.archive.org/web/20220505191251/https://blog.whatsapp.com/.

took a radically different approach to that of YouTube, again using a combination of constraints and, additionally, on-device editing tools to help users produce more creative content. These examples suggest that there are many areas of improvement and innovation in the way platforms permit the easy creation of content to be shared with more users, improving the potential for network effects, and that there is no reason to believe that there is not more innovation to come.[10]

Improving the discovery of user-generated content: Algorithmic filtering

Recommendation algorithms are one of the most criticized features of content-sharing platforms, seen as manipulative, operating under near total opacity, and ultimately prone to prop up unwanted content and negatively influence user attitudes. However, there is some information available about how they work, notably thanks to journalistic reporting, and is worth identifying what problems algorithms are meant to solve and the methods that are used to solve these problems.

As seen in the previous section, a successful content-sharing platform seeks to ensure and facilitate a large supply of content from and for its users. However, the counterpart of a large supply of content is that it becomes harder for users to find content they might enjoy. The practical, and technical, solution to this issue from the viewpoint of a digital platforms is algorithmic filtering – the use of software to serve users personalized recommendations based on their social ties and what can be inferred from their on-platform behavior.

This responds to the problem that users may get lost or miss out on important content, and, crucially, such solution is much more consistent with their business model than human curation.

10. A related way to boost network effects on platform is to add features unrelated to content-sharing that take advantage of the existing network of contacts on platform and exhibit network effects, such as a money transfer system, a messaging service, etc.

Developing software to match content to users in real time represents a fixed cost, and running the software has low marginal cost, which enables economies of scale on the supply side with cost per users going down with the number of users, mirroring network effects in which benefits to users increase with the number of users. Algorithmic filtering can be provided at scale, which would be impossible unless supplied by the users themselves, which defeats the purpose of providing an aid to navigating the content, although many platforms give users tools to edit that is presented to them, for instance by allowing to mute or block certain sources.

The most famous, and arguably, influential of these algorithms is Facebook's Feed (previously "New Feed"), which was rolled out in 2006 at a time when the experience of using Facebook's nascent network consisted of checking the individual profiles of connected users.[11] Centralizing the new changes in one place was an improvement in usability and immediately highlighted the issue of too much information to deal with. While the simplest algorithm is a chronological content feed, the questions of defining and prioritizing what matters from what does not and of presenting users with potentially information they may have missed has been at the heart of all the controversies associated with the use of feeds.[12] In particular, filters are suspected of creating "filter bubbles" whereby users are only served content that confirm their pre-existing opinions and inclinations, putting sharp focus on the role of private firms in the formation of public opinions.[13] The lack of transparency regarding how filters work makes it easy to pick up egregious examples where the filters

11. Farhad Manjoo, "Can Facebook Fix Its Own Worst Bug?," *The New York Times*, April 25, 2017, sec. Magazine, https://www.nytimes.com/2017/04/25/magazine/can-facebook-fix-its-own-worst-bug.html.

12. Robyn Caplan, "Algorithmic Filtering," in *Mediated Communication*, ed. Philip M. Napoli, vol. 7, Handbook of Communication Science (De Gruyter Mouton, n.d.), 561–83, https://doi.org/10.1515/9783110481129-030.

13. Lucy Bernholz, Hélène Landemore, and Rob Reich, "Introduction," in *Digital Technology and Democratic Theory*, ed. Lucy Bernholz, Hélène Landemore, and Rob Reich (Chicago: University of Chicago Press, 2020), 1–22.

seem to have made users reach extreme content that they would have avoided otherwise. This lets open the question of whether such instances are actually commonly occurring unanswered. However, a recent study of YouTube's algorithm suggests that the algorithm tends to serve mild content that is consistent with the viewer's ideology, creating narrow, but not extreme, echo chambers, and not leading users towards extreme content ("rabbit holes").[14]

The nature of these filters is hard to know since they are not public. They can use very simple rules about presenting content subscribed in chronological order, as in Twitter's "Latest Tweet" version of its feed. This can be augmented by giving more weight to content that has created reactions ("likes" on Facebook, or "comments", etc.) in a user's set of contacts, or in the entire network. A simple application is to insert in a feed past contributions that the user missed but that were identified as meaningful, for instance because they caused a lot of reactions. More involved algorithms will rely on machine learning to predict what users will be interested in based on past behavior. As with any machine learning-based application, such approach may yield very good predictions but why it does so may be impossible to understand even by its designers. Moreover, it will not perform well when new issues and behaviors arise. Indeed, a machine-learning based algorithm is optimized on past data and observed behaviors. But behaviors change, new type of issues come up, sometimes precisely to defeat the algorithm in unexpected ways (e.g., by using newly-created euphemisms to avoid censorship), implying a constant game of catching up by platforms as well as constant opportunities to undermine the recommendation algorithm.

In addition to being a pragmatic answer to the discovery of information, akin to how internet search algorithms, like Google's, superseded the early internet directory such as

14. Megan A. Brown et al., "Echo Chambers, Rabbit Holes, and Algorithmic Bias: How YouTube Recommends Content to Real Users," SSRN Scholarly Paper (Rochester, NY: Social Science Research Network, May 11, 2022), https://doi.org/10.2139/ssrn.4114905.

Yahoo's, filter algorithms also prove to be a feature of differentiation between platforms. Platforms are thus also competing on the algorithm they use to improve users' satisfaction and retention. In this regard, TikTok's recent success is noteworthy because it relies on user-generated content, but has few social features, no quasi-chronological "feed" as in previous generation platforms and relies exclusively on an algorithm that weighs the time spent by a user on a content, and proxies for the quality of the content, to serve new videos.[15] The fact that major established platforms (Meta included) are scrambling to react and imitate TikTok is a testament to the strength of the innovation.

MAKING A LIVING AND SURVIVING AS A PLATFORM: MONETIZATION AND PREEMPTION

Monetization and incentives to drive network effects

Creating and sustaining network effects is for naught if the platform cannot generate revenues to cover its costs. As mentioned above this is not necessarily easy and it often involves discriminating between sides, by making some participants who have a high willingness-to-pay to participate cover the costs for the others. This gives rise to advertising-based models – themselves already at the basis of the business model of print newspaper, whose revenues was for instance highly dependent on classified ads. Some commentators are making the ad-driven monetization one of the root causes of many perceived failings of platforms and the reason for their driving so hard for user acquisition and retention. However, that position arguably confuses cause and effect. Instead, one can argue that the root cause is platforms competing against each other to generate and exploit network effects and some of them using advertising as a means for monetization. Advertising may exacerbate the perceived

15. Ben Smith, "How TikTok Reads Your Mind," *The New York Times*, December 6, 2021, sec. Business, https://www.nytimes.com/2021/12/05/business/media/tiktok-algorithm.html.

failings of platforms, but even platforms that do not rely on advertising, or are not even for-profit, still must very carefully manage network effect and rely on tactics that are not that different to increase their user count.

Alternative monetization models that do not entirely rely on advertising, such as a mix of subscriptions, are available and sometimes chosen. For instance, the videoconference company Zoom, which is a platform enjoying same-side network effects, uses a combination of subscription for revenues and a free tier for facilitating user acquisition. These monetization models are contingent on the specifics of how the platform works. Zoom could have tried to have ads inserted in video calls, and conversely, Facebook could have a subscription-based model. What is fundamentally common to Facebook and Zoom is the reliance on network effects to jump start and maintain their business. The tactics, and the monetization align with that effort, but do not determine it. Put another way, one can argue that Facebook is using ads rather than subscription because it found that its users were not really bothered by ads, and thus that ads did not go in the way of network effects while asking for a participation fee would severely limit the reach of the platform and curtail network effect. Similarly, putting ads on Zoom would degrade drastically the quality of the product. Therefore, it is a mistake to argue that the root cause of the drive for user engagement by platforms is the advertising-based model. Instead, network effects come first while advertising, or another monetization methods, be it subscription-based, comes second.

The case of Signal, a non-profit organization founded to provide the best privacy possible in messaging, illustrates that competition for users still drives the addition of features that enable network effects even if the organization has no commercial motives. Observer of platforms Casey Newton reported that Signal had a goal to grow to 100 million users to get enough donations to pay for the cost of its operation.[16] To attract users,

16. Casey Newton, "Warning Signal: The Messaging App's New Features Are Causing Internal Turmoil," The Verge, January 25, 2021, https://www.theverge.com/22249391/signal-app-abuse-messaging-employees-violence-misinformation.

the platform plans to include features that are available in competing apps and boost network effects (e.g., payment system, groups communication, anonymous IDs). This created an outcry among employees, often highly motivated by the mission of the organization, who feared that these social features will possibly introduce more bad behavior and disinformation on the platform, without it having any mean to deal with it, since the content is full encrypted.

How secure are network effects?

Successful platforms got into their position by exploiting network effects, but they are at the same time acutely aware that their position is much less secure than what may seem to be. As a result, they first care about maintaining network effect and, as they are afraid of being displaced, they feel compelled to respond to competitors who come up with an innovative way of harnessing network effects. Again, understanding how platforms see threats to their business sheds light on how they allocate resources and attention, notably on the issues that concern scholars of disinformation.

Platforms' responses to competitive threats are two-fold: preemption by acquisition and imitation. Facebook's (now Meta) successive acquisitions of WhatsApp and Instagram are in line with this viewpoint. As new generations of platforms emerge that may eat into Facebook's usage, Facebook acquired them at a high premium which reflected not just the value of the asset and current and future user base (WhatsApp was acquired for $19 billions in 2014) but also the insurance that no one else than Facebook would be controlling it. The fast emergence of TikTok, thanks to a very different user experience, dented the revenues of Facebook and YouTube, prompting immediate imitation by these two incumbents. Twitter creates Spaces to compete with emerging platform Clubhouse which is based on voice conferencing.

The high tempo of product innovation and continuous attempts to enter suggest that the largest platforms are facing some tangible competition and do not take for granted their survival.[17]

One key reason for the fluidity of market shares is that the move to mobile phone-based internet usage has facilitated the "multi-homing" of users, i.e., users are simultaneously active on several platforms, depending on the need and the social circle they are involved with. Another reason is that there has been constant innovation in the format of content-sharing with limited time availability formats (Snapchat), short text (Twitter), short videos (TikTok), voice (Clubhouse), images saved for later (Pinterest), recommendations based on social circles, etc.

Keeping in mind that there has been a constant stream of innovations in terms of formats in content-sharing platforms matters especially because the discourse about the role of these platforms in the public sphere and whether and how they should be regulated is arguably very much influenced by what has been know of issues that have been commented on the largest and most established platforms (especially Facebook, YouTube, Twitter) while these are not necessarily representative of where the industry is headed.

17. This assessment is made purely based on their competitive strategy and is not meant to reflect whether any of these firms maybe contravening or not to competition law, which uses totally different standards to make such determination.

II. SUBVERTING NETWORK EFFECTS: CONTENT-SHARING PLATFORMS' AND INFORMATION MANIPULATIONS

The explanation of the internal logic of content-sharing platform, and the associated analytical toolbox, sets the scene for the analysis of attempts to subvert the platforms for political gain. The August 2018 joint CAPS-IRSEM report defines "information manipulation" as comprising three components: "a coordinated campaign, the diffusion of false information or information that is consciously distorted, and the political intention to cause harm."[1] The report pointed out that content-sharing platforms have become a leading conduit for information manipulation.

While it feels now self-evident that such platforms have taken a central role in most information manipulations, it is worth elaborating on how networks between users, as they are enabled by platforms, may matter to the diffusion of manipulated information, as well as the platforms' incentive to deal with this diffusion depending on which users are relaying manipulated information. Moreover, notwithstanding the importance of networks, it also matters to understand the extent to which modern information manipulations depend on digital content-sharing platforms in their modus operandi.

1. Jean-Baptiste Jeangène Vilmer et al., "Information Manipulation: A Challenge for Our Democracies" (Paris: Policy Planning Staff (CAPS) of the Ministry for Europe and Foreign Affairs; The Institute for Strategic Research (IRSEM) of the Ministry for the Armed Forces, August 2018), 21, https://www.gouvernement.fr/sites/default/files/locale/piece-jointe/2019/10/11_against_information_manipulation.pdf; Jean-Baptiste Jeangène Vilmer et al., "Les Manipulations de l'information : un défi pour nos démocraties" (Paris: Centre d'analyse, de prévision et de stratégie (CAPS) du ministère de l'Europe et des Affaires étrangères ; Institut de recherche stratégique de l'École militaire (IRSEM) du ministère des Armées, August 2018), https://www.diplomatie.gouv.fr/IMG/pdf/les_manipulations_de_l_information_2__cle04b2b6.pdf.

THE DIFFUSION OF INFORMATION ON PLATFORMS: INSIGHTS FROM NETWORK THEORY

An important framework to assess the diffusion of disinformation is Ben Nimmo's "Break out scale."[2] The scale seeks to provide a measure of the success of an information manipulation mainly along the dimensions of whether a piece of disinformation is carried across a single or multiple platforms and whether disinformation is relayed by celebrities or journalists. Nimmo specifically warns that individuals who have a large audience are particularly likely to be the target of information manipulators as they wield a large influence that can be exploited to spread disinformation.

It is useful to recast some of Nimmo's framework using the language of network theory to understand the incentives that content-sharing platforms must restrict the activities of some of their members and, conversely, to set up how the modus operandi of information manipulators attempts to take networks structure into account, drawing on the sociological analysis of networks.[3]

The first relevant concept is that of *small world*.[4] Social network, online and otherwise, very often have a structure whereby any two members of the network are very unlikely to have a direct tie, but that the average path length between two members, i.e., how many steps it takes to go from one user to another following direct links, is very low. This captures the idea of "six degrees of separation" – it only requires a few social intermediaries to

2. Ben Nimmo, "The Breakout Scale: Measuring the Impact of Influence Operations," *Brookings* (blog), September 25, 2020, https://www.brookings.edu/research/the-breakout-scale-measuring-the-impact-of-influence-operations/.

3. The sociological analysis of social networks offers many insights into the role of the structure of networks as determinant of social influence. For an introduction by a prominent contributor: Emmanuel Lazega, *Réseaux Sociaux et Structures Relationnelles*, Presses Universitaires de France, Que Sais-Je?, 2014.

4. Duncan J. Watts, *Small Worlds: The Dynamics of Networks between Order and Randomness*, Princeton Studies in Complexity (Princeton, N.J: Princeton University Press, 1999).

find a connection between two people picked randomly. A social structure that has a small world property is one where most individuals are embedded in small collections of sets of densely connected people ("clusters") and a few individuals have connections that span different clusters. These individuals are those who connect multiple clusters and who give the network structure a "small world" property.

To think about the role of these individuals, we can use two other concepts of network analysis: *structural holes* and *brokerage*. Because they sit between otherwise disconnected communities, these individuals are positioned in a structural hole. The structure of the network (not their formal position) gives them a gatekeeper role and sociological theory suggests that they may enjoy benefits from this position that allow them to engage in *brokerage* (e.g., information arbitrage, rent extraction).[5] Armed with these concepts we can see that being an influencer is not simply a matter of counts of connections. What also matters is whether these connections are confined to a well-connected cluster, or whether they are cutting across structural holes within a given network, or even between networks. In summary, heterogeneity among network members and their value for spreading disinformation is contingent on the structure of the network itself and on the member's position within the network. A celebrity has, by definition, a wide appeal and is likely to be bridging several networks, or clusters within a network. However, not all members who are in such position are necessarily celebrities, and they can still be very relevant to the diffusion of disinformation.

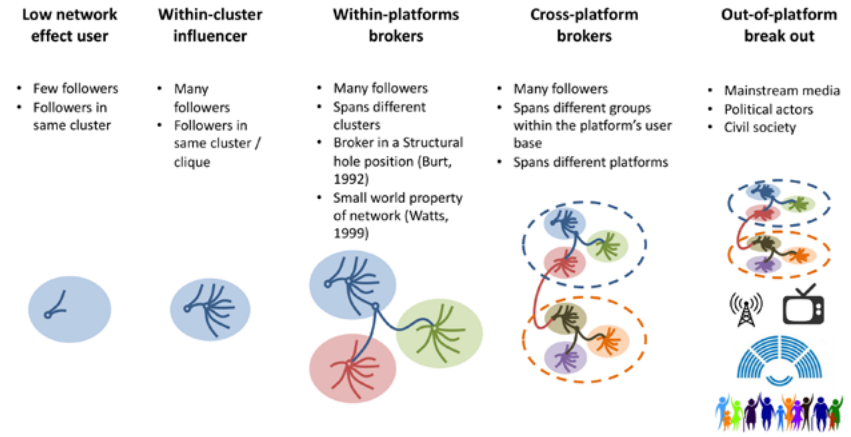
Figure 2 illustrates different stages of information diffusion in relationship with the structural network position of the network members, culminating, building on Nimmo's framework, with disinformation breaking out of digital networks and into the mainstream media and civil society. Framing the diffusion of disinformation in terms of network structure provides useful background for understanding how information manipulators'

5. Ronald S. Burt, *Structural Holes: The Social Structure of Competition* (Cambridge, Mass: Harvard University Press, 1992).

modus operandi depends on it and, conversely, how content-sharing platform depend on key individuals for network effects.

Figure 2

Network structure and influential positions for information dissemination



Adapted from B. Nimmo (2020) The breakout scale: Measuring the Impact of Influence Operations; Brookings

PLATFORMS AND NETWORK EFFECTS IN THE MODUS OPERANDI OF INFORMATION MANIPULATION

Information manipulators seek to spread their message for maximum effect, broadly and covertly. The intrinsic virality, due to network effects, of digital content-sharing platforms provides an unprecedented opportunity for spreading divisive messages. However, even information manipulations relying the most on digital platforms can go beyond supporting the diffusion of a specific message and pave the way for other covert operations (e.g., surveillance, cyber-espionage). Conversely, information manipulations that take advantage of digital platforms may be supported by a range of offline activities. While platforms, and those who study them, may concern themselves narrowly with

what is happening on digital platforms and the effect on users, it is important to attempt to see how actions on content-sharing platforms matter versus other means of action from the standpoint of those who seek to exploit them.

The contrast with a world without content-sharing platforms is instructive to seize what possibilities have been enabled by digital platforms. Pre-digital platform information manipulations, best exemplified by Soviet bloc "active measures", would follow the script of devising an embarrassing story as well as gathering a mix of true and forged documents to support it. The best smear is one that throw shades on the target and makes it hard to issue a credible denial either because it's hard to prove a negative or because there is no uninterested third party that can credibly vouch for the target. It should be kept in mind that the smear, or the lie, and disinformation itself is not a goal, it's merely a mean. The goal of a manipulation is political: sow the seeds of disunity within an adversary's political system to gain political advantage.

The next step is to get the smear out, and get it repeated and picked up by gatekeepers of information such as major news publications or well-respected personalities in politics or in civil society so that the story breaks through to the mainstream.[6] The main constraint for a successful manipulation is to be given credibility by these gatekeepers, who, unless they are complicit to the operation, need to believe the story to be genuine. In turn, this requires careful pre-diffusion research and targeting and often the crafting of elaborate forgeries, mixing a few fake elements with many true ones, to maximize the chance that they pass the filter of the gatekeepers.[7] As a result, a manipulation would often

6. Thomas Rid, *Active Measures: The Secret History of Disinformation and Political Warfare* (New York: Farrar, Straus and Giroux, 2020).

7. Ladislav Bittman, *The Deception Game: Czechoslovak Intelligence in Soviet Political Warfare*, [1st ed (Syracuse, N.Y.]: Syracuse University Research Corp, 1972), available at https://archive.org/details/400437397TheDeceptionGameLadislavBittman1972Pdf. Bittman, a defector from the Czechoslovak Communist-era intelligence services, writes: "For disinformation manipulations to be successful, they must at least

be a high-stake operation involving detailed planning with a relatively small initial target. This also demonstrates that the content of the information manipulation (message, forged evidence related to it) is not independent of the means of diffusion and in particular the anticipated scrutiny of gatekeepers.

Content-sharing platforms significantly change the calculus of information manipulation by making these operations cheaper to run and by reducing their dependency on information gatekeepers, expanding dramatically the set of valuable targets, but also possibly changing the importance and type of preparatory work.

For one thing the creation of content is easier and cheaper thanks to the popularity of digital creation tools made available by the progress of personal computing while putting content out is what platforms themselves are optimized to do. This enables tactics that seek to produce and spread less-developed fake information, relying on mass rather than quality[8] and is facilitated by the availability of third-party contractors providing infrastructure and support. Moreover, the question of breaking into the mainstream is posed in different terms. Traditional gatekeepers, while ultimately still crucial to the formation of public opinion, are no longer the only way to access the mainstream audience and are themselves more vulnerable to picking up sensational view considering their precarious business models and their increased competition. Another road to the mainstream would be to reach as much as possible of the audience on platform which, by some measures, literally constitutes the mainstream

partially correspond to reality or generally accepted views. A rational core is especially important when the recipient enemy or victim is a seasoned veteran in such matters, because without a considerable degree of plausible, verifiable information and facts it is impossible to gain his confidence. Not until this rational skeleton has been established is it fleshed with the relevant disnformation." (p. 21)

8. This is reminiscent of comments made by Bittman. See that author's description, in a pre-digital platform world, of what he disdainful calls "propaganda" (p. 23), a mode of operation where volume of output served to the press, however unsophisticated, is more important than any actual effect.

given their penetration in the population. Moreover, carefully targeting key influencers, instead of starting from low importance users could provide a powerful shortcut.

To do this, one mode of operation can be to get what seems to be grass-root mobilization of individuals ("AstroTurf"[9]) generated by false accounts passed to individuals who generate significant network effects on the platform ("influencers") who will then pass it to their own audience, out of gullibility, ideological proximity, or plain corruption.

At the inception stage, the manipulation consists of generating Astroturf, creating fake accounts fed by paid trolls or bots, attempting to trick the filtering algorithm into referring the story they push to more users. At this stage, a manipulation may easily die if the combination of the appeal of the story and the subversion of the algorithm is not enough to get traction. A platform will also have no qualms putting down these at this stage of discovery since it is but background noise among everything that happens on the platform.

At the pre-mainstreaming stage, the story has reached "influencers". Some maybe specialized in recycling controversial information as known "whistleblowers" or "hacktivists", but these can present the inconvenient of having small audiences and specialized reach. Minor celebrities or professional internet influencers are much more powerful because taking them down or downranking them creates a dilemma for the platforms since these influencers are, by definition, high contributors to network effects in their regular, non-information manipulation-related, activities.[10] In this case, the business incentives of the platforms are no longer neutral or aligned. If the disinformation is not credibly discredited, the platform may be wary to intervene as it runs the risk of alienating key contributors and users. This also implies that the ideal stories to be pushed needs to be outrageous

9. AstroTurf is a commercial name for a brand of artificial carpeting used in sport facilities. The term is used derisively to refer to covert efforts mimicking the efforts of concerned citizens, thus mimicking "grass root" citizen mobilization.

10. Chatain and Kaul, "No Easy Way Out? Platform-Mediated Political Externalities and Platform Strategy."

or lurid enough to be rapidly picked up while at the same time being hard to easily shoot down. Many conspiracy theories could fit this pattern. Finally, in the final stage of diffusion, the traditional gatekeepers, who are also paying attention to what is being said on key platforms, and out of fear of being left behind, are also reporting it. Here, the tension is that what may be interesting to repeat and hard to deny, may still not be worth reporting according to high journalistic standards.

A few recent examples of information manipulations show the variety of the ways they include content-sharing platforms in their modi operandi.

A most platform-centric example of operation was uncovered by a December 2021 investigation by the New York Times. The investigation started with examining publicly available request for proposals from the Shanghai police that sought to procure the services of an external agency to build and maintain a network of fake accounts on multiple Western social media platforms.[11] The details are consistent with an operation seeking to gradually build an audience for a general-purpose influence network. Specifically, the winner of the tender (eventually, a 20-employee media agency) was tasked with creating several hundred accounts under disguise, maintain them and ensure they attract some followers. Video content (2–3-minute videos) was to be produced and regularly updated.

The success metrics were the survival rate of the accounts against the platforms' take-down efforts, their ability to attract followers and especially their ability to climb among the list of recommendation. This modus operandi is consistent a run-of-the-mill operation that requires relatively little specific skill so that it can be the subject of a public procurement process. Moreover, the operation seemed aimed at creating low-influence users ("Low network effect users" in Figure 1) and hope that some will

11. Muyi Xiao, Paul Mozur, and Gray Beltran, "Buying Influence: How China Manipulates Facebook and Twitter," *The New York Times*, December 20, 2021, sec. Technology, https://www.nytimes.com/interactive/2021/12/20/technology/china-facebook-twitter-influence-manipulation.html.

graduate to "within-cluster influencer". Suppliers of such services are plentiful around the world and very cheap to contract according to several NATO Stratcom studies.[12] Interestingly, the request for proposal mentioned that part of the job was to find out the personal details of some real users posting certain type of online content (presumably considered disloyal), which appears to be a precursor to the more involved surveillance of users based in China. This last detail suggests that even the most routine information manipulations, those that are deployed entirely remotely and staffed by third-party personnel, can still be embedded in a larger intelligence operation.

Some manipulations may instead eschew the hard work of slowly building an audience. Indeed, platforms will remorselessly eliminate fake accounts if identified as such, so much that the mere survival of these accounts is a measure of success. As hinted by Figure 2, it might be more productive to directly enlist established influencers rather than creating new ones *ex nihilo*. Consistent with this are reports of YouTube influencers being approached by intermediaries who would supply the influencers with ready-made content and provide substantial compensation. The goal was to spread anti-vaccine messages during the Covid-19 epidemic for, and after the 2022 full-scale Russian invasion of Ukraine, to push pro-Russian talking points.[13] It is very hard to gauge the size of this phenomenon, since we do not know how many attempts were made and how many succeeded. This suggests however that additional covert capabilities, such as cut-outs to channel funds, can be mobilized to support information manipulation on platforms once operations more complex than

12. Rolf Fredheim and Martha Stolze, "Robotrolling 2022," *Robotrolling*, NATO Strategic Communications Centre of Excellence, no. 1 (2022), https://stratcomcoe.org/publications/robotrolling-20221/243.

13. "The YouTubers Who Blew the Whistle on an Anti-Vax Plot," *BBC News*, July 24, 2021, sec. BBC Trending, https://www.bbc.com/news/blogs-trending-57928647. "Guerre En Ukraine : Une Opération d'influence Russe Vise Des Youtubeurs Français," accessed June 10, 2022, https://www.marianne.net/monde/europe/guerre-en-ukraine-une-operation-dinfluence-russe-vise-des-youtubeurs-francais.

the stereotypical model of fake account creation and amplification are accounted for.

Indeed, some complex, ongoing operations can still involve social network only as part of a whole spectrum approach involving several types of media as well as covert on-the-ground activity. Scholars have speculated that GRU operations have been using classic methods based on planting news stories in the press, while the Internet Research Agency (IRA) tended to be more adept at tailoring message to specific on platform audiences, arguing that the latter was more successful than the former.[14] However, recent informational operations targeting multiple African countries by organizations associated with Russian tycoon Yevgeny Prigozhin (who is also the founder of the IRA) show that multiple media are mobilized and coordinated, including an online press agency, multiple online outlets, opinion polls, a private military company (the Wagner group) and even the production of feature length movie supporting the actions of Wagner.[15] In the context of Figure 2 and Nimmo's breakout scale, the influence operation is starting from the far right of the figure: multiple platforms and multiple traditional media are seeded simultaneously. One can speculate that the intention is to present a full-fledged information environment with consistent themes rather than undermine one that is already established.

Success of information manipulation is hard to define. For one thing, it should not be underestimated how hard it is to make them work. Since we only know about the attempts that have been detected, it is difficult to know how many attempts are failing. As such, the data available presents a severe selection

14. Renée DiResta, Shelby Grossman, and Alexandra Siegel, "In-House Vs. Outsourced Trolls: How Digital Mercenaries Shape State Influence Strategies," *Political Communication* 39, no. 2 (March 4, 2022): 222–53, https://doi.org/10.1080/10584609.2021.1994065.

15. Maxime Audinet and Colin Gérard, "Les « libérateurs » : comment la « galaxie Prigojine » raconte la chevauchée du groupe Wagner au Sahel," *Le Rubicon* (blog), accessed July 15, 2022, https://lerubicon.org/publication/la-galaxie-prigojine-promoteur-de-wagner-au-sahel/.

bias. They report attempts that have been detected and thus are likely to be more successful than the baseline (survivor bias), but also because platforms themselves are not consistent or forthcoming in sharing what they know (reporting bias).[16] What would be needed to assess fully the supply side of information manipulation would be internal documentation from one of the actors.[17] Moreover, the goals of an information manipulation may range from simply maintaining an online presence (as in the Shanghai police request for proposal case), to assist a complex on-the-ground operation involving hundreds of mercenaries (as in Prizhogine's involvement in Africa). All these involve an on-platform component to the operation that cannot be detached from the goals of the operation itself, which can be very diverse. Finally, because these operations can be cheaply run while achieving visibility even when they are uncovered, they may also be launched to answer the need to fulfill a bureaucratic imperative of appearing to be doing something.

As far as content-sharing platforms are concerned, the fact that the on-platform component of information manipulation can be part of a much larger picture only complicates matters because the platforms' content moderation capabilities are not primarily tuned to dealing with such threat.

16. Camille François and Evelyn Douek, "The Accidental Origins, Underappreciated Limits, and Enduring Promises of Platform Transparency Reporting about Information Operations," *Journal of Online Trust and Safety* 1, no. 1 (October 28, 2021), https://doi.org/10.54501/jots.v1i1.17.

17. The indictments related to the Russian meddling with the 2016 US presidential election are instructive because they revealed many attempts to combine online and offline actions (e.g., recruiting people on Facebook to join a made-up demonstration). Yet it is not clear that they are representative of current operations.

III. THE CHALLENGE OF CONTENT MODERATION ON PLATFORMS REGARDING INFORMATION MANIPULATIONS

For a platform, the challenge posed by information manipulations is only one among those related to content moderation and, more generally, "trust and safety." The activity of content moderation can represent a large part of the expenses and personnel for a platform and has become a key issue for their management teams. Information manipulations present specific characteristics that make them even costlier and harder to deal with through usual content moderation methods. To understand how the combat against disinformation works and the tradeoffs that platforms make, it is necessary to place them in the context the systems and tools that have been developed for content moderation at large. What is unique about content-moderation on content-sharing platforms is the scale needed to tackle this issue, reflecting the scale of the platforms themselves. Moreover, by using their discretion to moderate content at such scale, platforms are taking on what amounts to a regulatory role.[1]

"CONTENT MODERATION" AND "TRUST AND SAFETY"

"Content moderation" is the set of policies and processes by which platforms, as intermediaries, decide to remove or put qualifications on the content posted by users on their digital property. Content moderation is a subset of platforms' "trust and safety" policies and processes whose goal is to ensure that users are and feel safe using the platform. Trust and safety policies' aims to protect users from the threats that can be encountered while using a platform, including, e.g., financial fraud, hacking,

1. Romain Badouard, *Les Nouvelles Lois Du Web: Modération et Censure* (Paris: Seuil, La République des idées, 2020).

harassment, and harmful content (including disinformation and misinformation).

Platforms have long recognized that it is in their economic interest to exercise some content moderation, notably by removing the most objectionable content from the platform, even if they may not be legally obliged to.[2] By removing such content platforms can ensure the continuing involvement of their users. Kate Klonick's seminal work[3] describes the content moderation process as happening at different stages, as follows: moderation can happen *ex ante* by preventing the upload of content that is algorithmically recognized as harmful or illegal, thanks to the existence of specific databases. *Ex post* moderation deals with content already on the platform. Some of this moderation can done *proactively*, whereby problematic content is actively looked for, and some can be done *reactively*, in response to complaints. All moderation is based on a set of rules and guidelines that has developed over time, with variation between platforms based on the history of the challenges they faced, as well as the influence of the principles set by their management.

A form of soft content moderation can happen outside of the legal-like process described by Klonick. Content deemed potentially problematic form the viewpoint of the platform, but that does not fall into the existing internal guidelines justifying an immediate takedown can be deprioritized (downranked) in the recommendation engine of a platform. This can be done in a manner tailored to the preferences of a user.[4] Such soft moderation gives a lot of leeway to platforms for preemptively reducing the exposure given to content, even if there is not a formal determination that the content is not permitted per the guidelines.

2. Kate Klonick, "The New Governors: The People, Rules, and Processes Governing Online Speech," *Harv. L. Rev.* 131 (2017): 1598.

3. Klonick.

4. E.g., from Instagram: "How We Address Potentially Harmful Content on Feed and Stories," accessed March 29, 2022, https://web.archive.org/web/20220120213847/https://about.instagram.com/blog/announcements/how-we-address-harmful-content-on-feed.

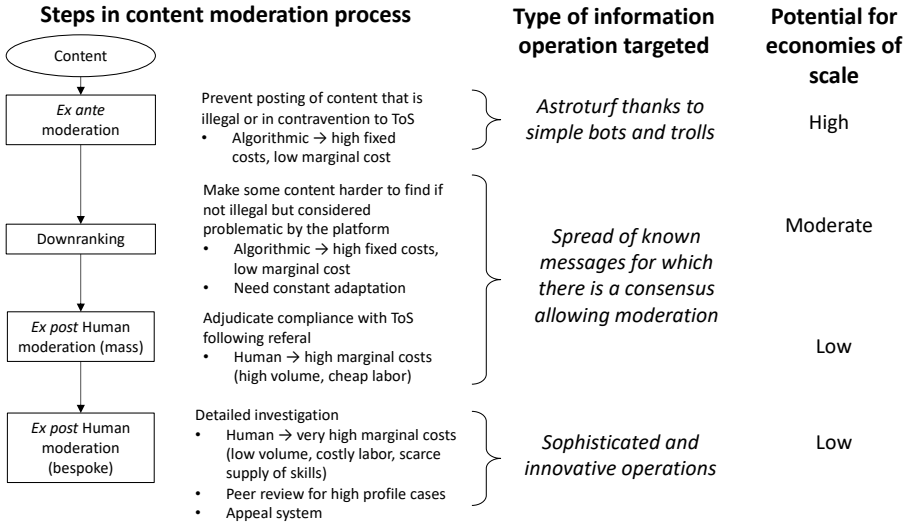
The IT systems that oversee downranking are notoriously opaque and not always working as intended. Facebook's own downranking system sometimes produced effects opposed to its intent from 2019 to 2021 due to a bug.[5] The issue took months to detect and fix. That such incident can happen in a well-resourced organization suggests that other such bugs can easily be active if their effects are not immediately evident, and that smaller platforms may have trouble setting up such systems in the first place. In any case, while downranking is very attractive in principle, it is neither cheap or simple to implement and may lack transparency, even for its own designers. It can also generate false positives, for instance by incorrectly taking down a news report because it refers to content that contravenes the terms of uses.[6] However, the analysis of the economics of platforms suggest that they have an interest in prioritizing this form of moderation, since it plays on their strengths in devising and deploying systems that are conducive to economies of scale. Platforms will tend to favor solutions that are amenable to economies of scale, in particular favoring using software solutions that can be deployed over many cases, over relying on human judgement to deal with individual cases. Figure 3 shows a stylized representation of the content moderation process and how it is conducive to economies of scale.

5. Alex Heath, "Facebook's Algorithm Was Mistakenly Elevating Harmful Content for the Last Six Months," The Verge, March 31, 2022, https://www.theverge.com/2022/3/31/23004326/facebook-news-feed-downranking-integrity-bug.

6. "How Silicon Valley Is Helping Putin and Other Tyrants Win the Information War," *Coda Story* (blog), April 14, 2022, https://www.codastory.com/authoritarian-tech/facebook-authoritarians-information-war/.

Figure 3

Content moderation process and economies of scale



THE NUTS AND BOLTS OF CONTENT MODERATION

A detailed examination of the content moderation process, consistent with what reporting and leaks has shown, suggests that that process can be very costly to implement and maintain for content-sharing platforms. Moderation seeking to combat disinformation and target instigators of information manipulations exhibits features making it even more resource-consuming than other forms of moderation.

The bureaucratic process of content moderation

The establishment and maintenance of terms of uses and community guidelines can consume a significant part of the top management's time and attention. This is necessarily a top management issue because it sits at the intersection of factors

fundamentally affecting the competitive position of the platform (what product to offer, for whom, drawing on which resources and capabilities) and factors affecting relations with stakeholders, such as governments and NGOs. Such tradeoffs can only be resolved by the top management. The Facebook Files[7] showed how content moderation issues, both pertaining to the establishment of rules, as well as high profile individual cases, were escalated up the organization and ended up being dealt with by the CEO and the close circle around him. There is wide variation in organizational resources depending on the size of the platform.[8]

The human review process requires a very large workforce, especially as digital platforms have very few employees relative to the size of their operations. Meta (then Facebook) declared in September 2021 that it had about 40,000 persons working on trust and safety issues, up from 10,000 in 2016.[9] This number likely include many external contractors and is to be compared with a headcount for Meta of about 72,000 at the end of 2021. Under the most extreme assumption that all trust and safety work is contracted out, this implies that for every two Meta employees

7. See, e.g., reporting on the "Facebook Files", leaked by Frances Haugen. The New York Times, "The Facebook Papers and Their Fallout.," *The New York Times*, October 25, 2021, sec. Business, https://www.nytimes.com/2021/10/25/business/facebook-papers-takeaways.html. Jeff Horwitz, "Facebook Says Its Rules Apply to All. Company Documents Reveal a Secret Elite That's Exempt.," *Wall Street Journal*, September 13, 2021, sec. Tech, https://www.wsj.com/articles/facebook-files-xcheck-zuckerberg-elite-rules-11631541353. Keach Hagey and Jeff Horwitz, "Facebook Tried to Make Its Platform a Healthier Place. It Got Angrier Instead.," *Wall Street Journal*, September 15, 2021, sec. Tech, https://www.wsj.com/articles/facebook-algorithm-change-zuckerberg-11631654215.

8. Robyn Caplan, "The Artisan and the Decision Factory: The Organizational Dynamics of Private Speech Governance," in *Digital Technology and Democratic Theory*, ed. Lucy Bernholz, Hélène Landemore, and Rob Reich (Chicago: University of Chicago Press, 2020), 167–90.

9. "Our Progress Addressing Challenges and Innovating Responsibly," *Meta* (blog), September 21, 2021, https://about.fb.com/news/2021/09/our-progress-addressing-challenges-and-innovating-responsibly/. https://web.archive.org/web/20210921163151/https://about.fb.com/news/2021/09/our-progress-addressing-challenges-and-innovating-responsibly/.

not working on trust and safety, there is roughly one Meta employee or contractor working on trust and safety. Meta also contended having spent more than $13 billions in personnel and technology in the same time span. Basic calculations show an increase of about 32% per year in personnel and peg the spending per employee or contractor involved per year at a little less than $100,000 USD. Given that Meta's minimum pay for its contracted reviewers is currently $22 per hour[10] (i.e., about $38,500 per year) while entry software engineers salary for Meta in the Silicon Valley are at about $150,000 (bonus included, excluding stock options), the spending figure suggests that a majority of the $13 billions goes into salaries and overhead rather than in technological investment.

This is consistent with recent information on TikTok's efforts to buttress its trust and safety activities in Europe. TikTok has been investing in attracting employees, competing with subcontractors (e.g., Accenture) and direct competitors (e.g., Meta) on wages and benefits to populate its Dublin trust and safety hub. Reporting from the Financial Times suggests that thousands of employees are working for TikTok on these issues in Dublin in 2022 while only a skeleton team was based there at the beginning of 2020. This large growth must have had a large impact on costs which the Financial Times correlated with an increase in financial losses in TikTok's European operations.[11]

Finally, which content should be moderated is typically country-, culturally- and language-dependent. This is a major challenge for any content-sharing platform that is spanning several countries. As a platform develops out of its home market most systems and rules are initially set up around the norms and preferences of the users in the home country, filtered by the

10. "An Update on Compensating and Supporting Facebook's Contractors," *Meta* (blog), May 13, 2019, https://web.archive.org/web/20200203181437/https://about.fb.com/news/2019/05/compensating-and-supporting-contractors/.

11. Cristina Criddle, "TikTok Poaches Content Moderators from Big Tech Contractors in Europe," *Financial Times*, February 15, 2022, https://www.ft.com/content/d03c945b-ed5b-425b-8817-acb236f60931.

potential biases of the personnel of the platform. This already creates cultural blind spots at home. For instance, Pinterest took steps in 2019 to downrank material related to "Plantation wedding" (wedding celebrations taking place in former plantations in the US South that exploited slave labor) on the platform after it was reported that this was considered offensive by many users.[12] Interestingly, no one within the company seemed to have raised this issue, which can be linked to the lack of diversity in tech firms,[13] and after this was pointed to them, the company treaded a fine line between downranking the material and yet keeping it on the platform since it was not formally going against the rules of the platform.

If this type of misstep is easy enough for a firm to make in its home market, it gets worse when platforms bring their model abroad, where cultures are different, the legal norms about free speech vary widely between countries and, of course, where languages are different. It seems paradoxical that a platform can succeed commercially while virtually not knowing the content its users post and consume and that the platform is distributing. Remember that in business model that is based on content-sharing and network effect, usage of a platform can become dominant in a locale even if the makers of the platform literally do not understand anything that is posted. Facebook became the prime social networking platform in Myanmar by 2015, but only had then four employees who spoke Burmese, based in Manila and Dublin, for 7.3 million active users in the country. Content moderation was largely done by English speakers, and the user interface (including the interface to report issue) was only translated

12. Heather Murphy, "Pinterest and The Knot Pledge to Stop Promoting Plantation Weddings," *The New York Times*, December 5, 2019, sec. Style, https://www.nytimes.com/2019/12/05/style/plantation-weddings-pinterest-knot-zola.html.

13. "To Increase Diversity, U.S. Tech Companies Need to Follow the Talent," accessed April 18, 2022, https://hbr.org/2020/12/to-increase-diversity-u-s-tech-companies-need-to-follow-the-talent.

to Burmese that year.[14] Facebook later hired local content moderators through contractors, basing some of them in Myanmar, but this example shows that such approach would need to be repeated for each country in which a content-sharing platform is active, all while platform revenues in many countries can be almost zero for lack of monetization.[15]

Human moderation also has a high human cost on the moderators themselves, creating a burden that is not adequately shared on the firms and by society. Theirs is a low-paying job with high productivity targets. It requires the employees to review material that can be extremely shocking, resulting in significant mental health issues that are not sufficiently addressed.[16]

Dilemmas of account closure: Preventing fake accounts creation and deplatforming influencial users

One major form of moderation does not focus on removing content, but instead removes a user (and its content) from the platform by closing the account if terms of uses, which typically include provisions about account being created by actual persons or organizations, are violated.

Platforms are caught between the incentive to make signing up easy, and the need to prevent the creation of fake and spam

14. "Why Facebook Is Losing the War on Hate Speech in Myanmar," *Reuters*, accessed April 18, 2022, https://www.reuters.com/investigates/special-report/myanmar-facebook-hate/.

15. This begs the question of why, e.g., Facebook, would bother to maintain a presence and "win" the market in countries that are unlikely to make a meaningful contribution to their bottom line in the foreseeable future. One explanation is that such firm, especially if still led by its founder, may be oriented towards the very long term. If the running costs are low, the option value may still be high. Furthermore, remaining active denies an opportunity a competitor, in keeping with the preemption logic intrinsic to platform businesses.

16. Casey Newton, "The Secret Lives of Facebook Moderators in America," The Verge, February 25, 2019, https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona.

accounts which are extensively used by troll farms. Providers of fake accounts (affiliated with governments or not) try to defeat the measures put in place to identify these accounts at the creation stage and stymie their creation. As a result, a cat-and-mouse game develops whereby platforms and fake account providers are trying to outsmart each other. Given this, it is not surprising that platforms report high rates of account prevention and closure, and that at the same time, providers of fake account are still able to deliver many viable accounts to their customers. A recent NATO Stratcom study found that private providers of fake accounts could provide numerous fake accounts on demand and that those were able to stay up for a relatively long time.[17] This does not mean that most attempts at creating fake accounts were successful, only that providers were successful enough to meet their customer's needs. The challenge for platforms is that, by definition, they do not see the successful attempts that they could not detect and can at the same time have prevented the creation of many fake accounts.

A different challenge arises for closing the accounts of real users when they breach terms of use. "Deplatforming" – removing the ability to use a platform to share one's content– can be handled by the regular content moderation process when an average user is concerned. However, deplatforming any user who has achieved an influencer position (i.e., one who has many followers, especially spanning otherwise disjointed clusters) is fraught for a platform. The platform will weigh the reputational damage of keeping the user active with the reduction in network effect from giving up on the user. This calculus involves different set of users in the platform and the platform's external stakeholders and arguably favors delaying deplatforming more influential users as compared to average ones.[18]

17. "StratCom | NATO Strategic Communications Centre of Excellence Riga, Latvia," accessed May 31, 2022, https://stratcomcoe.org/publications/social-media-manipulation-20212022-assessing-the-ability-of-social-media-companies-to-combat-platform-manipulation/242.

18. Chatain and Kaul, "No Easy Way Out? Platform-Mediated Political Externalities and Platform Strategy." HEC Paris Working Paper.

In summary, moderation is costly in the operations it involves to check accounts and activities, and in the potential network effect costs of removing content and deplatforming users. In the details, the costs of moderation can be divided into three categories: first, the costs related to the establishment and update of policies and of systems, second, the operating costs of the moderation function at the level of a country, and third, the country-specific investments that are necessary. The cost of creation and maintenance of content-moderation systems (IT and quasi-legal) are not necessarily fully proportional to the size of the network, which make them easier to bear by the largest platforms relative to the smaller ones, even if these costs are both monetary investment and a tax on the management's attention. However, human moderation costs are likely increasing in the size of the platform because it is related to the size of the user base.

Besides these informed speculations, there is little available hard evidence on the effect of scale on cost of moderation per users. One notable exception is a report by EY which was commissioned by the UK government.[19] This study focuses on video-sharing platforms that are active in the UK. It provides an estimate for the cost of moderation per user among three brackets of size of user bases (small: < 100k users worldwide, medium: between 100k and 10m users worldwide; large: more than 10m users worldwide). Cost per users for medium platforms were estimated approximately to be between in GBP 1.60 to GBP 3.40. By contrast, for large platforms, they were one order of magnitude lower, from GBP 0.25 to GBP 0.50. The authors mention that they have seen "evidence suggesting costs for some of the largest platforms may be materially lower". If these numbers are representative, this is consistent with very strong economies of scale, which suggests that the largest platforms can shift a large part of the content moderation tasks to automated systems.

19. "Understanding How Platforms with Video-Sharing Capabilities Protect Users from Harmful Content Online," GOV.UK, accessed May 17, 2022, https://www.gov.uk/government/publications/understanding-how-platforms-with-video-sharing-capabilities-protect-users-from-harmful-content-online.

MODERATING DISINFORMATION AND DETECTING INFORMATION MANIPULATIONS

After a detour into the inner workings of content moderation, we can address the question of how the basic architecture of content moderation adapts when the problems are the spread of disinformation and the occurrence of information manipulations, in which covert actors organize and coordinate the release of damaging material, some of which straight disinformation, to attain political objectives.

Compared to the posting of illegal material or the curbing of individual-based bad behavior, it is more challenging to characterize disinformation[20] as this requires additional judgement and investigation. Acting on disinformation requires an ability to stiff through potentially immense amounts of material, as well as the management's will to take a stand on the issue. Some of the most successful information manipulations combined a lot of true material with only a limited amount of forgery.[21] Others are simply intended to sow doubt by bringing up tendentious, yet plausible, facts.

For purposes of algorithmic downranking and *ex ante* moderation, a platform would need a large and accurate training sample to devise an accurate filter that removes disinformation but also does not create too many false positive, i.e., content that is not disinformation but is flagged as such. For instance, Pinterest recently decided to remove climate change denialist material from its platform.[22] What is expensive in such case may not be

20. Some authors make a distinction between disinformation and misinformation. Both consist of supplying or repeating incorrect or misleading information, but disinformation implies the awareness the information is incorrect and the intent to mislead while misinformation does not. This distinction is often made by platforms to allow differentiated treatment in terms-of-use: disinformation is considered more serious than misinformation. For instance, disinformation may, e.g., entail taking down a post or closing an account while misinformation may not.

21. Rid, *Active Measures*.

22. "Pinterest Bans All Climate Change Misinformation," *TechCrunch* (blog), accessed April 21, 2022, https://social.techcrunch.com/2022/04/06/pinterest-bans-all-climate-change-misinformation-on-its-platform/.

the technical aspect, but the constitution of an appropriate training sample to teach the classifying algorithm to catch the relevant material while avoiding false positive (legit material that is labelled as infringing the rules). In practice, this can require asking, and paying, many experts to evaluate the appropriateness of a wide variety of material[23] – not a cheap proposition, especially if the media is rich and requires time to be reviewed (e.g., a video). This also requires regular updates as producers of disinformation are adept at using euphemisms to go around these barriers.

Detecting information manipulations also requires different set of skills and resources to be identified. Information manipulations are typically mimicking patterns of real, grass root, information diffusion. The difficulty is to tell a pattern of activity that is truly the result of good faith collective behavior among users from one that is fostered by a set of covert actors. One key notion is that of "inauthentic coordinated behavior" advanced first by Facebook, implicit in the CAPS-IRSEM report,[24] and generalized into the triad of "manipulative actors, deceptive behaviors, harmful content" put forward by Camille François.[25] As we have seen, content can be hard to assess without involving human judgement, while assessing intent and behavior require extensive human work, even if there are tools to assist the investigation. The exact amount of resources that is devoted to this issue is not known, but even for a rich platform, the most binding

23. Daisuke Wakabayashi, "YouTube Moves to Make Conspiracy Videos Harder to Find," *The New York Times*, January 25, 2019, sec. Technology, https://www.nytimes.com/2019/01/25/technology/youtube-conspiracy-theory-videos.html.

24. Jeangène Vilmer et al., "Information Manipulation: A Challenge for Our Democracies"; Jeangène Vilmer et al., "Les Manipulations de l'information : un défi pour nos démocraties. »

25. Camille François, "Actors, Behaviors, Content: A Disinformation ABC – Highlighting Three Vectors of Viral Deception to Guide Industry & Regulatory Responses," Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression (Santa Monica, California, September 20, 2019), https://www.ivir.nl/publicaties/download/ABC_Framework_2019_Sept_2019.pdf.

constraint is arguably not so much money but the workforce available and the time that the specialists have for each case. The supply of specialist skills (OSINT, forensic investigation, computer science, language skills) is limited to a relatively narrow community and the required on-the-job training for effective investigation is extensive, having little to do with basic content moderation activities.

A more subtle issue is the mismatch between how an information manipulation depends on its on-platform component, and how easily detectable it is by the platform. A purely online operation relying on fake accounts and artificial amplification is likely the most detectable and the most vulnerable to moderation. Platforms tools and incentives are best aligned to deal with brute force approaches. How successful they are will depend on the balance of means and ingenuity between attackers and defenders. However, more complex operation that do not fully rely on the use of platforms may be better adapted at avoiding scrutiny if the operation is not exposed elsewhere.

These biases in coverage are exacerbated by the international reach of many platforms. Language and cultural differences exacerbate the mismatch between the resources (especially people) available and the informational terrains on which operations are mounted. This may reinforce a bias toward better coverage of Western, English-speaking countries, to the detriment of other locales.

IV. THE CONSEQUENCES OF THE UPCOMING REGULATION OF CONTENT-SHARING PLATFORM: CONJECTURES AND SCENARIOS

Armed with our analysis of the economics of content-moderation and how they fit in the business model of content-sharing platforms, we can make some informed speculations regarding the impact of the push for the regulation of platforms in Western countries. The largest, US-based, content-sharing platforms have developed content moderation policies mostly in response to commercial interests,[1] and possibly to forestall future regulation, as US law broadly exempts from responsibility over the third party content they make available thanks to "Section 230" provisions.[2] However, and early on, other countries' courts started to ask platforms to comply with local laws as in the Yahoo! Case in which a French court obliged Yahoo!, a US firm, to remove Nazi memorabilia from its auction site on the ground that this contravened French law and that French users could use Yahoo!'s services.[3] Finally, the importance the political impact of on-platform activities and the urge not to leave it to self-regulating firms came to the fore after the revelations of attempts to manipulate the 2016 US presidential election by Russian covert actors as well as the alleged role of platform data to influence the outcome of the Brexit vote.

As of the writing of this note, the most consequential effort that may come to fruition is the Digital Service Act (DSA) to be enacted by the European Union. The DSA is not yet finalized but its broad shapes are already known, and it is possible to consider how competition between content-sharing platforms may be affected. This may in turn change which platforms are going

1. Klonick, "The New Governors: The People, Rules, and Processes Governing Online Speech."

2. Section 230 is a section of Title 47 of the US code which regulates communications.

3. "Yahoo Loses Nazi Memorabilia Case," *Financial Times*, January 13, 2006.

to increase their efforts to combat disinformation, and we can speculate about how malicious actors may themselves react in the design of their operations.

THE EU'S DIGITAL SERVICE ACT AND THE MARKET STRUCTURE OF CONTENT-SHARING PLATFORMS

The European Union wields a significant regulatory influence over businesses worldwide notably because of the size of its market. As a result, the EU, when it passes legislation, is one of the most consequential setters of regulatory standards.[4] The DSA is a wide-ranging set of regulations aiming, among other things, to provide a unified framework to deal with the safety of users of digital platforms, notably regarding online disinformation.[5] For the purpose of this analysis, we can highlight the following features.

The DSA's scope is all platforms active in the European Union, including content-sharing platform. The DSA creates obligations of due process (e.g., information and appeal possibilities), transparency, and reporting for dealing with complaints related to illegal and harmful content, including cooperation with third-party "trusted flaggers" who could collect complains of users. Moreover, for the very large online platforms (VLOPs), who serve more than 10% of the EU's population (corresponding to a threshold of 45 million monthly active users), additional reporting, especially on advertising targeting and recommendation systems is required, as well as more comprehensive and proactive content-moderation.[6] From a cost perspective, this implies more costs for all platforms, and even more so for platforms above the 45 million monthly users threshold. The question of the cost of

4. Anu Bradford, "The Brussels Effect," *Northwestern University Law Review* 107 (2013 2012): 1.

5. "The Digital Services Act Package | Shaping Europe's Digital Future," accessed May 17, 2022, https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package.

6. Daphne Keller, "What Does the DSA Say?," accessed May 17, 2022, https://cyberlaw.stanford.edu/blog/2022/04/what-does-dsa-say-0.

compliance for the smaller platforms is a real one because they may not have the financial and organizational capability to put in place the systems necessary to comply with the DSA. Specialized vendors may step in to provide such solutions at a cheaper cost, but it is too early to know how penalizing the DSA will be for the smaller firms.

From a market structure perspective, an increase in costs for all firms favors the firms that have a larger scale, especially if the costs are fixed rather than variable, as seems to be the case for content-moderation.[7] In this respect the DSA might freeze the established positions of the largest platforms especially if key provisions of the DSA can be implemented via automatic systems which involve large, fixed investments. Moreover, the existence of a threshold for the more severe obligations will in effect dissuade smaller platforms to cross this threshold unless they expect to grow substantially above the threshold. In short, while the DSA may reduce the profitability of the very large platforms, they are also making them less vulnerable to competition for the long run.

There are more subtle effects to expect regarding the new types of platforms that will be created. For one thing, an increase in costs, in this case related to mandated content-moderation, will likely reduce entry of new innovative platforms, even if costs are apparently low, as seen in the case of the GDPR.[8] For another, this perversely incentivizes platform innovators to seek business models for which moderation costs are intrinsically low, for instance by insisting on privacy and end-to-end encryption. In an encrypted environment, only the sender and receiver can read communication, but not intermediaries, so external scrutiny and investigation of mediated content is drastically reduced.

Finally, a possibility is the increased breaking down of platforms boundaries along national and regional regulatory lines.

7. "Understanding How Platforms with Video-Sharing Capabilities Protect Users from Harmful Content Online."

8. Rebecca Janßen et al., "GDPR and the Lost Generation of Innovative Apps," May 2022, https://doi.org/10.3386/w30028.

While some platforms have early on taken the opportunities to develop across borders, they are now caught between different, and possibly conflicting regulatory regimes. New, emerging platforms may decide to choose a camp, and stick to it. TikTok's difficulties to separate its US data from Chinese access is illustrative of the practical hurdles for internal separation.[9]

MEDIUM-TERM IMPLICATIONS FOR INSTIGATORS AND SUPPLIERS OF INFORMATION MANIPULATION

The first-order implication of DSA are more resources and efforts poured into content moderation by the major platforms, which will reduce the relative effectiveness of information manipulations on those platforms. While it is not clear that efforts exerted in response to regulation are the most effective because of their intrinsic blind spots or the most thoroughly designed possible, they may be still be more intense and effective than what platforms would have done solely in response to their own commercial incentives.

Given an expected reduced effectiveness, instigators of information manipulations face several dilemmas. The first question is whether to allocate resources away from sheer manipulation and more toward overt, and legal, forms of influence that would not be moderated away. However, the fact that these malicious actors resort to manipulation is already evidence that they have less to offer in terms of overt positive message. In any case, this requires a large investment on the part of these actors as well as possibly different organizations.

The path of least resistance, which does not require drastic organizational changes for instigators of information manipulations, is to adapt operations to the new environment and to put relatively more efforts toward deploying operations on smaller and less moderated, platforms. The big drawback is that this reduces the opportunities to get a message into the mainstream.

9. Hannah Murphy, "TikTok Says It Is Working to 'Safeguard' US Data and National Security," *Financial Times*, July 1, 2022.

One response to this issue is to design operations that are systematically cross-platforms, funneling attention of targeted users from the mainstream platforms with mostly anodyne content (e.g., Facebook, Twitter) towards less moderated or partially encrypted platforms, where more malicious material can be delivered.[10] This exploits a blind spot in that there seems to be little formalized coordination forum between platforms on these issues although taking down notices have implied that there has been some cooperation when investigators from one platform found that malicious actors were also active on another. With a narrower base of potential target users, the goals may accordingly be adapted toward disinforming a few people more rather than disinforming many people a little. This would be consistent with messages seeking to motivate an active, more extremist, minority.

While very speculative, this discussion underscores the main causal path that is explored in this paper. Disinformation tactics and activities on content-sharing platforms are influenced by cost-benefit considerations that depend on which and how much content-moderation is enacted on platforms. These content-moderation efforts are themselves framed by the mechanics and technical possibilities that make moderation possible. In turn, platforms will tend to favor moderation solutions that are most compatible with their business model and their competitive concerns.

10. Interview with anonymous content moderation professional.

CONCLUSION

This note sought to contribute to our understanding of how private organizations managing content-sharing platforms create opportunities for information manipulation by fostering network effects, allowing the exploitation of the platforms by malicious actors. An analysis of the economic logic of content-sharing platforms provides insights on when they are more likely to act effectively against disinformation as well as the limits of the systems they put in place to deal with this issue. Thinking in terms of the business logic of the platforms also permits setting out scenarios for how disinformation on platforms may evolve in the wake of the enactment of the EU's Digital Service Act. Beyond the case of content-sharing platforms, this analysis gives an example of the insights that can be gleaned in our understanding of the role played by private organizations in their enabling, or inhibiting, new forms of inter-state conflictuality that are initiated in peace time in lieu of armed conflicts.

REPORT – No. 99

THE BUSINESS MODEL OF CONTENT-SHARING PLATFORMS AND THE SUPPLY OF CONTENT MODERATION

IMPLICATIONS FOR COMBATING INFORMATION MANIPULATIONS

Olivier Chatain

New forms of conflictuality below the threshold of violence often unfolds in spaces that are created and administered by private organizations, yet the roles played by these organizations in shaping the context in which conflicts happen, and their motivations, is rarely explored in security studies. This note explores the role played by content-sharing digital platforms in shaping the environment conducive to information manipulations. The note clarifies the economic incentives and constraints under which platforms operate. These incentives and constraints shape the essential design choices made by platforms, especially regarding the potency of network effects. This makes content-sharing platforms attractive targets for information manipulators who adapt their tactics to this new domain, but also affects the platforms' ability and incentives to conduct effective content moderation to counter manipulations. Using this conceptual toolbox, the note makes a preliminary assessment of the potential impact of the forthcoming Digital Service Act prepared by the European Union on platforms' efforts to moderate content, and the possible responses of malicious actors.

IRSEM, École militaire
1, place Joffre, case 38 - 75700 Paris SP 07
irsem.fr

IRSEM
INSTITUT DE RECHERCHE STRATÉGIQUE DE L'ÉCOLE MILITAIRE